UNIVERSIDADE DO ALGARVE

ESTIMAÇÃO EM PEQUENOS DOMÍNIOS COM MODELOS ESPACIOTEMPORAIS DE NÍVEL ÁREA

LUÍS MIGUEL SOARES NOBRE DE NORONHA E PEREIRA

DOUTORAMENTO EM
MÉTODOS QUANTITATIVOS APLICADOS À
ECONOMIA E À GESTÃO
NA ESPECIALIDADE DE ESTATÍSTICA

2009

UNIVERSIDADE DO ALGARVE

ESTIMAÇÃO EM PEQUENOS DOMÍNIOS COM MODELOS ESPACIOTEMPORAIS DE NÍVEL ÁREA

LUÍS MIGUEL SOARES NOBRE DE NORONHA E PEREIRA

DOUTORAMENTO EM
MÉTODOS QUANTITATIVOS APLICADOS À
ECONOMIA E À GESTÃO
NA ESPECIALIDADE DE ESTATÍSTICA

Tese orientada por: Professor Doutor Pedro Simões Coelho
Professor Doutor Rui Sousa Nunes

2009

Às mulheres da minha vida:

Lara e Ariana...

“Doing research is a creative process, there is no one way or right way of doing it; you need to discover what strategies work best for you.”

“Then one day that random variable of all random variables, the mind, puts things together in a slightly different way and it is solved. The solution often then seems simple and obvious.”

Hamada e Sitter (2004)

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS

AIC	Critério de Informação de Akaike
AICC	Critério de Informação de Akaike corrigido
ANOVA	Análise de Variância
AR(1)	Auto-regressivo de primeira ordem
BLUE	*Best Linear Unbiased Estimator*
BLUP	*Best Linear Unbiased Predictor*
BP	*Best Predictor*
BRAM	Enviesamento Relativo Absoluto Médio
BRN	Enviesamento Relativo Negativo
CAR	*Conditional Autoregressive Model*
CV	Coeficiente de Variação
CVM	Coeficiente de Variação Médio
DW	Durbin-Watson
EBLUP	*Empirical Best Linear Unbiased Predictor*
EBP	*Empirical Bayes Predictor*
EPR	Erro Padrão Relativo
EPRM	Erro Padrão Relativo Médio
EQM	Erro Quadrático Médio
EQMP	Erro Quadrático Médio de Predição
EQMRM	Erro Quadrático Médio Relativo Médio
ERAM	Erro Relativo Absoluto Médio
gl	graus de liberdade
IABH	Inquérito aos Preços de Avaliação Bancária na Habitação
IC	Intervalo de Confiança
iid	independente e identicamente distribuído
ind	independente
INE	Instituto Nacional de Estatística
IPTH	Inquérito aos Preços Médios de Transacção na Habitação
KS	Kolmogorov-Smirnov
MINQUE	*Minimum Norm Quadratic Unbiased Estimation*
MV	Máxima Verosimilhança

MVR	Máxima Verosimilhança Restrita
NUTSII	Nível II da Nomenclatura das Unidades Territoriais para Fins Estatísticos
NUTSIII	Nível III da nomenclatura das Unidades Territoriais para Fins Estatísticos
p	valor-*p*
RBAM	Rácio de Enviesamento Absoluto Médio
RV	Razão de Verosimilhanças
SAR	*Simultaneous Autoregressive Model*
SAS	*Statistical Analysis System*
SIPCH	Sistema de Indicadores de Preços na Construção e Habitação
STAR	*Simultaneous Spatio-temporal Autoregressive Model*
SW	Shapiro-Wilk
TCDBM	Taxa de Cobertura do Intervalo de Confiança *design-based* Média

LISTA DE NOTAÇÕES

$\mathbf{1}_n$: vector de uns de dimensão $n\times 1$.

$\mathbf{I}_n$: matriz identidade de ordem n.

$\mathbf{J}_n$: matriz de uns de ordem n, ou $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n'$.

$\mathbf{0}_n$: matriz de zeros de ordem n.

$\mathbf{0}_{n\times m}$: matriz de zeros de ordem $n\times m$.

$\mathbf{a}$: vector coluna

$\mathbf{a}'$: transposta do vector $\mathbf{a}$.

$\mathbf{A}$: matriz de ordem $n\times m$.

$\mathbf{A} = \{a_{ij}\}$: matriz $\mathbf{A}$ com o (i,j)-ésimo elemento da matriz representado por a_{ij}.

$\mathbf{A}'$: transposta da matriz $\mathbf{A}$.

$\mathbf{A}^{-1}$: inversa da matriz $\mathbf{A}$.

$\mathbf{A}^{-}$: inversa generalizada de Moore-Penrose da matriz $\mathbf{A}$.

$r(\mathbf{A})$: característica da matriz $\mathbf{A}$.

$tr(\mathbf{A})$: traço da matriz $\mathbf{A}$.

$col_{1\le i\le m}(\mathbf{a}_i)$: vector coluna $(\mathbf{a}_1',\ldots,\mathbf{a}_m')'$, onde $\mathbf{a}_1,\ldots,\mathbf{a}_m$ são vectores coluna; note-se que esta representação também inclui o vector coluna quando $a_1,\ldots,a_m$ são escalares.

$diag_{1\le i\le m}(\mathbf{A}_i)$: matriz diagonal por blocos com as matrizes $\mathbf{A}_1,\ldots,\mathbf{A}_m$ na sua diagonal; note-se que esta representação também inclui a matriz diagonal quando $A_1,\ldots,A_m$ são escalares.

$\mathbf{A}(\eta)$: os elementos da matriz $\mathbf{A}$ são função do parâmetro η.

$\frac{\partial \mathbf{A}(\boldsymbol{\eta})}{\partial \eta_i}; \frac{\partial \mathbf{A}}{\partial \eta_i}$: derivada parcial da matriz A em relação a η_i.

$o(D^{-s})$: infinitésimo de ordem inferior a D^{-s}.

$[O(D^{-s})]_{n\times m}$: matriz $n\times m$ com elementos de ordem $o(D^{-s})$.

$E(\boldsymbol{\xi})$: valor esperado do vector aleatório $\boldsymbol{\xi}$.

$Cov(\xi;\zeta)$: covariância entre as variáveis aleatórias ξ e ζ.

$V(\boldsymbol{\xi})$: Matriz de covariâncias de ordem n do vector aleatório $\boldsymbol{\xi}$ definido como $V(\boldsymbol{\xi}) = \left[Cov(\xi_i, \xi_j)\right]_{1 \le i, j \le n}$, onde ξ_i representa o i-ésimo elemento do vector $\boldsymbol{\xi}$ de dimensão $n \times 1$.

$N(\boldsymbol{\mu}; \boldsymbol{\Sigma})$: distribuição Normal multivariada, onde $\boldsymbol{\mu}$ representa o vector da média e $\boldsymbol{\Sigma}$ a matriz de covariâncias.

$\nabla g(\boldsymbol{\eta})$: gradiente do vector g em relação a $\boldsymbol{\eta}$.

$o_p(1)$: termo que converge em probabilidade para zero.

$O_p(1)$: termo que é limitado em probabilidade.

$\otimes$: produto directo (ou de Kronecker).

AGRADECIMENTOS

Ao Professor Doutor Pedro Simões Coelho, orientador desta tese, por quem me sinto em dívida, por uma paciência infinita, por toda a orientação científica, assim como pelas suas qualidades de compreensão excepcionais. Desejo ainda, expressar sinceros agradecimentos pela sua disponibilidade e amabilidade demonstrada ao longo destes anos de contacto frequente, no sentido de ajudar, aconselhar, corrigir e apontar a direcção certa para a investigação. Mas, a minha maior gratidão está cativa da sua amizade e companheirismo, pois são privilégios de que sempre gozei. A sua orientação foi, de facto, imprescindível neste trabalho e constitui uma grande dívida que contraí.

Ao Professor Doutor Rui Sousa Nunes, co-orientador desta tese, por todo o apoio, sugestões e críticas, que foram de incalculável valor e que contribuíram para o enriquecimento deste trabalho. Desejo também, expressar sinceros agradecimentos pela sua disponibilidade e amabilidade, pela sua compreensão e, em especial, pelas suas palavras de incentivo sempre presentes nos momentos mais difíceis.

À Fundação para a Ciência e a Tecnologia (FCT) pela Bolsa de Doutoramento que me atribuiu (referência SFRH/BD/36764/2007). Esta bolsa ajudou-me a suportar financeiramente o pagamento das propinas, a compra de bibliografia, a execução gráfica da tese, a frequência de vários cursos de curta duração e a participação em congressos nacionais e internacionais, fundamentais na minha formação, no estabelecimento de contactos com os melhores investigadores na área da estimação em pequenos domínios e na descoberta de novos caminhos de investigação.

À Escola Superior de Gestão, Hotelaria e Turismo da Universidade do Algarve, onde desenvolvo a minha actividade profissional, pelos apoios financeiros concedidos para participação na 56^{th} *Session of the International Statistical Institute* (2007) e no XVI Congresso Anual da Sociedade Portuguesa de Estatística (2008), bem como para as deslocações a Lisboa para ter reuniões científicas com o Professor Doutor Pedro Simões Coelho.

Ao Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, pelo acolhimento e pela disponibilização da licença do programa estatístico

Statistical Analysis System (SAS), sem o qual teria sido muito mais difícil realizar esta tese.

Ao Instituto Nacional de Estatística pela disponibilização dos dados que permitiram a realização do estudo empírico.

À minha mãe, por todo o amor, incentivo, paciência e incentivo durante toda a minha vida, e especialmente durante esta difícil fase por ter aberto mão das poucas horas que já tínhamos juntos em prol do cumprimento de mais esta meta.

Um agradecimento muito especial aos meus sogros, que são simplesmente excepcionais. Sem eles não teria de todo sido possível dedicar-me a 100% a esta investigação. A sua ajuda foi extremamente preciosa pelo facto de terem tomado conta da sua netinha, desde os primeiros meses de vida, como se fosse eu, deixando-me tempo a mim e à Lara para nos dedicarmos, com toda a energia, aos nossos trabalhos de investigação para doutoramento, que decorreram em simultâneo durante grande parte do tempo.

À Lara, companheira de vida e de trabalho, pelo amor, carinho, atenção e paciência, mas principalmente, pelo seu grande incentivo e inspiração, pelo constante apoio e pelas palavras de confiança, sem os quais este projecto não teria chegado ao fim. Não posso também deixar de expressar o meu agradecimento pelas sugestões e discussões relacionadas com alguns aspectos do estudo empírico deste trabalho, e por fim, pela sua leitura cuidadosa e crítica do produto final, cujas correcções subsequentes tiveram um peso relevante na versão final da tese.

Mas, a minha maior dívida está cativa da minha querida princesa Ariana, pelas incontáveis horas de mimos, colo, brincadeiras e passeios que lhe roubei em prol do cumprimento desta meta, e que nunca lhe poderei pagar… Posso apenas afiançar-lhe que vou tentar compensá-la daqui para a frente e prometer-lhe grandes passeios, muitos mimos e imensas brincadeiras…

RESUMO

Nesta tese são apresentados desenvolvimentos metodológicos ao nível da estimação em pequenos domínios no âmbito das sondagens, quando os dados amostrais têm uma natureza espacial e/ou cronológica. É proposto um estimador EBLUP (*Empirical Best Linear Unbiased Predictor*) assistido por um modelo linear misto de nível área, que permite especificar explicitamente a ligação entre os parâmetros de interesse e a informação auxiliar disponível. É também proposta uma metodologia que permite a introdução de restrições na estimação, garantindo a consistência interna na publicação das estimativas. São ainda propostas medidas de precisão *model-based* dos estimadores EBLUP, derivadas pela metodologia delta e por metodologias por reamostragem. Todos estes desenvolvimentos metodológicos são aplicados na estimação em pequenos domínios do preço médio de transacção da habitação em Portugal.

São realizados dois estudos empíricos por simulação de Monte Carlo. No primeiro estudo, por simulação *design-based*, é avaliada a qualidade dos estimadores propostos, com e sem restrições, relativamente a outros estimadores habitualmente utilizados na estimação em pequenos domínios. Os resultados deste estudo permitem concluir que os estimadores EBLUP são os que apresentam propriedades estatísticas de melhor qualidade. Em particular, o estimador EBLUP espaciotemporal proposto é o que apresenta melhores propriedades *design-based* na estimação do preço médio de transacção da habitação. Os resultados permitem também constatar que a garantia da consistência interna na publicação das estimativas conduz ao aumento do enviesamento e a perdas de eficiência dos estimadores, comparativamente a um estimador equivalente que não garante essa consistência. Por outro lado, a garantia da consistência interna a um nível mais agregado tem um efeito mais significativo sobre as propriedades do estimador que garante essa consistência, do que sobre o estimador que a garante a um nível geográfico mais desagregado. No entanto, se a consistência interna for garantida a um nível geográfico pouco agregado, então o efeito sobre a variância do estimador é quase insignificante.

No outro estudo por simulação *model-based*, é avaliado o desempenho dos estimadores do erro quadrático médio de predição (EQMP) dos EBLUP temporal e espaciotemporal.

Os resultados alcançados em ambos os casos permitem concluir que os estimadores baseados em métodos por reamostragem apresentam um desempenho muito bom, quando comparado com o do respectivo estimador analítico, podendo ser recomendados como uma alternativa a esse estimador analítico no contexto de modelos longitudinais mais complexos de estimação em pequenos domínios.

Palavras-chave: EBLUP; estimação com restrições; estimação do EQMP do EBLUP; estimação em pequenos domínios; métodos por reamostragem; modelo linear misto.

ABSTRACT

In this thesis methodological developments are presented at the small area estimation level in the framework of sampling surveys, namely when considering spatial and/or chronological sample data. An empirical best linear unbiased predictor (EBLUP) assisted by an area level linear mixed model is proposed. This estimator allows to explicitly specifying the link between the parameters of interest and the auxiliary information. It is as well proposed an approach that allows the introduction of restrictions in the estimation in order to guarantee internal consistency in the publication of estimates. In addition, model-based precision measures of the EBLUP estimators are proposed. These measures of precision are derived using the delta methodology and resampling methods. All of these methodological developments are applied in the estimation of the mean price of habitation transaction in Portugal.

Two empirical studies are carried out by Monte Carlo simulation. The first study uses a design-based simulation where the quality of the proposed estimators with and without restrictions is evaluated relatively to other estimators usually used in small area estimation. The results show that the EBLUP estimators present higher quality statistical properties. In particular the proposed spatio-temporal EBLUP estimator presents the best design-based properties in the small area estimation of the mean price of habitation transaction. From the results it is also possible to conclude that the guarantee of internal consistency in the publication of estimates conducts to an increase of bias and loss of efficiency of the estimators, comparatively to an equivalent estimator not internally consistent. The results also show that the guarantee of internal consistency at a more aggregated level has a more significant effect on the properties of the estimator that guarantees that consistency than on the estimator that guarantees the consistency at a more disaggregated geographic level. Nonetheless if internal consistency is guaranteed at a less aggregated geographic level, the effect on the estimator variance will be almost insignificant.

The second study uses a model-based simulation where the performance of the mean squared prediction error (MSPE) of the temporal and spatio-temporal EBLUP estimators is evaluated. The results demonstrate that in both cases the resampling-based

estimators have a very good performance when compared to the respective analytical estimator. Those estimators can be recommended as an alternative to that analytic estimator in the context of more complex longitudinal models for small area estimation.

Key-words: EBLUP; estimation with restrictions; estimation of the MSPE of the EBLUP; small area estimation; resampling methods; linear mixed model.

1. INTRODUÇÃO

1.1 ENQUADRAMENTO E RELEVÂNCIA

As decisões económicas e políticas são cada vez mais baseadas em indicadores estatísticos. Antigamente os decisores económicos e políticos suportavam as suas decisões em estatísticas nacionais, *i.e.*, ao nível de país ou em estatísticas referentes a grandes áreas geográficas. Contudo, nos últimos anos tem-se verificado por todo o mundo um grande crescimento na procura de estatísticas fiáveis com um maior nível de desagregação, geralmente ao nível de pequenas regiões. Este facto deve-se à utilização crescente deste tipo de estatísticas, por parte de organismos públicos e privados, com o objectivo de proporcionar informação que permita a formulação de melhores programas e políticas, a repartição de recursos mais equitativa e um melhor planeamento regional. Como consequência, a maioria dos produtores de estatísticas oficiais tem vindo a sentir a necessidade de publicação sistemática de indicadores estatísticos para níveis mais desagregados. Desta forma, as sondagens inicialmente planeadas para assegurarem a produção de estimativas directas[1] com uma determinada precisão ao nível de grandes áreas, são cada vez mais utilizadas para fornecer estimativas para domínios de diversa natureza e dimensão. Contudo, na prática raramente é possível garantir dimensões amostrais suficientemente grandes para suportarem estimativas directas com precisão aceitável em todos esses domínios de interesse. Por outro lado, frequentemente não é possível antecipar todas as utilizações dos dados amostrais no momento de planeamento da sondagem. O facto dos domínios de interesse não serem todos planeados no momento do desenho do plano de sondagem, leva a que as dimensões amostrais no

[1] As estimativas directas são produzidas através de estimadores directos, os quais utilizam apenas as observações da variável de interesse pertencentes ao domínio de estudo e referentes ao período de tempo em análise (Rao, 2003).

interior desses domínios sejam por vezes muito pequenas (ou mesmo nulas). Para tais domínios não planeados, os estimadores directos tradicionais apresentam normalmente uma precisão inaceitável, sendo mesmo impossível nalguns casos fazer a estimação através desses estimadores. Na literatura, são designados por pequenos domínios aqueles que representam uma fracção muito pequena da população e para os quais não é possível produzir estimativas directas com precisão aceitável. Nestes domínios, torna-se necessária a utilização de estimadores indirectos[2] que combinem informação auxiliar de fontes exteriores à sondagem, muitas vezes de natureza censitária ou administrativa, com a informação amostral disponível da variável de interesse. Esta informação auxiliar refere-se, em geral, a outras variáveis que apresentam correlação com as variáveis que são objecto de estudo ou às próprias variáveis em estudo em diferentes momentos do tempo.

A grande aplicação prática e interesse teórico da estimação em pequenos domínios têm atraído a atenção de alguns investigadores, os quais apresentaram importantes avanços metodológicos nesta área nos últimos anos. Estes novos desenvolvimentos materializaram-se sobretudo na proposta de novos estimadores indirectos para pequenos domínios, baseados em modelos de nível área e em modelos de nível unidade, e de estimadores do seu erro quadrático médio de predição (EQMP), os quais têm sido aplicados com sucesso a uma grande variedade de problemas de estimação em pequenos domínios por todo o mundo. Uma apresentação geral destes métodos e da sua aplicação pode ser encontrada em Rao (2003). Algumas das principais extensões aos modelos de nível área e aos modelos de nível unidade básicos, que têm sido propostas na literatura ao longo do tempo, incluem os modelos com erros amostrais correlacionados, os modelos que utilizam informação seccional e cronológica e os modelos que utilizam informação espacial.

Nos últimos anos foram propostos estimadores indirectos para pequenos domínios baseados em modelos que utilizam informação espacial/seccional e cronológica: Coelho (2000) propôs estimadores indirectos baseados em modelos lineares mistos de nível unidade e considerando restrições na estimação; e Singh *et al.* (2005) propuseram estimadores indirectos baseados em modelos de nível área do tipo *state space*,

[2] Os estimadores indirectos utilizam também observações da variável de interesse de fora do domínio de estudo ou do período de tempo considerado (Rao, 2003).

utilizando o filtro de Kalman para a estimação dos seus parâmetros. Por outro lado, ao longo destes anos também têm vindo a ser apresentados importantes desenvolvimentos metodológicos no sentido de se garantir a consistência interna na publicação das estimativas referentes a pequenos domínios: Pfeffermann e Tiller (2006) apresentaram estimadores indirectos para pequenos domínios baseados em modelos de nível área do tipo *state space* e considerando restrições na estimação; Wang *et al.* (2008) derivaram a forma do melhor preditor linear centrado (BLUP)[3] com restrições; e Ugarte *et al.* (2009) sistematizaram uma metodologia de introdução de restrições na estimação sob um modelo linear misto de nível unidade.

A utilização da estimação em pequenos domínios nas aplicações reais é muito variada, podendo destacar-se a sua crescente aplicação a problemas de estimação nas áreas da agricultura (*e.g.* produção de cereais), da saúde (*e.g.* características relacionadas com a saúde em grupos sócio-demográficos, mapeamento de doenças, taxas de mortalidade e de incidência) e da economia (*e.g.* rendimento médio *per capita*, taxa de desemprego, nível de pobreza, valor acrescentado bruto). Em particular, Girouard e Blondal (2001) e Diewert (2006) referem que alguns países desenvolvidos já criaram ou estão a criar índices de preços de transacção da habitação, alguns deles com grande nível de desagregação, os quais são necessários do ponto de vista económico, uma vez que a evolução dos preços da habitação apresenta normalmente uma correlação muito forte com o ciclo económico. Para além disso, e tal como sublinham Girouard e Blondal (2001), os preços da habitação têm um impacto positivo significativo no consumo privado e têm um efeito importante no investimento privado. Neste sentido, os preços da habitação podem ser utilizados como um *input* na medição da inflação no consumo privado, como deflatores, como indicadores de riqueza e como indicadores para analisar as pressões da procura.

Uma vez que o preço da habitação constitui um indicador muito importante na avaliação do estado da actividade económica de um país, tem-se verificado nos últimos anos um forte aumento na procura de estatísticas fiáveis de preços de transacção da habitação, especialmente por parte dos agentes dos sectores imobiliário e financeiro. A procura

[3] *Best Linear Unbiased Predictor* pode ser traduzido para português por melhor preditor linear centrado. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (BLUP), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

dessa informação tem-se dirigido sobretudo para pequenas subpopulações, isto é, para pequenas regiões geográficas e/ou para as várias tipologias da habitação numa dada região, aqui denominadas por domínios de estudo. Ao contrário do que se verifica em alguns países desenvolvidos (*e.g.* Austrália, Canadá, Dinamarca, Espanha, Estados Unidos da América, Noruega, Suécia), em Portugal ainda não existe um índice de preços de transacção da habitação.

Em Portugal, a Secretaria de Estado da Habitação e Comunicações, solicitou no final do último século ao Instituto Nacional de Estatística (INE), o desenvolvimento de um Sistema de Indicadores de Preços na Construção e Habitação (SIPCH). Este sistema visava a produção de informação estatística oficial sobre o mercado da habitação, em ópticas tão diversas como os custos de construção de habitação nova, os preços de transacção da habitação, os valores de avaliação bancária, os preços de manutenção e reparação e o financiamento da habitação. Passada uma década, o INE publica periodicamente um vasto conjunto de indicadores integrados no SIPCH, embora ainda não publique um indicador dos preços de transacção da habitação a qualquer nível de agregação. A inexistência desse indicador deve-se sobretudo à escassez de informação amostral, a qual inviabilizou a produção, com precisão aceitável, de estimativas directas do preço médio de transacção da habitação por metro quadrado, para níveis mais desagregados do que as regiões de Portugal classificadas ao nível II da Nomenclatura das Unidades Territoriais para Fins Estatísticos (NUTSII). A referida informação amostral utilizada era proveniente do Inquérito aos Preços Médios de Transacção na Habitação (IPTH), que era um inquérito longitudinal com rotação, da responsabilidade do INE. Ainda neste contexto, era intenção do INE a publicação de estatísticas que verificassem a consistência interna, ou seja, que as estimativas produzidas para um determinado nível de agregação coincidissem com o resultado do agrupamento das estimativas produzidas para níveis de agregação inferiores (pequenos domínios). Naturalmente que a impossibilidade de produção de estatísticas para pequenos domínios inviabilizou qualquer tentativa de garantia da consistência interna.

De acordo com Diewert (2006), já existe um conjunto de países a desenvolver e a aplicar metodologias de estimação do preço médio de transacção da habitação, embora nenhuma delas responda ao problema da escassez de informação amostral ao nível dos pequenos domínios de estudo (*e.g.* ao nível III da Nomenclatura das Unidades

Territoriais para Fins Estatísticos - NUTSIII - ou ao nível de concelho). Este trabalho de investigação pretende contribuir para colmatar esta falta. De facto, este trabalho assenta na convicção que a obtenção de estimativas mais precisas, e que verifiquem a consistência interna, para o parâmetro de interesse nas referidas subpopulações, passa pela utilização de estimadores indirectos para pequenos domínios baseados em modelos e, naturalmente, pelo recurso a todo o tipo de informação auxiliar disponível.

Relembrando-se uma opinião de Rao (2005*b*), que segundo o qual a estimação em pequenos domínios é um exemplo notável de ligação entre a teoria e a prática, que apesar de ter apresentado avanços teóricos importantes, necessita ainda de desenvolvimentos para muitos problemas de índole prática, então decidiu aproveitar-se esta ideia. Um dos temas apontados por Rao (2005*b*) que carece de mais investigação teórica consiste na garantia da consistência interna dos estimadores baseados em modelos, ou seja, no processo de introdução de restrições na estimação de forma a assegurar que as estimativas produzidas ao nível dos pequenos domínios possam ser agregadas, igualando estimativas directas produzidas para um nível de agregação superior.

Pela sua importância e actualidade, pelo facto de responder a um problema de ordem prática quando se trabalha com estatísticas oficiais e por ainda não ter sido alvo de investigação até ao momento, neste trabalho pretende estudar-se e apresentar-se desenvolvimentos metodológicos ao nível dos modelos de nível área para estimação em pequenos domínios com dados espaciais e cronológicos, sob o modelo linear misto. Pretende também incluir-se restrições na estimação, de forma a se alcançar a consistência interna dos estimadores baseados nesse tipo de modelos. Neste estudo, os desenvolvimentos teóricos serão aplicados na estimação do preço médio de transacção da habitação por metro quadrado de área útil, ao nível das NUTSIII de Portugal continental. Procurar-se-á, desta forma, dar resposta a um problema de produção de informação estatística para pequenos domínios com precisão adequada, ainda não resolvido pelo INE.

A metodologia proposta para estimação do preço médio de transacção da habitação por metro quadrado para pequenos domínios (*i.e.* ao nível de NUTSIII), pretende explorar toda a informação amostral disponível, nas suas vertentes espacial e temporal. Em particular, será explorada, por um lado, a informação auxiliar fornecida pelo Inquérito

aos Preços de Avaliação Bancária na Habitação (IABH) e os dados de painel fornecidos pelo IPTH, e por outro lado, a eventual associação espacial existente entre domínios. Uma vez que não é possível fazer a ligação dos dados relativos à variável de interesse e à variável auxiliar ao nível individual, isto é, ao nível de transacção, a investigação será circunscrita a modelos ao nível de área, que estabelecem a associação das várias fontes de dados ao nível de domínio de estudo.

Por último, é ainda de sublinhar que o trabalho de investigação desenvolvido no âmbito desta tese está orientado para os problemas de estimação/predição, apesar dos problemas de inferência em pequenos domínios, no âmbito de uma sondagem, deverem ser analisados numa perspectiva integrada, logo na fase de planeamento da sondagem. A escolha deste enfoque deve-se ao facto de se pretender, com este trabalho, dar resposta a um problema de produção de informação estatística em pequenos domínios, numa fase em que os dados amostrais já estão disponíveis e os domínios de interesse só foram identificados *a posteriori*. Quando é possível a identificação *a priori* dos domínios para os quais se pretende obter a informação, ou seja, quando os domínios são planeados, e os recursos disponíveis são suficientes, então o plano de sondagem deverá ser desenhado de forma a permitir a produção de estimativas de precisão adequada para os domínios de interesse (Coelho, 2000).

1.2 OBJECTIVOS

Com o trabalho de investigação desenvolvido nesta tese, pretende dar-se um contributo metodológico na área da estimação em pequenos domínios no âmbito de inquéritos por amostragem com dados de nível área de natureza espacial e/ou cronológica. Em simultâneo, pretende resolver-se um problema de estimação de índole prático, através da proposta de uma metodologia de estimação, ao nível de NUTSIII, do preço médio de transacção da habitação por metro quadrado de área útil, em Portugal continental, com precisão aceitável. No âmbito da aplicação prática, pretende ainda garantir-se a consistência aritmética na publicação dessas estimativas, com as disponíveis para níveis mais agregados. Neste sentido, definem-se cinco objectivos específicos para este estudo, enumerados de seguida:

1. Propor estimadores de parâmetros de interesse em pequenos domínios, alternativos aos estimadores tradicionais que têm vindo a ser utilizados na estimação em domínios (estimadores directos e indirectos), e que possam constituir uma opção válida para inferência em pequenos domínios do preço médio de transacção da habitação por metro quadrado de área útil, em Portugal continental. Mais especificamente, pretende propor-se um modelo de nível área de estimação em pequenos domínios, que tire partido de dados de natureza espacial e/ou cronológica, bem como deduzir as expressões explícitas dos estimadores combinados assistidos por esse modelo. Com os novos estimadores tenciona-se alcançar melhores níveis de precisão e/ou de enviesamento, do que aqueles que podem ser obtidos pela utilização do estimador directo, ou de estimadores indirectos frequentemente utilizados no âmbito da estimação em pequenos domínios.

2. Apresentar uma metodologia que permita a introdução de restrições nos modelos de estimação em pequenos domínios, de forma a garantir a consistência aritmética na publicação das estimativas do preço médio com as publicadas a níveis mais agregados (*e.g.* NUTSII ou Portugal continental). Em complementaridade, a introdução de restrições nos modelos de estimação em pequenos domínios pretende conferir aos estimadores uma maior robustez contra possíveis falhas e/ou más especificações dos modelos que assistem a estimação.

3. Propor estimadores das componentes de variância do modelo espaciotemporal de nível área que assiste a estimação. Apesar da estimação das componentes de variância não constituir um objectivo central deste trabalho, elas têm que ser estimadas de forma a ser possível estimar e avaliar a variabilidade dos estimadores dos parâmetros de interesse em pequenos domínios.

4. Apresentar medidas de precisão *model-based* dos estimadores combinados (com e sem restrições) propostos neste estudo, derivadas a partir do modelo espaciotemporal que assiste essa estimação, bem como de outros estimadores combinados existentes na literatura, mas utilizados no âmbito deste estudo. Mais especificamente, pretende propor-se estimadores do EQMP de estimadores combinados dos parâmetros de interesse em pequenos domínios.

5. Avaliar, por simulação de Monte Carlo, o desempenho dos estimadores propostos para estimar o preço médio de transacção da habitação, relativamente a diversos outros estimadores directos e indirectos habitualmente utilizados na estimação desse tipo de parâmetros de interesse em pequenos domínios. Em complementaridade a este objectivo, pretende ainda avaliar-se, por simulação de Monte Carlo, o desempenho dos estimadores do EQMP propostos para medir a incerteza de estimadores combinados de parâmetros de interesse sem restrições, utilizados no âmbito deste estudo.

1.3 METODOLOGIA

Neste estudo pretende desenvolver-se uma metodologia de estimação do preço médio de transacção da habitação por metro quadrado ao nível de NUTSIII, para Portugal. A metodologia de estimação em pequenos domínios baseada em modelos (*model-based*) afigura-se como a mais adequada porque se pretende estimar preços médios ao nível de subpopulações, de dimensão amostral não controlável e na maior parte dos casos muito pequena, inviabilizando a estimação directa desses preços médios com boa precisão. As subpopulações ou domínios de estudo são definidos pelo nível de desagregação geográfica que se pretende para a estimação do preço médio de transacção da habitação, tendo-se escolhido o nível NUTSIII por razões de ordem prática e por ser o primeiro nível de desagregação a não proporcionar, em todos os domínios, a produção de estimativas directas do parâmetro de interesse com precisão aceitável para publicação pelos organismos oficiais. Para se alcançarem os objectivos definidos, a realização deste estudo seguirá a seguinte metodologia:

1. Pretende-se que os novos estimadores do preço médio de transacção da habitação permitam incorporar informação relativa a variáveis auxiliares, muitas vezes de natureza censitária ou administrativa, bem como informação amostral de natureza espacial e/ou cronológica, exógena aos domínios de interesse, com o objectivo de auxiliar a estimação a esse nível. Neste sentido, os estimadores propostos serão assistidos por um modelo espaciotemporal de nível área, que se enquadra no contexto do modelo linear misto, e é suficientemente flexível para representar fenómenos que possam ser descritos através de dados de natureza espacial e/ou cronológica. De

facto, parece potencialmente interessante explorar toda a flexibilidade do modelo linear misto para representar fenómenos que possam originar dados de natureza espacial e/ou cronológica, não só através da inclusão de efeitos fixos e de efeitos aleatórios, mas também através da possível especificação de estruturas de covariância espacial e cronológica arbitrárias. Os estimadores propostos, assistidos por esse modelo, podem ser considerados como melhores preditores lineares centrados empíricos (EBLUP)[4] dos parâmetros nos domínios de interesse. Pretende utilizar-se os modelos lineares mistos e não os modelos do tipo *state space* para incorporar informação de natureza espacial e/ou cronológica no modelo e estimar os seus parâmetros, porque os primeiros permitem contemplar estruturas de covariância que não são passíveis de representação num modelo do tipo *state space*, que é apenas um dos seus inúmeros casos particulares. Ao contrário dos modelos do tipo *state space*, que utilizam informação referente aos períodos anteriores para auxiliar unicamente a estimação no período presente, a modelação conjunta de dados de diversos períodos através de um modelo linear misto, permite também a estimação de parâmetros em períodos passados, utilizando a informação referente a períodos mais recentes.

2. No quadro de uma abordagem de estimação *model-based* ou *model-assisted*, é vantajoso que os estimadores do preço médio de transacção da habitação incluam mecanismos que assegurem alguma protecção contra possíveis falhas e/ou más especificações dos modelos que os suportam. Por outro lado, quando se trabalha com estatísticas publicadas, e em particular com estatísticas oficiais, é exigido frequentemente que as estimativas produzidas ao nível dos pequenos domínios (*e.g.* NUTSIII) possam ser agregadas, igualando estimativas directas suficientemente precisas para níveis de agregação mais elevados (*e.g.* NUTSII ou Portugal continental). Estas preocupações poderão ser superadas através da introdução de restrições na estimação, constituindo uma forma implícita de utilização de "informação emprestada" dos domínios de interesse. Pretende assim garantir-se que são produzidas estimativas do preço médio de transacção da habitação adequadas e consistentes, para um determinado nível de agregação, mesmo que o modelo

[4] *Empirical Best Linear Unbiased Predictor* pode ser traduzido para português por melhor preditor linear centrado empírico. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (EBLUP), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

postulado falhe ou não esteja perfeitamente especificado. A introdução de restrições na estimação no âmbito do modelo linear misto será inspirada no conhecido problema da econometria clássica de estimação dos parâmetros de uma regressão sob restrições lineares.

3. Apesar da estimação das componentes de variância não constituir um dos objectivos principais no âmbito da estimação em pequenos domínios, elas têm que ser estimadas de forma a se obterem estimativas dos parâmetros de interesse. Na literatura pode ser encontrado um vasto leque de métodos de estimação das referidas componentes de variância no contexto do modelo linear misto, desde os métodos que não exigem nenhuns pressupostos até aos métodos mais exigentes a esse nível. Neste estudo, a estimação das componentes de variância será baseada no método dos momentos. Pretende estimar-se as componentes de variância através do método dos momentos e não através de outros métodos eventualmente mais populares, como os conhecidos métodos de verosimilhança, por os primeiros não exigirem pressupostos, nomeadamente o da normalidade dos erros do modelo, que muitas vezes não se verificam na realidade.

4. A proposta de medidas de precisão *model-based* dos estimadores propostos/utilizados para estimar o preço médio de transacção da habitação será efectuada através da dedução/apresentação de aproximações do EQMP desses estimadores dos parâmetros de interesse. Este trabalho de investigação, que tem sido considerado como um dos maiores desafios no âmbito da estimação em pequenos domínios, assentará numa metodologia baseada em desenvolvimentos em série de Taylor (normalmente denominada por metodologia delta) e em metodologias por reamostragem (métodos *jackknife* e *bootstrap*), por serem aquelas que têm vindo a ser utilizadas em outros trabalhos de investigação na área da estimação em pequenos domínios. Para além disso, as metodologias por reamostragem são consideradas actualmente como uma boa alternativa ao método delta, devido à sua simplicidade conceptual, à sua facilidade de aplicação mesmo a modelos mais complexos e, geralmente, à menor exigência de pressupostos.

5. A avaliação do desempenho dos estimadores será efectuada através da realização de estudos empíricos por simulação, de diferentes naturezas:

a) será efectuado um estudo por simulação do tipo *design-based* para avaliar o desempenho dos estimadores de parâmetros de interesse em pequenos domínios. Decidiu utilizar-se este tipo de estudo empírico, baseado em amostragem repetida de uma pseudo-população finita (fixa) gerada a partir de uma amostra aleatória de dados reais, por ser efectuado no contexto de uma população real e de um método de amostragem realista. A avaliação do desempenho dos estimadores dos parâmetros de interesse será efectuada através de um conjunto de medidas de enviesamento, de precisão e de eficiência, sob uma perspectiva de amostragem repetida.

b) será efectuado um estudo por simulação do tipo *model-based* para avaliar o desempenho dos estimadores do EQMP dos estimadores combinados de parâmetros de interesse sem restrições em pequenos domínios. Decidiu utilizar-se este tipo de estudo empírico, baseado em sucessivos conjuntos de dados gerados a partir de um modelo de superpopulação postulado *a priori*, pois só é possível avaliar a qualidade daquele tipo de estimadores com dados gerados artificialmente com as características desejadas. A avaliação do desempenho dos referidos estimadores será efectuada através de um conjunto de medidas de enviesamento e de precisão, sob uma perspectiva *model-based*.

No estudo empírico do tipo *design-based*, será igualmente abordado o problema da robustez dos estimadores de parâmetros de interesse em pequenos domínios, analisando as consequências da introdução de restrições na estimação sobre a precisão e enviesamento desses estimadores. Neste tipo de estudo empírico por simulação de Monte Carlo, serão utilizados dados reais de painel, fornecidos pelo IPTH e pelo IABH, ambos da responsabilidade do INE. As tarefas de manipulação de dados e de cálculo serão realizadas no programa estatístico *Statistical Analysis System* (SAS), desenvolvendo-se para tal programas na sua linguagem. Em todos os ensaios de hipóteses efectuados no âmbito deste trabalho, será considerada uma probabilidade de rejeitar uma hipótese que de facto é verdadeira de 5%. Para além disso, todos os intervalos de confiança (IC) serão construídos com um grau de confiança de 95%.

1.4 PLANO DA TESE

Esta tese encontra-se organizada em oito capítulos e três anexos, sendo complementada por 33 apêndices. Neste primeiro capítulo que termina com o plano da tese, foi apresentado um enquadramento e exposta a relevância do problema de investigação, foram definidos os objectivos deste trabalho e foi sumariada a metodologia a seguir.

No segundo capítulo são apresentados alguns conceitos fundamentais, notação e definições. O conceito de pequeno domínio é introduzido neste capítulo. São também apresentadas as abordagens de estimação *design-based* e *model-based*, assim como algumas classificações que podem ser atribuídas aos modelos de estimação em pequenos domínios. Neste capítulo são ainda apresentados alguns estimadores combinados habitualmente utilizados na estimação em domínios.

O terceiro capítulo é dedicado à revisão bibliográfica sobre predição em modelos lineares mistos. Neste capítulo é apresentado o modelo linear misto geral, o modelo *state space* linear geral como caso particular do primeiro, e alguns modelos espaciais gerais. No âmbito do modelo linear misto, que assiste os estimadores para pequenos domínios propostos neste trabalho, é dado particular destaque ao problema da medição da incerteza associada ao EBLUP.

Os principais modelos de nível área para estimação em pequenos domínios presentes na literatura são apresentados no capítulo quarto. Para além do modelo básico de nível área para dados seccionais, ao qual é dado maior enfoque, são também apresentados os modelos básicos para dados seccionais e cronológicos, para dados espaciais e para dados simultaneamente espaciais/seccionais e cronológicos. Por facilidade de exposição, neste capítulo são também propostas duas metodologias por reamostragem para medição da incerteza associada ao EBLUP temporal no âmbito do conhecido modelo longitudinal de Rao e Yu (1994), e um estimador, pelo método dos momentos, da componente de variância do modelo espacial de Salvati (2004).

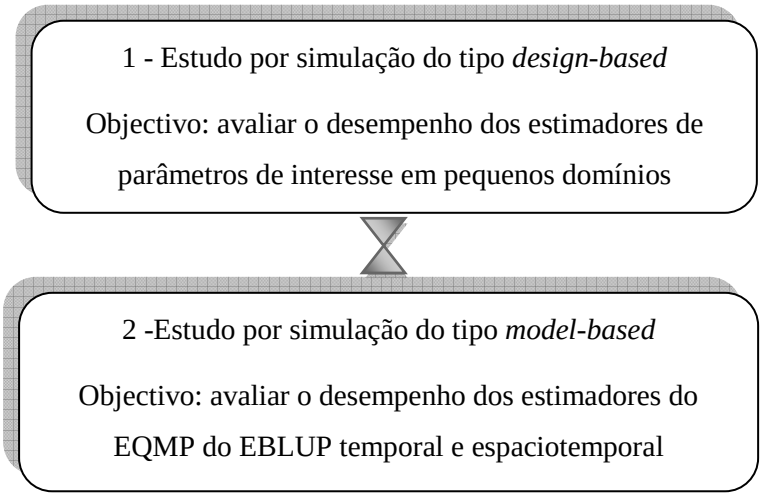
No quinto capítulo são propostos estimadores para os parâmetros de interesse em pequenos domínios, assistidos por um modelo espaciotemporal de nível área que se enquadra no contexto do modelo linear misto. Esses estimadores podem incorporar simultaneamente dados de natureza espacial/seccional e cronológica, tirar partido de

informação auxiliar relativa a outras variáveis conhecidas sobre a população, bem como acomodar estruturas de covariância cronológica e espacial auto-regressivas entre os dados amostrais. Neste capítulo são também propostos estimadores dos parâmetros de variância do modelo espaciotemporal. Na parte final do quinto capítulo é discutido o importante problema de medição da incerteza associada aos estimadores dos parâmetros de interesse, sendo propostas três abordagens de estimação do EQMP do EBLUP espaciotemporal.

O sexto capítulo é dedicado à estimação com restrições. Na primeira parte deste capítulo é efectuada uma revisão bibliográfica sobre a garantia da consistência interna no contexto da estimação em pequenos domínios. Na segunda parte do capítulo é introduzido um modelo linear misto geral com restrições para estimação em pequenos domínios, no âmbito do qual é deduzido o EBLUP com restrições. Neste contexto, são também propostas duas metodologias de estimação do EQMP do EBLUP com restrições. Por último, é apresentado, como caso particular, um modelo com restrições para estimação em pequenos domínios com dados espaciais e cronológicos.

A apresentação dos estudos empíricos, por simulação de Monte Carlo, é feita no sétimo capítulo. O capítulo inicia-se com uma apresentação da sua estrutura, na qual é revelada a ordem pela qual são apresentados os diferentes estudos empíricos. Em seguida, é efectuada uma descrição detalhada de cada dos estudos empíricos, ao que se segue uma apresentação e discussão dos resultados. A figura 1.4.1 ilustra a organização dos estudos empíricos apresentados no capítulo sétimo.

Figura 1.4.1: Organização dos estudos empíricos



No oitavo e último capítulo, apresentam-se as principais conclusões e limitações deste estudo, tendo em conta os objectivos definidos. São ainda apresentados alguns caminhos de investigação futura, estando já alguns deles a ser trilhados.

Nesta tese existem três anexos, nos quais são apresentados, respectivamente, o IPTH, o IABH e o mapa das NUTSIII de Portugal continental. Depois dos anexos é apresentada a bibliografia da tese.

Os 33 apêndices da tese constituem um volume autónomo. Nos apêndices encontram-se algumas informações relativas aos estudos empíricos, bem como resultados dos estudos por simulação de Monte Carlo e os programas desenvolvidos no SAS que suportam esses estudos por simulação.

2. CONCEITOS FUNDAMENTAIS E ESTIMADORES COMBINADOS PARA DOMÍNIOS

2.1 INTRODUÇÃO

Neste capítulo é definido o conceito de pequeno domínio (subcapítulo 2.2) e são apresentadas algumas noções básicas da teoria das sondagens, tais como população, amostra, plano de sondagem e estimador (subcapítulo 2.3). Nos subcapítulos 2.4 e 2.5, são apresentadas as abordagens de estimação *design-based* e *model-based*[5], bem como as propriedades estatísticas dos estimadores no contexto dessas abordagens de estimação. No subcapítulo 2.6 é efectuada uma introdução aos estimadores combinados para domínios. Por último, no subcapítulo 2.7. são identificadas algumas classificações que podem ser atribuídas aos modelos de estimação em pequenos domínios.

2.2 CONCEITO DE PEQUENO DOMÍNIO

Designa-se por domínio de estudo uma subpopulação de dimensão desconhecida, para a qual se pretendem estimar parâmetros, ou seja, uma subpopulação para a qual se pretendem produzir estimativas pontuais ou por intervalos para uma característica de interesse, antes ou depois da fase de planeamento da sondagem. Se for possível

[5] A tradução para português de *design-based* e de *model-based* pode ser baseado no desenho e baseado no modelo, respectivamente. No entanto, decidiu utilizar-se nesta tese as referidas expressões em inglês por aquelas não serem traduções oficiais para a língua portuguesa, não constando no Glossário Estatístico Inglês-Português da Sociedade Portuguesa de Estatística (SPE) e da Associação Brasileira de Estatística (ABE) (SPE e ABE, 2007).

identificar os domínios de estudo antes de a amostra ser recolhida, então o plano de sondagem a adoptar deverá ter em consideração esses domínios planeados, considerando-os como estratos, grupos de estratos ou conglomerados[6]. Nesta situação, a dimensão da amostra em cada domínio deverá ser suficientemente grande para produzir estimativas directas com precisão aceitável. Nos casos em que é impossível (por exemplo, por ausência de uma base de sondagem) ou impraticável (por exemplo, devido aos custos elevados) tratar esses domínios como estratos, grupos de estratos ou conglomerados ao nível do desenho da amostra, então eles designam-se por domínios não planeados. Os domínios de estudo não planeados também podem resultar de uma identificação tardia (após a selecção da amostra) de necessidades de inferência ao nível de determinadas subpopulações. Contudo, mesmo na situação de domínios planeados, é frequente encontrarem-se domínios com poucas observações, ou mesmo com nenhumas (Särndal *et al.*, 1992).

De acordo com Rao (2003), um domínio é considerado grande se a amostra específica desse domínio for suficientemente grande para produzir estimativas directas com precisão adequada. Pelo contrário, um domínio é considerado pequeno se a amostra específica desse domínio não for suficientemente grande para suportar estimativas directas com precisão adequada, ou seja, à expressão "pequeno domínio" está associado o problema da ausência de uma amostra de dimensão significativa nesse domínio[7].

A expressão "pequeno domínio" define, portanto, um qualquer domínio que represente uma fracção muito pequena da população e para o qual não é possível produzir estimativas directas com precisão aceitável. Segundo Rao (2003), a expressão "pequeno domínio" é frequentemente utilizada como designação de áreas geográficas pequenas, como por exemplo, secções estatísticas, freguesias, concelhos ou até mesmo NUTSIII; ou como referência a subpopulações pequenas. As subpopulações pequenas podem ser o resultado de grupos populacionais pequenos (por exemplo, as minorias étnicas ou os

[6] Através da redefinição das unidades de amostragem em amostragens multi-etápicas.

[7] Contudo, existem outros autores que definiram outros tipos de domínios no que se refere à sua dimensão. Purcell e Kish (1979) definiram quatro tipos de domínios em função da dimensão relativa de cada domínio, $P_d=N_d/N$, onde N é a dimensão populacional global e N_d a dimensão populacional do domínio d: se $P_d \geq 0{,}1$, o domínio é considerado grande; se $0{,}01 \leq P_d < 0{,}1$, o domínio é considerado pequeno; se $0{,}0001 \leq P_d < 0{,}1$, o domínio é considerado muito pequeno; e se $P_d < 0{,}0001$, o domínio é considerado escasso.

portadores de uma doença muito específica ou rara), mas também podem ser o resultado de subgrupos da população resultantes de variáveis categóricas (por exemplo, um determinado sector de actividade económica), bem como de cruzamentos de variáveis categóricas (por exemplo, "classe etária × nível de instrução" ou "sector de actividade económica × classe de número de colaboradores").

2.3 NOÇÕES BÁSICAS DA TEORIA DAS SONDAGENS

2.3.1 População, amostra e variável de interesse

Uma população designa um conjunto de elementos em relação aos quais se pretende conhecer uma determinada característica. Considere-se uma população alvo ou universo de referência com dimensão finita e conhecida. A população finita denota-se por *U* e a sua dimensão por *N*. A cada elemento da população, geralmente designado por unidade estatística ou indivíduo, pode ser associado um índice *j* (*j*=1, ..., *N*), sendo cada elemento representado por u_j. Por uma questão de simplicidade de notação, vai denotar-se o *j*-ésimo elemento da população, u_j, pelo respectivo índice *j*, podendo representar-se a população alvo finita como $U = \{1, ..., j, ..., N\}$.

No âmbito de um inquérito por amostragem, uma amostra é um subconjunto de elementos de uma população *U*, a partir da qual são recolhidos dados de forma a permitir a sua extrapolação para o conjunto da população. Uma amostra particular é geralmente denotada por *s* e a sua dimensão por n. Uma amostra de n elementos retirada da população *U*, pode então representar-se por $s = \{1, ..., n\}$.

Considere-se que a população finita com *N* unidades estatísticas se encontra dividida em *m* subpopulações mutuamente exclusivas, U_i com *i*=1, ..., *m*, de dimensões N_i, que se denominam por domínios de estudo. Normalmente, as dimensões dos domínios na população, N_i, são desconhecidas. Tem-se então que $U = \bigcup_{i=1}^{m} U_i$, $U_i \cap U_{i'} = \emptyset, \quad i \neq i'$ e $N = \sum_{i=1}^{m} N_i$. Também por uma questão de simplicidade de notação, vai denotar-se o

i-ésimo domínio da população, U_i com *i*=1, ..., *m*, pelo respectivo índice *i*. Quando cada elemento populacional, *j*, está referenciado ao domínio a que pertence, *i*, pode igualmente representar-se pelo respectivo índice *ij*.

Considere-se agora que as unidades estatísticas que compõem a população finita se encontram agrupadas em *A* subpopulações exaustivas e mutuamente exclusivas, U_g com *g*=1, ..., *A*, de dimensões N_g, designadas por unidades primárias ou conglomerados. Neste caso, os conglomerados formam uma base de amostragem de onde é seleccionada a amostra, pelo que a população de conglomerados pode ainda ser representada simbolicamente por[8] $U_G = \{1, ..., g, ..., A\}$. Pelo agrupamento das N unidades estatísticas em A conglomerados, tem-se que $U = \bigcup_{g=1}^{A} U_g$, $U_g \cap U_{g'} = \emptyset, \quad g \neq g'$, e $N = \sum_{g=1}^{A} N_g$.

Considere-se ainda que a população de *A* conglomerados se encontra previamente estratificada em *H* grupos homogéneos, exaustivos e mutuamente exclusivos, U_h com *h*=1, ..., *H*, sendo o número de conglomerados em cada um desses grupos representado por A_h. Geralmente estes grupos são estratos definidos *a priori* ou pós-estratos que podem atravessar os *m* domínios de interesse. Mais ainda, considerando-se que a população de conglomerados está dividida em *H* grupos e em *m* domínios que se intersectam, designa-se por U_{hi}, com *h*=1, ..., *H*, *i*=1, ..., *m* (ou pelo índice *hi*), a intersecção na população do estrato $U_{h\bullet}$ com o domínio $U_{\bullet i}$, de dimensão A_{hi}. Tem-se que[9] $U = \bigcup_{h=1}^{H} U_{h\bullet} = \bigcup_{i=1}^{m} U_{\bullet i} = \bigcup_{i=1}^{m} \bigcup_{h=1}^{H} U_{hi}$ e $A = \sum_{h=1}^{H} A_h$.

A cada unidade estatística da população está associado o valor de uma variável de interesse, representada por *Y*. A variável de interesse representa a característica da população que se pretende estudar. O valor da variável de interesse no *j*-ésimo elemento da população representa-se por y_j (*j*=1, ..., *N*), de modo que o conjunto dos valores de *Y* assumidos pelos *N* elementos da população pode ser representado por $Y = \{y_1, ..., y_N\}$,

[8] Para simplificação de notação, vai representar-se o conglomerado U_g apenas pelo índice *g* (*g*=1, ..., *A*).

[9] Para simplificação de notação, vai representar-se $U_{h\bullet}$ e $U_{\bullet i}$ apenas por U_h e U_i, respectivamente.

sendo o conjunto de valores observados na variável de interesse junto das *n* unidades amostrais representado por $y_s = \{y_1, ..., y_n\}$.

Quando o valor da variável de interesse está referenciado ao *j*-elemento populacional do domínio *i* no período *t*, (*i*=1, ..., *m*; *t*=1, ..., *T*), representa-se esse valor por y_{tij}. Desta forma, considera-se que a variável de interesse está referenciada a *T* períodos de tempo, os quais correspondem a momentos de observação.

2.3.2 Plano de sondagem

No âmbito de um inquérito repetido no tempo, seja S_t o conjunto de todas as possíveis amostras da mesma dimensão que é possível extrair de uma população *U* no período *t*, e $p_t(.)$ a função de probabilidade de S_t. Desta forma, $p_t(s_t)$ é a probabilidade de extrair a amostra s_t no período *t* de entre o conjunto de todas as amostras possíveis. O plano de sondagem associado ao período *t*, pode então ser entendido como um par $(S_t, p_t(.))$, onde $p_t(s_t) > 0$ para toda a amostra $s_t \in S_t$.

Considere-se que no período *t* é retirada uma amostra aleatória s_t, de dimensão n_t, da população *U* de acordo com um determinado plano de sondagem, $p_t(s_t)$. Seja s_{ti} o conjunto dos elementos de s_t que intersectam o domínio *i*, isto é, $s_{ti} = s_t \cap U_i$ (*t*=1, ..., *T*; *i*=1, ..., *m*). Segundo Särndal *et al*. (1992), a dimensão de s_{ti}, denotada por n_{ti}, é aleatória e por vezes é extremamente reduzida, admitindo-se que em algumas casos possa ser nula. O plano de sondagem ou, mais propriamente, as probabilidades de inclusão de primeira ordem, irão determinar se é de esperar uma boa ou má dimensão amostral no *i*-ésimo domínio. Neste caso, tem-se que $s_t = \bigcup_{i=1}^{m} s_{ti}$ e $n_t = \bigcup_{i=1}^{m} n_{ti}$.

Represente-se agora por s_{Gt} uma amostra de conglomerados extraída a partir de *U* no período *t*, de dimensão a_t, seleccionada aleatoriamente de acordo com um determinado plano de sondagem, $p_{Gt}(\cdot)$. Numa sondagem aleatória por conglomerados são observadas todas as unidades estatísticas (unidades secundárias) dos conglomerados (unidades primárias) seleccionados para a amostra s_{Gt}. No caso em que é possível

identificar estratos ou pós-estratos na população de conglomerados, também a amostra s_{Gt} pode ser dividida em subamostras s_{Gth}, t=1, ..., T; h=1, ..., H. O conjunto de unidades secundárias que são observadas no estrato h continua a designar-se por s_{th} e o conjunto total de unidades secundárias que são observadas no período t, por s_t, ou seja, $s_t = \bigcup_{h=1}^{H} \bigcup_{g \in s_{Gth}} U_g$. A dimensão da amostra de unidades secundárias no estrato h (aleatória) no período t, é dada por $n_{th} = \sum_{g \in s_{Gth}} N_g$, sendo a dimensão da amostra de unidades secundárias (também aleatória) no mesmo período, dada por $n_t = \sum_{h=1}^{H} n_{th} = \sum_{g \in s_{Gt}} N_g$.

Uma amostra s_t não é seleccionada de entre o conjunto de todas as amostras possíveis, S_t (Särndal *et al.*, 1992). A amostra é constituída pela selecção das unidades que a compõem através das probabilidades de inclusão de cada elemento da população. Estas probabilidades dependem da forma como os elementos da população são seleccionados, ou seja, dependem do plano de sondagem adoptado[10]. É, portanto, essencial determinar essas probabilidades de inclusão para se fazer inferência sobre parâmetros da população, a partir das observações realizadas junto das n_t unidades pertencentes à amostra.

Numa sondagem aleatória por conglomerados previamente estratificados, uma amostra s_t contém todos os elementos dos conglomerados que são seleccionados para a amostra do período t. Nesta situação, tem-se que a probabilidade do j-ésimo elemento da população pertencente ao conglomerado g ser seleccionado para a amostra s_t, denominada por probabilidade de inclusão de primeira ordem, é dada por:

$$\pi_{(t,\, j,\, g)} = P\left(j \in s_t\right) = P\left(g \in s_{Gt}\right) = \pi_{tg}\,. \qquad (2.3.1)$$

As probabilidades de inclusão de segunda ordem, ou seja, as probabilidades dos j-ésimo e k-ésimo elementos da população serem seleccionados simultaneamente para a amostra s_t, são dadas por:

[10] Para um dado plano de sondagem, $p(s)$, a probabilidade do j-ésimo elemento ser incluído na amostra s é dada por $\pi_j = \sum_{j \in s} p(s)$, e a probabilidade dos elementos j e k pertencerem simultaneamente à amostra s é dada por $\pi_{jk} = \sum_{j,k \in s} p(s)$.

$$\pi_{(t,\,j,\,g)(t,\,k,\,g)} = P\left(j \in s_t \wedge k \in s_t\right) = P\left(g \in s_{Gt}\right) = \pi_{tg} \tag{2.3.2}$$

se ambos os elementos populacionais *j* e *k* pertencem ao mesmo conglomerado *g*, e por:

$$\pi_{(t,\,j,\,g)(t,\,k,\,g^*)} = P\left(j \in s_t \wedge k \in s_t\right) = P\left(g \in s_{Gt} \wedge g^* \in s_{Gt}\right) = \pi_{tgg^*} \tag{2.3.3}$$

se os elementos *j* e *k* pertencem a diferentes conglomerados *g* e g^*, respectivamente.

No caso em que o plano de sondagem corresponde a um painel puro, no qual a amostra seleccionada no período 1, s_{G1}, é observada ao longo de *T* períodos de tempo, tem-se que $\pi_{tg} = \pi_g$ (t = 1, ..., T; g = 1, ..., G) e $\pi_{tgg^*} = \pi_{gg^*}$ (t = 1, ..., T; g, g^* = 1, ..., G; $g \neq g^*$).

Em seguida, apresentam-se as probabilidades de inclusão de primeira e de segunda ordem, no caso particular em que o plano de sondagem adoptado corresponde a um painel puro, no qual a amostra é seleccionada através de uma sondagem aleatória por conglomerados previamente estratificados[11]. Este é o plano de sondagem que está subjacente ao estudo empírico apresentado no sétimo capítulo. A probabilidade de um elemento *j* pertencente ao conglomerado *g* do estrato *h* ser seleccionado para a amostra global, é igual à probabilidade do conglomerado a que pertence esse elemento ser seleccionado para a amostra dos conglomerados do estrato *h*, s_{Gh}, ou seja,

$$\pi_{(j,g,h)} = \pi_{(g,h)} = \frac{a_h}{A_h}, \tag{2.3.4}$$

onde $\pi_{(j,g,h)} = P\left(j \in s\right)$ e $\pi_{(g,h)} = P\left(g \in s_{Gh}\right)$.

A probabilidade dos elementos *j* e *k*, pertencentes ao mesmo conglomerado *g* do estrato *h*, serem seleccionados para a amostra global, é igual à probabilidade do conglomerado a que pertencem esses elementos ser seleccionado para a amostra dos conglomerados do estrato *h*, s_{Gh}, ou seja,

$$\pi_{(j,g,h)(k,g,h)} = \pi_{(g,h)} = \frac{a_h}{A_h}, \tag{2.3.5}$$

[11] Supõe-se que as unidades primárias populacionais são seleccionadas sem reposição.

onde $\pi_{(j,g,h)(k,g,h)} = P(j \in s \wedge k \in s)$.

A probabilidade dos elementos *j* e *k*, pertencentes a diferentes conglomerados *g* e g^*, respectivamente, do estrato *h*, serem seleccionados para a amostra global, é igual à probabilidade dos conglomerados *g* e g^* serem seleccionados para a amostra dos conglomerados do estrato *h*, s_{Gh}, ou seja,

$$\pi_{(j,g,h)(k,g^*,h)} = \pi_{(g,h)(g^*,h)} = \frac{a_h(a_h - 1)}{A_h(A_h - 1)}, \tag{2.3.6}$$

onde $\pi_{(j,g,h)(k,g^*,h)} = P(j \in s \wedge k \in s)$ e $\pi_{(g,h)(g^*,h)} = P(g \in s_{Gh} \wedge g^* \in s_{Gh})$.

A probabilidade do elemento *j* pertencente ao conglomerado *g* do estrato *h* pertencer à amostra s_{Gh} retirada desse estrato, e do elemento *k* pertencente ao conglomerado g^* do estrato h^* pertencer à amostra s_{Gh^*} retirada do estrato h^*, é igual à probabilidade do conglomerado *g* ser seleccionado para a amostra s_{Gh} e do conglomerado g^* ser seleccionado para a amostra s_{Gh^*}, ou seja,

$$\pi_{(j,g,h)(k,g^*,h^*)} = \pi_{(g,h)(g^*,h^*)} = \frac{a_h}{A_h}\frac{a_{h^*}}{A_{h^*}}, \tag{2.3.7}$$

onde $\pi_{(j,g,h)(k,g^*,h^*)} = P(j \in s \wedge k \in s)$ e $\pi_{(g,h)(g^*,h^*)} = P(g \in s_{Gh} \wedge g^* \in s_{Gh^*})$.

2.3.3 Parâmetros e estimadores

Um parâmetro descritivo de uma população finita é uma função dos *N* valores da variável de interesse na população. Por sua vez, um estimador de um parâmetro da população é uma função dos dados da amostra que tem como objectivo a produção de valores aproximados a esse parâmetro desconhecido, designados por estimativas. Espera-se que na maior parte das amostras, essas estimativas estejam situadas numa vizinhança próxima do verdadeiro valor do parâmetro a estimar. Geralmente um parâmetro desconhecido da população, *U*, denota-se por θ, e um seu estimador denota-

se por $\hat{\theta}$. Da mesma forma, um parâmetro desconhecido num domínio *i* da população *U*, denota-se por θ_i, e um seu estimador denota-se por $\hat{\theta}_i$.

Tal como o parâmetro θ é função dos *N* valores que a variável de interesse, *Y*, pode assumir na população, isto é, $\theta = \theta(y_1, ..., y_N)$, também o parâmetro θ_i é função dos N_i valores que a variável de interesse, *Y*, pode assumir no *i*-ésimo domínio da população, isto é, $\theta_i = \theta_i(y_{i1}, ..., y_{iN_i})$.

Os parâmetros total, média e variância de uma variável, *Y*, num domínio *i* da população, que se denotam, respectivamente, por τ_i, μ_i e σ_i^2, *i*=1, ..., *m* (ou por τ_{yi}, μ_{yi} e σ_{yi}^2, caso se pretenda deixar explícito a que variável estas quantidades se referem), podem ser apresentados na seguinte forma:

$$\theta_i^1 = \tau_i = \sum_{j \in U_i} y_j \,, \tag{2.3.8}$$

$$\theta_i^2 = \mu_i = \frac{1}{N_i} \sum_{j \in U_i} y_j = \frac{\tau_i}{N_i}, \tag{2.3.9}$$

$$\theta_i^3 = \sigma_i^2 = \frac{1}{N_i - 1} \sum_{j \in U_i} (y_j - \mu_i)^2 \,. \tag{2.3.10}$$

A estimação de parâmetros de interesse em domínios pode ser efectuada através de diferentes tipos de estimadores, dependendo da dimensão amostral disponível nos domínios de estudo, bem como da informação auxiliar disponível. Os estimadores para domínios podem ser classificados em estimadores directos e em estimadores indirectos.

Os estimadores directos são baseados apenas nas observações da variável de interesse pertencentes ao domínio de estudo e ao período de tempo em análise. Este tipo de estimadores pode usar informação auxiliar conhecida de dentro ou de fora do domínio. Os estimadores directos são centrados ou aproximadamente centrados e poderão apresentar boa precisão, particularmente se utilizarem informação auxiliar, ou se os domínios de estudo tiverem dimensões amostrais razoáveis. No entanto, quando existe um grande número de domínios de pequena dimensão as variâncias aproximadas poderão ser elevadas e os enviesamentos significativos (Rao, 2003).

Por sua vez, os estimadores indirectos utilizam observações da variável de interesse, bem como de variáveis auxiliares, de fora do domínio de estudo e/ou do período de tempo considerado. Neste último tipo de estimadores, que é conhecido por "pedir informação emprestada" a outros domínios de forma a aumentar a dimensão "efectiva" da amostra do domínio de estudo, é ainda possível fazer uma subclassificação em estimadores directos modificados, estimadores sintéticos e estimadores combinados (compostos ou compósitos) (Rao, 2003).

Alguns estimadores indirectos que verificam propriedades estatísticas desejáveis do ponto de vista do plano da sondagem (como o não enviesamento aproximado ou a consistência) são denominados por muitos investigadores por estimadores directos modificados. Contudo, Coelho (2000) sublinha a sua preferência em designar estes estimadores por estimadores sintéticos corrigidos, porque considera que podem ser encarados como estimadores sintéticos aos quais é adicionado um factor de correcção do enviesamento. Os estimadores directos modificados são aproximadamente centrados no desenho e apresentam normalmente menores níveis de variância do que os estimadores directos alternativos.

Por sua vez, os estimadores indirectos cujas propriedades estatísticas dependem das hipóteses do modelo postulado são denominados por estimadores sintéticos. Segundo Gonzalez (1973), um estimador é designado por estimador sintético se for utilizado um estimador directo para um domínio de maior dimensão ou um conjunto de domínios que contenha o pequeno domínio de estudo, para derivar um estimador indirecto para esse pequeno domínio, assumindo que os pequenos domínios têm as mesmas características que os grandes domínios. O pressuposto implícito na estimação sintética é, portanto, o de que a relação existente entre a variável de interesse e as variáveis auxiliares seja idêntica no domínio de estudo e no domínio de maior dimensão ou no conjunto de domínios. Os estimadores sintéticos apresentam normalmente menores níveis de variância do que os estimadores directos alternativos, mas podem ser fortemente enviesados se as hipóteses subjacentes estiverem erradas (Coelho, 2000).

Por último, os estimadores combinados são estimadores que resultam de uma combinação (normalmente linear) entre um estimador directo (ou directo modificado) e um estimador sintético (Rao, 2003).

Pelo facto dos estimadores directos, directos modificados e sintéticos para domínios não constituírem o objecto central de investigação neste trabalho, remete-se o leitor para Särndal *et al.* (1992), Thompson (1992), Coelho (2000) e Rao (2003). Nestas referências são apresentados, de forma detalhada, todos esses estimadores para um plano de sondagem genérico e para os planos de sondagem mais utilizados, bem como discutidas as suas propriedades estatísticas. Aos estimadores combinados para domínios é dada especial atenção neste trabalho, sobretudo nos capítulos quarto e quinto.

Quando se pretende fazer estimação em domínios sobre parâmetros da população, θ_i, e se dispõe de vários estimadores alternativos, $\hat{\theta}_i$, é necessário escolher o estimador mais adequado, no sentido em que este deverá fornecer estimativas com o menor erro amostral possível. Para se avaliar a qualidade de um estimador $\hat{\theta}_i$ de θ_i, bem como para se fazer a comparação entre estimadores alternativos recorre-se, normalmente, às seguintes propriedades dos estimadores: enviesamento e precisão. Existe, contudo, outra propriedade dos estimadores que deve ser avaliada, sempre que possível, no âmbito da teoria das sondagens: a consistência. As supracitadas propriedades são apresentadas no subcapítulo 2.5.

2.4 TIPOS DE INFERÊNCIA

Existem duas grandes abordagens de estimação quando é efectuada amostragem sobre uma população finita: a abordagem *design-based* e a abordagem *model-based*.

Segundo Rao (2003), na estimação *design-based* o vector dos valores tomados pela variável de interesse, $\mathbf{y} = (y_1, ..., y_N)$, na população finita, *U*, é formado por quantidades fixas e as probabilidades de inclusão dadas pelo desenho da amostra são utilizadas para determinar valores esperados, variâncias, enviesamentos e outras propriedades dos estimadores. Assume-se então que a característica de interesse, *Y*, associada ao *j*-ésimo elemento da população pode ser medida exactamente através da observação desse elemento *j*, não existindo erros de medida. A componente probabilística é apenas introduzida pelo plano de sondagem utilizado, $p(s)$.

Na abordagem de estimação *design-based* pretende efectuar-se inferência sobre parâmetros descritivos da população finita (como por exemplo totais e médias), os quais são encarados como funções de $\mathbf{y} = (y_1, ..., y_N)$. Esta inferência é normalmente efectuada através de estimadores directos, sendo a sua variância geralmente obtida do ponto de vista do plano da sondagem. Desta forma, as probabilidades de inclusão são utilizadas não só na estimação dos parâmetros, mas também na avaliação das propriedades dos estimadores.

Segundo Särndal *et al.* (1992), na estimação *model-based* o vector dos valores tomados pela variável de interesse sobre a população finita, $\mathbf{y} = (y_1, ..., y_N)$, é considerado como uma realização de *N* variáveis aleatórias $\mathbf{Y} = (Y_1, ..., Y_N)$, cuja distribuição conjunta é especificada por um modelo ζ. Este tipo de modelo é também conhecido por modelo de superpopulação. No entanto, a observação dos valores da variável de interesse está igualmente limitada às unidades de uma amostra seleccionada a partir da população finita. Existe assim um primeiro nível de aleatoriedade introduzido pelas hipóteses do modelo ζ, e um segundo nível resultante do plano de sondagem adoptado, $p(s)$.

Na segunda abordagem de estimação pode ser efectuada inferência sobre os parâmetros da superpopulação ou sobre os parâmetros descritivos da população finita. Quando a inferência é feita sobre os parâmetros descritivos da população finita, normalmente através de estimadores indirectos, na estimação e na avaliação das propriedades dos estimadores são por vezes utilizados critérios que têm em conta não só o modelo postulado, mas também o plano de sondagem.

Coelho (2000) considera que as situações em que a inferência despreza o plano de sondagem $p(s)$ e se baseia apenas no modelo, são situações de inferência *model-based* pura. Contudo, para aquele autor, quando o estimador e a avaliação das suas propriedades se baseiam num critério que tem em conta o plano de sondagem $p(s)$, a abordagem pode ser designada por *model-assisted*[12]. Nesta abordagem, o modelo assiste a estimação, porém as propriedades estatísticas do estimador são avaliadas do ponto de vista do plano de sondagem. Coelho (2000) acrescenta ainda que nesta abordagem não

[12] Também se optou por manter na língua inglesa a designação *model-assisted*, por não se conhecer uma tradução oficial para a língua portuguesa desta expressão. Uma tradução possível poderá ser assistido pelo modelo.

se pressupõe que a população finita tenha sido efectivamente gerada pelo modelo ζ postulado, mas somente que esta pode ser aproximadamente descrita por esse modelo de superpopulação.

Segundo Coelho (2000), nas abordagens *model-based* e *model-assisted* a qualidade estatística da inferência fica dependente da verificação das hipóteses subjacentes aos modelos postulados, mas no último caso procuram-se habitualmente estimadores que mantenham algumas propriedades básicas *design-based* que não dependam desses pressupostos. Por esta razão, a selecção e a validação dos modelos têm um papel muito importante na estimação, e em particular na estimação *model-based*. Se os modelos postulados não se ajustarem bem aos dados, então os estimadores *model-based* serão enviesados no modelo, mesmo com amostras grandes, o que provoca consequências graves ao nível da estimação.

Na estimação em pequenos domínios, é frequentemente seguida uma abordagem *model-based* ou *model-assisted*, em oposição à abordagem *design-based* seguida na estimação em domínios de maiores dimensões amostrais. Segundo Rao (2003), a estimação em pequenos domínios segundo uma abordagem *model-based* ou *model-assisted* tem muitas vantagens, sendo os ganhos de precisão a vantagem mais importante. Outra das vantagens desta abordagem apontada por este autor, consiste na derivação de estimativas "óptimas"[13] dos parâmetros de interesse, bem como das medidas de variabilidade a elas associadas.

2.5 PROPRIEDADES ESTATÍSTICAS DOS ESTIMADORES NAS ABORDAGENS *DESIGN-BASED* E *MODEL-BASED*

No contexto das sondagens, as propriedades estatísticas mais frequentemente utilizadas para aferir a qualidade de um estimador, $\hat{\theta}$, são o enviesamento e a precisão. Estas propriedades, que geralmente são função do valor esperado e da variância dos estimadores, podem ser determinadas tendo por base os pressupostos do modelo ou o plano de sondagem adoptado. No primeiro caso, são designadas por valor esperado e

[13] No sentido em que são calculadas através de estimadores EBLUP.

variância no modelo (ou *model-based*), enquanto no segundo caso são designadas por valor esperado e variância no desenho (ou *design-based*).

Para se fazer a distinção entre estas duas abordagens, supõe-se que se está interessado em estimar um parâmetro populacional de uma característica de interesse, *Y*. Para tal, foi seleccionada uma amostra, *s*, de dimensão *n*, de acordo com um plano de sondagem específico, *p*(*s*), a partir de uma população de dimensão *N*. Supõe-se também que os dados amostrais $y_s = \{y_1,..., y_n\}$ foram recolhidos assumindo que não existem erros não amostrais. Admita-se que o problema de inferência consiste na obtenção de um estimador $\hat{\theta} = \hat{\theta}_s$, da sua variância, denotada por $V(\hat{\theta})$, e de um IC para o parâmetro θ ao nível de confiança $100(1-\alpha)\%$, a partir dos dados amostrais e assumindo que *n* é suficientemente grande para justificar a aproximação à distribuição Normal.

2.5.1 Abordagem *design-based*

Seja $\hat{\theta}$ um estimador *design-based* de θ. O enviesamento no desenho de um estimador $\hat{\theta}$, denotado por $B_d(\hat{\theta})$, é definido como a diferença entre o valor esperado no desenho do estimador $\hat{\theta}$ e o verdadeiro valor do parâmetro a estimar, θ, isto é:

$$B_d(\hat{\theta}) = E_d(\hat{\theta}) - \theta, \tag{2.5.1}$$

onde o valor esperado no desenho de $\hat{\theta}$ é definido em função da distribuição de probabilidade originada pelo plano de sondagem:

$$E_d(\hat{\theta}) = \sum_{s \in S} p(s)\hat{\theta}_s, \tag{2.5.2}$$

sendo $p(s)$ a probabilidade de seleccionar a amostra *s* e $\hat{\theta}_s$ a estimativa produzida por $\hat{\theta}$ a partir da amostra *s*. Uma propriedade desejável para um estimador $\hat{\theta}$ é que este seja capaz de produzir estimativas centradas do verdadeiro valor de θ, ou seja, que $B_d(\hat{\theta}) = 0$. Um estimador é, portanto, não enviesado no desenho quando $E_d(\hat{\theta}) = \theta$.

A precisão de um estimador $\hat{\theta}$ é uma medida da dispersão das estimativas produzidas por esse estimador em torno do verdadeiro parâmetro. A precisão de um estimador pode

ser avaliada através da sua variância, se $\hat{\theta}$ for centrado, ou através do seu erro quadrático médio (EQM), se $\hat{\theta}$ for enviesado[14]. A variância no desenho de um estimador $\hat{\theta}$, representada por $V_d\left(\hat{\theta}\right)$, mede o grau de dispersão das estimativas em torno da sua média e calcula-se como o valor esperado no desenho do quadrado da diferença entre o estimador $\hat{\theta}$ e a sua esperança matemática:

$$V_d\left(\hat{\theta}\right) = E_d\left[\hat{\theta} - E_d\left(\hat{\theta}\right)\right]^2 = \sum_{s \in S} p(s)\left[\hat{\theta}_s - E_d\left(\hat{\theta}\right)\right]^2 . \qquad (2.5.3)$$

Naturalmente que o desvio padrão de um estimador $\hat{\theta}$ (ou erro padrão, como é frequentemente denominado no contexto das sondagens) é dado pela raiz quadrada positiva da variância, $\sigma_{\hat{\theta}} = \sqrt{V_d\left(\hat{\theta}\right)}$, e o seu coeficiente de variação (CV) é dado pelo quociente entre o desvio padrão e o valor esperado do estimador, $CV\left(\hat{\theta}\right) = \sigma_{\hat{\theta}} / E_d\left(\hat{\theta}\right)$.

Um estimador da variância no desenho de $\hat{\theta}$, denotado por $\hat{V}_d\left(\hat{\theta}\right)$, admite-se poder ser obtido a partir dos dados amostrais. Segundo Särndal *et al.* (1992) e Rao (2003), na prática utiliza-se como indicador da precisão de um estimador $\hat{\theta}$ centrado ou quase centrado, uma estimativa (enviesada) do CV teórico, dada por $cv\left(\hat{\theta}\right) = \hat{\sigma}_{\hat{\theta}} / \hat{\theta}$. Desta forma, um estimador centrado de θ será tanto mais preciso no desenho quanto menor for o valor dessa estimativa do CV.

O EQM no desenho de um estimador $\hat{\theta}$, denotado por $EQM_d\left(\hat{\theta}\right)$, mede o grau de dispersão das estimativas em torno do verdadeiro valor de θ e calcula-se como o valor esperado no desenho do quadrado da diferença entre o estimador $\hat{\theta}$ e o verdadeiro valor do parâmetro θ:

[14] Em língua inglesa, existem os termos "*precision*" e "*accuracy*", que significam respectivamente precisão e exactidão de um estimador (SPE e ABE, 2007), sendo a precisão avaliada com base na variância e a exactidão com base no EQM (Kish, 1995). Contudo, em língua portuguesa não é habitual falar-se na "exactidão de um estimador". Desta forma, decidiu utilizar-se apenas a terminologia "precisão de um estimador" para ambos os casos. Naturalmente que se deve ter a noção que a variância não mede a precisão de um estimador enviesado, mas apenas o grau de dispersão das estimativas em torno da sua média.

$$EQM_d\left(\hat{\theta}\right) = E_d\left(\hat{\theta} - \theta\right)^2 = \sum_{s \in S} p(s)\left(\hat{\theta}_s - \theta\right)^2 . \qquad (2.5.4)$$

A partir de (2.5.4) deduz-se que $EQM_d\left(\hat{\theta}\right) = V_d\left(\hat{\theta}\right) + \left[B_d\left(\hat{\theta}\right)\right]^2$, donde se verifica que se $\hat{\theta}$ for um estimador centrado de θ, então o $EQM_d\left(\hat{\theta}\right) = V_d\left(\hat{\theta}\right)$.

Na abordagem *design-based*, assume-se que o plano de sondagem assegura probabilidades de inclusão positivas de primeira ordem, $\pi_j > 0$, $\forall j \in U$, e de segunda ordem, $\pi_{jk} > 0$, $\forall\, j \neq k \in U$. Nestas condições, esta abordagem de estimação permite que os estimadores de θ e da sua variância, $V_d\left(\hat{\theta}\right)$, sejam centrados no desenho.

Aquando da escolha do melhor estimador $\hat{\theta}$ para o mesmo parâmetro θ, normalmente escolhe-se aquele que é o mais eficiente, ou seja, aquele que possui o EQM mais pequeno. Contudo, pode ser perigoso utilizar o EQM como único critério de avaliação da qualidade de um estimador, pois corre-se o risco de se estar a aceitar um enviesamento elevado, desde que a redução na variância seja compensadora. Deste modo, é desejável que um estimador tenha não só um EQM pequeno, mas também um enviesamento pequeno (Särndal *et al*., 1992).

Um critério de escolha entre um estimador enviesado, $\hat{\theta}_1$, e um estimador centrado, $\hat{\theta}_2$, é dado pela comparação entre o EQM do primeiro estimador e a variância do segundo. Se o valor do EQM de $\hat{\theta}_1$ for inferior ao valor da variância de $\hat{\theta}_2$, isto é, $V_d\left(\hat{\theta}_1\right) + \left[B_d\left(\hat{\theta}_1\right)\right]^2 < V_d\left(\hat{\theta}_2\right)$, então o estimador enviesado, $\hat{\theta}_1$, é preferível ao estimador centrado, $\hat{\theta}_2$.

Segundo Särndal *et al.* (1992), nalgumas situações, a utilização de um estimador com um enviesamento moderado é preferível à utilização de um estimador centrado, pelos seguintes motivos: (*i*) muitos parâmetros têm uma estrutura formal que dificulta a determinação de um estimador centrado; e (*ii*) um estimador com um enviesamento moderado pode ter variância, e até mesmo EQM, inferior à variância de um estimador centrado. Contudo, deve evitar-se a utilização de estimadores que sejam consideravelmente enviesados, uma vez que nestas situações os rácios de

enviesamento[15] são geralmente elevados, o que conduz a afastamentos significativos na probabilidade de cobertura dos IC *design-based*.

Um IC *design-based* para θ a $100(1-\alpha)\%$, pode ser construído a partir de estimativas da variância no desenho de $\hat{\theta}$, sendo dado por:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{V}_d(\hat{\theta})}, \tag{2.5.5}$$

onde a probabilidade $(1-\alpha)$ é designada por grau ou coeficiente de confiança, e $P(Z \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) = 1-\alpha/2$ com $Z \sim N(0;1)$. Contudo, quando se trabalha com estimadores enviesados, os IC *design-based* baseados em estimativas da variância no desenho de $\hat{\theta}$ são geralmente inválidos devido aos seus elevados rácios de enviesamento. Segundo Särndal *et al.* (1992) e Rao (2003), diz-se que um IC *design-based* para θ é válido se num grande número de amostras da mesma dimensão, *s*, extraídas de uma população *U* segundo o mesmo plano de sondagem, cerca de $100(1-\alpha)\%$ dos intervalos contém o verdadeiro valor (fixo) do parâmetro θ. Por outro lado, diz-se que um IC *design-based* para θ é inválido se num grande número de amostras da mesma dimensão, *s*, extraídas de uma população *U* segundo o mesmo plano de sondagem, a percentagem de intervalos que contém o verdadeiro valor (fixo) do parâmetro θ for significativamente diferente de $100(1-\alpha)\%$.

De facto, quando $\hat{\theta}$ é um estimador enviesado para θ e é possível[16] dispor-se de uma estimativa do EQM no desenho de $\hat{\theta}$, $eqm_d(\hat{\theta})$, pode construir-se um IC *design-based* para θ a $100(1-\alpha)\%$ do tipo:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{eqm_d(\hat{\theta})}\,. \tag{2.5.6}$$

Segundo Coelho (2000), este tipo de IC apresenta uma probabilidade de cobertura que cresce com o rácio de enviesamento, sendo sempre maior ou igual ao nível de confiança

[15] O rácio de enviesamento é dado por $RB(\hat{\theta}) = B_d(\hat{\theta}) \Big/ \sqrt{V_d(\hat{\theta})}$.

[16] É normalmente impossível estimar o EQM *design-based* de um estimador enviesado $\hat{\theta}$, dada a habitual dificuldade em produzir estimativas adequadas do seu enviesamento.

definido *a priori*. Em particular, para rácios de enviesamento elevados (geralmente superiores a um), Coelho (2000) mostrou que as probabilidades de cobertura dos IC *design-based* construídos a partir de estimativas do EQM no desenho podem ser significativamente superiores a $100(1-\alpha)\%$, e que as probabilidades de cobertura dos IC *design-based* construídos a partir de estimativas da variância no desenho podem ser significativamente inferiores a esse nível de confiança definido *a priori*.

Segundo Coelho (2000), estas limitações motivam a procura de estimadores do parâmetro de interesse com enviesamentos moderados, de forma a assegurar que as variâncias constituam medidas de precisão com significado[17], bem como proporcionem probabilidades válidas para os IC baseados nessas variâncias. Esta motivação é fortalecida pela dificuldade de obtenção de estimativas do enviesamento dos estimadores, e consequentemente das respectivas estimativas do EQM no desenho.

Outra propriedade desejável para um estimador é a sua consistência. Contudo, a propriedade da consistência da teoria geral da inferência estatística clássica exige que se faça uma adaptação para o caso da teoria das sondagens, pois esta definição não pode ser aplicada directamente aos estimadores definidos sobre amostras de uma população finita. Se se pretender estimar um parâmetro sobre uma característica de uma população *U* de dimensão *N*, a partir das *n* observações amostrais da variável de interesse, com $n \leq N$, naturalmente que não é possível calcular um limite quando *n* tende para infinito. A referida adaptação a esta propriedade pode ser encontrada em Hansen *et al*. (1953). Outra adaptação da definição de consistência pode ser encontrada em Cochran (1977), que afirma que um estimador é consistente se, quando se iguala a dimensão amostral à dimensão populacional, a estimativa é exactamente igual ao verdadeiro valor do parâmetro na população. Särndal *et al.* (1992) sublinham ainda que um estimador consistente deve ser aplicado com alguma precaução, pois pode não ser satisfatório em amostras de pequena dimensão. Rao (2003) sintetiza que um estimador $\hat{\theta}$ de θ é um estimador consistente no desenho se for centrado no desenho (ou o seu enviesamento tende para zero com o aumento da dimensão amostral) e a sua variância no desenho tender para zero com o aumento da dimensão amostral. A consistência no desenho de

[17] Em particular, as referidas limitações motivam a procura de estimadores com pequenos rácios de enviesamento.

um estimador da variância do estimador do parâmetro de interesse, $\hat{V}_d\left(\hat{\theta}\right)$, é definida da mesma forma. Se o estimador $\hat{\theta}$ e o seu estimador da variância, $\hat{V}_d\left(\hat{\theta}\right)$, forem ambos consistentes no desenho, então a abordagem de estimação *design-based* fornece inferências válidas sobre *Y* independentemente dos valores da população, no sentido de que $t = \left(\hat{\theta} - \theta\right) \big/ \hat{\sigma}_{\hat{\theta}}$ converge em distribuição para uma variável $Z \sim N(0;1)$, à medida que a dimensão da amostra aumenta (Rao, 2003).

2.5.2 Abordagem *model-based*

Segundo Särndal *et al.* (1992), na abordagem *model-based*, os valores tomados pela variável de interesse sobre a população finita, y_j, considerados como uma realização de *N* variáveis aleatórias Y_j, são observados para os elementos $j \in s$, e não observados para os elementos $j \in (U - s)$, j=1, ..., N. Os registos dos dados amostrais são representados por $y_s = \{y_1, ..., y_n\}$. Utilizando as especificações do modelo ζ, a distribuição de $\left(\hat{\theta} - \theta\right)$ pode ser derivada para uma dada amostra *s*. Note-se que nesta abordagem de estimação, a inferência está vinculada a uma particular amostra *s* que foi observada, e não a quaisquer outras amostras que possam ser extraídas da população *U*.

Seja $\hat{\theta}$ um estimador *model-based* de θ. Diz-se que um estimador $\hat{\theta}$ é não enviesado no modelo se para uma dada amostra *s*,

$$E_m\left[\left(\hat{\theta} - \theta\right) | s\right] = 0. \tag{2.5.7}$$

A medida que avalia a variabilidade de $\hat{\theta}$ é o EQMP *model-based*, que é dado por:

$$EQMP_m\left(\hat{\theta}\right) = E_m\left[\left(\hat{\theta} - \theta\right)^2 | s\right]. \tag{2.5.8}$$

Neste caso denomina-se a medida de incerteza por EQMP porque na abordagem *model-based* o parâmetro θ é ele próprio uma variável aleatória. Um estimador do EQMP *model-based* é denotado por $eqmp_m\left(\hat{\theta}\right)$. Se o modelo apresentar uma boa qualidade de ajustamento aos dados, então esta medida pode ser considerada como uma medida realista da precisão obtida, podendo ser utilizada na construção de intervalos de

predição *model-based*. Um intervalo de predição *model-based* para θ a $100(1-\alpha)\%$, para uma particular amostra *s*, pode ser construído a partir do estimador $\hat{\theta}$ não enviesado no modelo e do $eqmp_m(\hat{\theta})$:

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{eqmp_m(\hat{\theta})}. \tag{2.5.9}$$

Ao contrário do IC (2.5.5), o intervalo de predição (2.5.9) não garante frequentemente a cobertura do verdadeiro valor do parâmetro desconhecido θ numa proporção $(1-\alpha)$, em amostras repetidas da mesma dimensão, *s*, extraídas segundo o mesmo plano de sondagem de uma população, independentemente da forma da população. Segundo Särndal *et al.* (1992), a interpretação do intervalo de predição (2.5.9) é a seguinte: dada uma amostra *s*, a diferença entre o estimador *model-based* e o parâmetro de interesse, $(\hat{\theta} - \theta)$, está contida no intervalo $\pm z_{\alpha/2} \sqrt{eqmp_m(\hat{\theta})}$ para aproximadamente $100(1-\alpha)\%$ de todos os vectores $\mathbf{y} = (y_1, ..., y_N)$ que podem ser gerados sob a especificação de um modelo ζ *N*-dimensional.

Por último, no âmbito da estimação *model-based* é ainda de realçar que a derivação do EQMP *model-based* dos estimadores dos parâmetros de interesse é feita a partir da teoria estatística subjacente ao modelo que suporta a estimação (*vide* secção 3.2.5 para o caso do modelo linear misto). Uma vez que os objectivos desta investigação se centram na estimação em pequenos domínios segundo uma abordagem *model-based* ou *model-assisted*, e que a avaliação das propriedades dos estimadores é naturalmente efectuada segundo esta abordagem, daqui em diante omite-se o índice inferior *m* no valor esperado e no EQMP, neste contexto de inferência.

2.6 ESTIMADORES COMBINADOS PARA DOMÍNIOS

Os estimadores combinados para domínios são estimadores que resultam de uma combinação (normalmente linear) entre um estimador directo (ou directo modificado) e um estimador sintético. Estes estimadores podem ser encarados como uma forma natural de equilibrar o enviesamento potencial do estimador sintético, representado por

$\hat{\theta}_i^{\sin}$, com a instabilidade do estimador directo, representado por $\hat{\theta}_i^{dir}$, procurando desta forma evitar que a qualidade do estimador fique totalmente dependente da veracidade do modelo postulado. Um estimador combinado para o parâmetro de interesse θ no domínio *i* pode ser escrito como (Rao, 2003):

$$\hat{\theta}_i^{com} = \gamma_i \hat{\theta}_i^{dir} + (1-\gamma_i)\hat{\theta}_i^{\sin}, \qquad (2.6.1)$$

onde γ_i é um ponderador adequado $(0 \le \gamma_i \le 1)$. Os estimadores combinados enquadram-se no tipo de inferência *model-based* (ou *model-assisted*), uma vez que não é assegurada a propriedade do não enviesamento do ponto de vista do plano da sondagem. De qualquer modo, as propriedades estatísticas dos estimadores combinados podem ser avaliadas do ponto de vista do plano da sondagem ou exclusivamente com base no modelo que se admite ter gerado a população finita.

Tal como acontece com os estimadores directos e com os outros tipos de estimadores indirectos, a produção de estimativas com base nos estimadores combinados tem subjacente a hipótese que as probabilidades de inclusão são positivas, $\pi_j > 0, \forall j \in U$. É, ainda, de salientar que a produção de uma estimativa para o domínio *i* com base num estimador combinado é possível mesmo quando $n_i = 0$. Neste caso, o estimador combinado reduz-se, na maior parte dos casos, a um estimador sintético.

Os estimadores combinados poderão ser classificados em três grandes tipos, de acordo com a forma de definição dos pesos, γ_i: pesos fixados à partida, pesos dependentes da dimensão da amostra e pesos dependentes dos dados. Em seguida é apresentada uma descrição sucinta de cada um destes tipos de estimadores.

2.6.1 Pesos fixados à partida

Uma abordagem simplista de definição dos pesos consiste em fixar os pesos à partida, de acordo com as expectativas ou convicções do investigador acerca das propriedades dos estimadores que fazem parte da combinação linear (2.6.1).

O EQM do estimador combinado (2.6.1), do ponto de vista *design-based*, é dado por (Coelho, 2000):

$$EQM_d\left(\hat{\theta}_i^{com}\right) = \gamma_i^2 EQM_d\left(\hat{\theta}_i^{dir}\right) + (1-\gamma_i)^2 EQM_d\left(\hat{\theta}_i^{\sin}\right) + 2\gamma_i(1-\gamma_i)Cov_d\left(\hat{\theta}_i^{dir}, \hat{\theta}_i^{\sin}\right)$$

$$\approx \gamma_i^2 EQM_d\left(\hat{\theta}_i^{dir}\right) + (1-\gamma_i)^2 EQM_d\left(\hat{\theta}_i^{\sin}\right), \qquad (2.6.2)$$

assumindo que o termo de covariância é nulo ou aproximadamente nulo.

Esta forma de definição dos pesos tem a desvantagem de não ter explicitamente em consideração a informação disponível, e em particular a fiabilidade observada do estimador directo e o enviesamento do estimador sintético. Dado que a dimensão da amostra nos domínios de estudo é uma variável aleatória, então será conveniente reflectir nos pesos, γ_i, tais dimensões.

2.6.2 Pesos dependentes da dimensão da amostra

Os estimadores combinados com pesos dependentes da dimensão da amostra são estimadores cujos pesos dependem geralmente das dimensões populacionais nos domínios, N_i e $\hat{N}_i$, embora também possam depender dos totais de uma variável auxiliar X nos domínios, $\tau_{x,i}$ e $\hat{\tau}_{x,i}$. Estes estimadores foram originalmente desenvolvidos para fazer estimação em domínios para os quais o valor esperado da dimensão amostral é suficientemente grande para tornar o estimador directo com precisão aceitável, mas a dimensão efectiva da amostra nesses domínios é inferior ao seu valor esperado (Drew *et al.*, 1982).

Drew *et al.* (1982) propuseram um estimador combinado com os seguintes pesos dependentes da dimensão da amostra:

$$\gamma_i = \begin{cases} 1 & \text{se } \hat{N}_i \geq \delta N_i \\ \hat{N}_i / \delta N_i & \text{se } \hat{N}_i < \delta N_i \end{cases}, \qquad (2.6.3)$$

onde $\hat{N}_i$ é um estimador directo da dimensão conhecida do domínio N_i e δ é escolhido subjectivamente para controlar a contribuição do estimador sintético[18].

[18] No *Canadian Labour Force Survey* é utilizado o estimador combinado com pesos dependentes dos dados com δ=2/3, para produzir estimativas ao nível das *census divisions* (Ghosh e Rao, 1994).

O princípio subjacente a este tipo de pesos é fácil de compreender mesmo intuitivamente e torna-se mais evidente quando se enquadra o problema no âmbito de uma sondagem aleatória simples. Considerando $\delta = 1$ neste caso, o peso γ_i é igual a 1 quando $n_i \geq E\left(n_i^s\right)$, onde n_i representa a dimensão efectiva da amostra no domínio de estudo. Isto significa que o estimador combinado com os pesos dependentes da dimensão da amostra se reduz ao estimador directo, se a dimensão efectiva da amostra no domínio de estudo não for menor do que o valor esperado da amostra nesse domínio, mesmo que ele seja muito pequeno. Neste caso, o estimador poderá não ser tão preciso quanto é desejável porque n_i é pequena. Contudo, este problema poderá ser resolvido através da consideração de um $\delta > 1$. De facto, será natural que o peso atribuído à componente directa do estimador seja tanto maior quanto maior for a dimensão efectiva da amostra. Por seu lado, será atribuído um peso considerável à componente sintética quando a dimensão da amostra no domínio é muito pequena relativamente ao seu valor esperado, $n_i << E\left(n_i^s\right)$. Este peso é gradualmente transferido para o estimador directo à medida que a referida dimensão amostral cresce, de tal forma que o estimador combinado seja consistente. Implicitamente considera-se que o valor esperado de n_i^s corresponde a uma dimensão amostral suficiente para produzir estimativas directas com um nível de precisão aceitável. Este raciocínio terá obviamente especial sentido quando se trabalha com domínios planeados, caso em que a sondagem pode ser planeada de forma a que as amostras nos domínios tenham dimensão com valor esperado adequado.

Drew *et al.* (1982) também propuseram um estimador combinado com pesos dados pela utilização de $\hat{\tau}_{x,i} / \tau_{x,i}$ em vez de $\hat{N}_i / N_i$, na expressão (2.6.3). Neste caso, estes autores usaram o estimador pós-estratificado pelo quociente como o estimador directo, $\hat{\theta}_i^{dir}$, e o estimador sintético pelo quociente como o estimador sintético, $\hat{\theta}_i^{\sin}$.

É, ainda, de referir que o estimador combinado com pesos dados por (2.6.3), adaptado ao caso da regressão, tem tido grande utilização prática, sendo habitual definir-se o parâmetro δ no intervalo $\left[2/3;3/2\right]$, dependendo obviamente do risco de enviesamento que se pretenda correr.

Särndal e Hidiroglou (1989) propuseram uma modificação aos pesos do estimador combinado, dados neste caso por:

$$\gamma_i = \begin{cases} 1 & \text{se } \hat{N}_i \geq N_i \\ \left(\hat{N}_i / N_i\right)^{\text{h-1}} & \text{se } \hat{N}_i < N_i \end{cases}, \qquad (2.6.4)$$

onde h é uma constante positiva escolhida subjectivamente. Estes autores sugeriram como valor de utilização geral $h = 2$, tendo considerado o estimador directo modificado pela regressão como o estimador directo, $\hat{\theta}_i^{dir}$, e o estimador sintético pela regressão como o estimador sintético, $\hat{\theta}_i^{\text{sin}}$.

Segundo Rao (2003), a estimação do EQM dos estimadores combinados com pesos dependentes dos dados apresenta alguma dificuldade. Por esse motivo, este autor sugere que se utilize a seguinte abordagem *ad hoc* proposta por Särndal e Hidiroglou (1989):

$$EQM_d\left(\hat{\theta}_i^{com}\right) \approx \gamma_i^2 EQM_d\left(\hat{\theta}_i^{dir}\right) + \left(1-\gamma_i\right)^2 EQM_d\left(\hat{\theta}_i^{\text{sin}}\right). \qquad (2.6.5)$$

As duas principais vantagens associadas aos estimadores combinados com pesos dependentes da dimensão da amostra são a sua simplicidade de aplicação e a possibilidade de utilização dos mesmos pesos para a estimação de parâmetros de várias variáveis de interesse. Desta forma, é possível garantir uma consistência interna na estimação que se torna manifestamente útil em situações em que esteja em causa a estimação de um grande número de variáveis, como é habitual na prática[19].

A principal desvantagem dos estimadores combinados com pesos dependentes da dimensão da amostra reside no facto dos pesos não reflectirem a informação dada pelas observações da variável de interesse, não tendo em consideração, em particular, nem a (boa) precisão do estimador directo, nem o enviesamento (eventualmente significativo) do estimador sintético. Note-se, ainda, que segundo esta metodologia de definição de pesos, os pesos são iguais para todas as variáveis de interesse, não reflectindo as suas diferenças no que respeita à variabilidade entre domínios de estudo. Tal é possível através de estimadores combinados com pesos dependentes dos dados.

[19] Na verdade, é possível conceber estimadores combinados com pesos dependentes dos dados, onde a consistência interna seja igualmente assegurada por via de procedimentos multivariados. Tais procedimentos não cabem no âmbito deste trabalho, mas podem ser encontrados em Fuller e Harter (1987) e Fay (1987).

2.6.3 Pesos dependentes dos dados

Os estimadores combinados com pesos dependentes dos dados são estimadores cujos pesos dependem da dimensão da variabilidade entre domínios de estudo, quando comparada com a variabilidade da variável de interesse dentro desses domínios. Desta forma, os pesos óptimos a atribuir a cada uma das componentes do estimador combinado dependem naturalmente dos seus EQM e da sua covariância.

Uma forma de determinação dos pesos, γ_i, consiste em minimizar o EQM do estimador combinado relativamente a γ_i. Assumindo que $Cov_d\left(\hat{\theta}_i^{dir}, \hat{\theta}_i^{\sin}\right)$ é zero ou é muito pequena quando comparada com $EQM_d\left(\hat{\theta}_i^{\sin}\right)$, tem-se que (Rao, 2003):

$$\gamma_i \approx \frac{EQM_d\left(\hat{\theta}_i^{\sin}\right)}{EQM_d\left(\hat{\theta}_i^{dir}\right) + EQM_d\left(\hat{\theta}_i^{\sin}\right)}. \tag{2.6.6}$$

Estes pesos são habitualmente desconhecidos, podendo ser derivados teoricamente com relativa facilidade. Assumindo que o estimador directo associado ao domínio *i* é não enviesado ou aproximadamente não enviesado no desenho, então um estimador para este peso é dado por:

$$\hat{\gamma}_i^1 = 1 - \frac{\hat{V}_d\left(\hat{\theta}_i^{dir}\right)}{\left(\hat{\theta}_i^{\sin} - \hat{\theta}_i^{dir}\right)^2}. \tag{2.6.7}$$

Contudo, este estimador pode ser muito instável: será de esperar que no caso em que a dimensão das amostras nos domínios é insuficiente para produzir boas estimativas directas, também não será adequada para produzir boas estimativas das variâncias e enviesamentos correspondentes. Esta dificuldade poderá ser ultrapassada através da utilização de estimativas médias de pesos para um conjunto de domínios, para um conjunto de variáveis ou para ambos.

Neste sentido, outra forma de determinação dos pesos, γ_i, consiste em minimizar a média dos EQM do estimador combinado para todos os domínios de estudo (ou para um grupo de domínios), relativamente a γ_i. Esta abordagem conduz a um peso óptimo comum para todos os domínios dado por (Purcell e Kish, 1980):

$$\gamma = \frac{\sum_{i=1}^{m} EQM_d\left(\hat{\theta}_i^{\sin}\right)}{\sum_{i=1}^{m}\left[EQM_d\left(\hat{\theta}_i^{dir}\right) + EQM_d\left(\hat{\theta}_i^{\sin}\right)\right]}. \tag{2.6.8}$$

Assumindo as mesmas hipóteses que no caso anterior, mas para todos os domínios de estudo, então um estimador para este peso é dado por:

$$\hat{\gamma}^2 = 1 - \frac{\sum_{i=1}^{m} \hat{V}_d\left(\hat{\theta}_i^{dir}\right)}{\sum_{i=1}^{m}\left(\hat{\theta}_i^{\sin} - \hat{\theta}_i^{dir}\right)^2}. \tag{2.6.9}$$

Este estimador é estável, mas o uso de um único estimador para todos os domínios de estudo não é, naturalmente, uma opção satisfatória, uma vez que as variâncias dos estimadores directos da variável de interesse podem variar significativamente de domínio para domínio.

Desta forma, uma outra alternativa possível de determinação dos pesos γ_i, passa pela modelação do enviesamento da parte sintética do estimador combinado, conduzindo à construção de estimativas indirectas para estes parâmetros. Com esse objectivo, têm vindo a ser desenvolvidos estimadores indirectos baseados em modelos explícitos, os quais envolvem efeitos aleatórios específicos de domínio de forma a contemplar a variabilidade existente entre domínios, não explicada pelas variáveis auxiliares do modelo.

Grande parte desses modelos de estimação em pequenos domínios pode ser encarada como um caso particular do modelo linear misto envolvendo efeitos fixos e aleatórios[20]. Sob esta abordagem, o parâmetro de interesse em cada domínio *i* envolve uma combinação linear desses efeitos, pelo que pode ser obtida através da metodologia EBLUP. Em particular, no terceiro capítulo são introduzidos os modelos lineares mistos, bem como a predição segundo a metodologia EBLUP utilizando este tipo de modelos.

[20] Alguns modelos de estimação em pequenos domínios podem ser ainda considerados casos particulares do modelo linear misto generalizado, tais como o modelo logit ou o modelo poisson-gama, entre outros, os quais não fazem parte dos objectivos deste trabalho de investigação.

2.7 MODELOS DE ESTIMAÇÃO EM PEQUENOS DOMÍNIOS

Segundo Rao (2003), os modelos que suportam a estimação em domínios podem ser classificados em modelos de ligação implícita (*implicit linking models*) e em modelos de ligação explícita (e*xplicit linking models*). Os modelos de ligação explícita permitem especificar da forma desejada a ligação entre o parâmetro de interesse e os diferentes tipos de informação disponível, bem como a incorporação da forma desejada de efeitos aleatórios específicos de domínio que expliquem a variabilidade inter-domínio, não explicada pelos efeitos fixos do modelo. Pelo contrário, nos modelos de ligação implícita não é possível especificar da forma desejada a ligação entre o parâmetro de interesse e os diferentes tipos de informação disponível, e assume-se que não existe variabilidade inter-domínios, para além da explicada pelos efeitos fixos do modelo. Na classe dos modelos de ligação implícita, que estabelecem a ligação entre pequenos domínios relacionados através de dados suplementares, encontram-se os estimadores indirectos tradicionais (estimadores directos modificados e sintéticos). Por outro lado, na classe dos modelos de ligação explícita para pequenos domínios, que tomam especificamente em consideração a variação inter-domínio, incluem-se os estimadores *model-based* puros e os estimadores *model-assisted*.

Segundo Rao (2003), as propriedades dos estimadores indirectos tradicionais baseados em modelos de ligação implícita são geralmente avaliadas do ponto de vista *design-based*, sendo as suas variâncias *design-based* normalmente inferiores às variâncias *design-based* dos estimadores directos. Todavia, esses estimadores indirectos são geralmente enviesados do ponto de vista do plano de sondagem, não decrescendo esse enviesamento com o aumento da dimensão total da amostra. Se o modelo implícito de ligação for bom, então o enviesamento do ponto de vista do plano de sondagem será pequeno, conduzindo a um EQM *design-based* significativamente mais pequeno do que o EQM do estimador directo. A redução do EQM é, portanto, a principal razão para a utilização dos estimadores indirectos na estimação em pequenos domínios.

Nos modelos de ligação explícita, a variabilidade inter-domínio é explicada, não só pelas variáveis auxiliares incluídas no modelo, mas também pelos efeitos aleatórios específicos de domínio. Neste tipo de modelos, define-se, portanto, a forma como a informação auxiliar é incorporada no processo de estimação. Por sua vez, os modelos de

ligação explícita podem ser classificados em dois grandes grupos[21]: os modelos de nível unidade (*unit level models*) e os modelos de nível área (*area level models*).

De acordo com Rao (2003), os modelos de nível unidade fazem a ligação entre os valores individuais da variável de interesse e os valores individuais das variáveis auxiliares, para cada um dos elementos da população. Se não for possível fazer tal ligação a este nível, são utilizados os modelos de nível área. Estes últimos modelos estabelecem a ligação entre um parâmetro[22] da variável de interesse e um parâmetro das variáveis auxiliares, para cada um dos pequenos domínios da população.

Ainda segundo Rao (2003), o uso de modelos de ligação explícita apresenta várias vantagens quando comparado com o uso de modelos de ligação implícita:

- permitem a utilização de medidas de diagnóstico da qualidade dos modelos, com o objectivo de encontrar os modelos que se ajustam melhor aos dados;
- permitem o cálculo de medidas de precisão específicas associadas às estimativas do parâmetro de interesse produzidas para cada pequeno domínio, em oposição às medidas de precisão globais frequentemente utilizadas nos estimadores sintéticos;
- permitem a tomada em consideração de modelos lineares mistos, assim como modelos lineares mistos generalizados (para tratar dados binários, dados de contagem, *etc*.);
- permitem a utilização de estruturas de dados mais complexas, como os dados de natureza temporal e/ou de natureza espacial;
- permitem a utilização dos desenvolvimentos mais recentes sobre modelos de efeitos aleatórios para realizar inferências mais precisas nos pequenos domínios.

[21] Na verdade, é também possível conceber modelos que compreendem simultaneamente variáveis auxiliares de nível unidade e variáveis auxiliares de nível área, designados por modelos de dois níveis (*two-level models*). Estes modelos não se inserem no âmbito deste estudo. Mais informações sobre estes modelos podem ser encontradas em Moura e Holt (1999) e em Rao (2003).

[22] O parâmetro que estabelece a ligação entre a variável de interesse e as variáveis auxiliares ao nível desagregado pode ser uma média populacional, μ_i, um total populacional, τ_i, ou qualquer função do tipo $\theta_i = g(\mu_i)$.

Apesar de existir uma grande variedade de modelos para estimação em pequenos domínios, é de realçar que devem ser os investigadores ou utilizadores finais a escolher o modelo a aplicar ao caso em estudo, e em especial, devem ser eles a escolher as variáveis auxiliares. O sucesso de qualquer método de estimação *model-based* ou *model-assisted* depende, sobretudo, da existência de boa informação auxiliar.

No contexto de estimação *model-based*, neste trabalho de investigação é dada atenção especial aos estimadores EBLUP derivados a partir de modelos de ligação explícita de nível área, e considerados como casos especiais do modelo linear misto.

Mais detalhes sobre os assuntos revistos neste capítulo podem ser encontrados, por exemplo, em Särndal *et al.* (1992) e em Rao (2003).

3. PREDIÇÃO EM MODELOS LINEARES MISTOS

3.1 INTRODUÇÃO

Nos últimos anos tem-se assistido a um interesse crescente pelo estudo de modelos lineares mistos, o qual tem sido acompanhado pelo alastrar da sua aplicação a diversas áreas do conhecimento, e em particular à área da estimação em pequenos domínios. Este capítulo foca-se na apresentação do modelo linear misto geral e de alguns dos seus casos particulares usados na estimação em pequenos domínios. Em particular, são apresentados o modelo linear misto (subcapítulo 3.2), o modelo *state space* linear geral (subcapítulo 3.3) e alguns modelos espaciais gerais (subcapítulo 3.4).

3.2 MODELO LINEAR MISTO GERAL

3.2.1 Introdução

Os modelos lineares mistos são largamente utilizados como suporte para a inferência acerca de parâmetros de interesse relativos a pequenos domínios. Enquanto a teoria acerca da predição dos efeitos mistos no contexto de modelos lineares mistos é amplamente conhecida, a literatura acerca da medição da incerteza do EBLUP está ainda dispersa. Neste subcapítulo são apresentados sucintamente os tipos de modelos lineares mistos, revistos os métodos de estimação de componentes de variância e expostos os principais resultados acerca da predição dos efeitos mistos. Na última secção deste subcapítulo é ainda efectuada uma revisão bibliográfica sobre a medição da

incerteza do EBLUP. A notação e terminologia introduzidas neste subcapítulo são iguais às utilizadas ao longo de todo o trabalho de investigação.

Mais detalhes sobre estes temas, bem como sobre o espectro da aplicação prática do modelo linear misto, podem ser encontrados, por exemplo, em Verbeke e Molenberghs (2000), em McCulloch e Searle (2001) e em Jiang (2007).

3.2.2 Tipos de modelos lineares mistos

A melhor forma de apresentar o modelo linear misto consiste em relembrar o conhecido modelo de regressão linear. Dispondo-se de um conjunto de *n* observações, o modelo de regressão linear pode ser apresentado como $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, onde $\mathbf{y}$ é um vector $n\times1$ de observações da variável dependente, $\mathbf{X}$ é uma matriz $n\times p$ de observações de variáveis independentes, $\boldsymbol{\beta}$ é um vector $p\times1$ de coeficientes de regressão desconhecidos e $\boldsymbol{\varepsilon}$ é um vector $n\times1$ de variáveis residuais (não observáveis). Neste modelo, os coeficientes de regressão são fixos. Contudo, existem casos nos quais faz sentido assumir que alguns desses coeficientes são aleatórios. Esses casos ocorrem tipicamente quando as observações da variável dependente estão autocorrelacionadas. Por exemplo, quando se dispõe de dados recolhidos ao longo do tempo sobre um mesmo conjunto de indivíduos, ou seja, dados longitudinais, é razoável assumir que as observações apresentam autocorrelação cronológica ou temporal. Da mesma forma, quando se dispõe de dados espaciais[23], é também razoável assumir que as observações apresentam dependência ou associação espacial. Estes tipos de correlações/associações podem ser introduzidos na estimação através da utilização de modelos lineares mistos particulares denominados, respectivamente, por modelos longitudinais e por modelos espaciais. Outra situação em que é prática habitual considerar-se um efeito como aleatório é quando não se dispõe de observações exaustivas de todos os possíveis níveis desse efeito, para o qual se deseja efectuar inferência. O modelo linear misto mais geral pode ser apresentado como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \tag{3.2.1}$$

onde $\mathbf{y}$ é um vector $n\times1$ de observações da variável dependente, $\mathbf{X}$ é uma matriz $n\times p$ de observações de variáveis independentes, $\boldsymbol{\beta}$ é um vector $p\times1$ de coeficientes de

[23] Os dados espaciais são introduzidos na secção 3.4.2.

regressão desconhecidos, os quais são frequentemente denominados efeitos fixos, $\mathbf{Z}$ é uma matriz $n \times h$ conhecida (matriz de desenho), $\boldsymbol{\upsilon}$ é um vector $h \times 1$ de efeitos aleatórios (não observáveis) e $\boldsymbol{\varepsilon}$ é um vector $n \times 1$ de variáveis residuais (também não observáveis). Quando comparado com o modelo de regressão linear, a diferença reside em $\mathbf{Z}\boldsymbol{\upsilon}$, a qual pode assumir um elevado número de formas diferentes, criando desta forma uma classe de modelos muito rica. As hipóteses básicas do modelo (3.2.1) são as seguintes: (*i*) os efeitos aleatórios têm média nula e variância finita; (*ii*) os resíduos têm média nula e variância finita; (*iii*) os efeitos aleatórios e os resíduos não são correlacionados. Em geral, as matrizes de variâncias-covariâncias[24] $\mathbf{G} = \mathbf{G}(\boldsymbol{\psi}) = V(\boldsymbol{\upsilon})$ e $\mathbf{R} = \mathbf{R}(\boldsymbol{\psi}) = V(\boldsymbol{\varepsilon})$ envolvem alguns parâmetros de dispersão desconhecidos denominados componentes de variância. O vector das componentes de variância é representado por $\boldsymbol{\psi} = (\psi_1, ..., \psi_q)'$. Assume-se que $\boldsymbol{\psi}$ pertence a um subconjunto do espaço euclidiano *q*-dimensional de tal forma que $\mathbf{V} = \mathbf{V}(\boldsymbol{\psi}) = V(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ é uma matriz não singular para todos os elementos de $\boldsymbol{\psi}$ pertencentes a esse subconjunto. Note-se que a matriz de covariâncias de $\mathbf{y}$ também depende de $\boldsymbol{\psi}$. Os desenvolvimentos históricos do modelo linear misto podem ser encontrados em McCulloch e Searle (2001).

Existem diferentes tipos de modelos lineares mistos, bem como diferentes formas de serem classificados. Uma forma de classificação é baseada na admissão, ou não, da hipótese da normalidade dos efeitos aleatórios e dos resíduos, comummente denominados por termos de erro do modelo. Se se assumir que os efeitos aleatórios e os resíduos estão normalmente distribuídos, então o modelo (3.2.1) é denominado modelo linear misto Gaussiano. Pelo contrário, quando não se assume que os efeitos aleatórios e os resíduos estão normalmente distribuídos, então o modelo (3.2.1) é denominado modelo linear misto não Gaussiano. Por este motivo, a distribuição conjunta pode não estar totalmente especificada para um conjunto de parâmetros no contexto do modelo linear misto não Gaussiano. Se por um lado a normalidade proporciona maior flexibilidade à modelação, por outro lado os modelos não Gaussianos são mais robustos à violação das hipóteses distribucionais (Jiang, 2007).

[24] Para simplificação da linguagem, denominam-se ao longo do texto as matrizes de "variâncias-covariâncias" simplesmente por matrizes de "covariâncias".

Um caso particular do modelo (3.2.1), frequentemente utilizado na área da estimação em pequenos domínios, é o seguinte modelo linear misto longitudinal:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\upsilon}_i + \boldsymbol{\varepsilon}_i \ , \tag{3.2.2}$$

onde $\mathbf{X}_i\ (n_i \times p)$ e $\mathbf{Z}_i\ (n_i \times b_i)$ são matrizes conhecidas, $\boldsymbol{\upsilon}_i$ e $\boldsymbol{\varepsilon}_i$ estão independentemente distribuídos com $\boldsymbol{\upsilon}_i \overset{ind}{\sim} (\mathbf{0}, \mathbf{G}_i)$ e $\boldsymbol{\varepsilon}_i \overset{ind}{\sim} (\mathbf{0}, \mathbf{R}_i)$, e *i* representa um dos *m* pequenos domínios (*i*=1, …, *m*). Também neste caso se assume que as matrizes de covariâncias $\mathbf{G}_i = \mathbf{G}_i(\boldsymbol{\psi})(b_i \times b_i)$ e $\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\psi})(n_i \times n_i)$ possam depender de um vector de componentes de variância, $\boldsymbol{\psi} = (\psi_1, ..., \psi_q)'$. Note-se que o modelo (3.2.2) pode ser reescrito tal como o modelo (3.2.1) utilizando a seguinte notação: $\mathbf{y} = col_{1\le i\le m}(\mathbf{y}_i)$, $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$, $\boldsymbol{\upsilon} = col_{1\le i\le m}(\boldsymbol{\upsilon}_i)$, $\boldsymbol{\varepsilon} = col_{1\le i\le m}(\boldsymbol{\varepsilon}_i)$, $\mathbf{Z} = diag_{1\le i\le m}(\mathbf{Z}_i)$, $\mathbf{G} = diag_{1\le i\le m}(\mathbf{G}_i)$, $\mathbf{R} = diag_{1\le i\le m}(\mathbf{R}_i)$, $n = \sum_{i=1}^{m} n_i$ e $b = \sum_{i=1}^{m} b_i$. Neste caso particular tem-se uma estrutura de covariâncias de $\mathbf{y}$ diagonal por blocos: $\mathbf{V} = diag_{1\le i\le m}(\mathbf{V}_i)$, onde $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i' + \mathbf{R}_i$.

3.2.3 Estimação de componentes de variância

A estimação das componentes de variância é um problema importante na análise de um modelo linear misto. Em algumas áreas, como por exemplo na genética quantitativa, o objectivo principal consiste em estimar as componentes de variância (Shaw, 1987). Em muitas outras áreas, a estimação das componentes de variância não constitui um dos objectivos principais, mas elas têm que ser estimadas de forma a avaliar a variabilidade dos estimadores de outras grandezas que fazem parte dos objectivos principais, como por exemplo os estimadores dos efeitos fixos e dos efeitos aleatórios. A análise de dados longitudinais (Diggle *et al.*, 1996) e a estimação em pequenos domínios incluem-se neste último caso (Jiang, 2007), razão pela qual a estimação das componentes de variância assume um papel crucial neste trabalho. Contudo, uma revisão detalhada de todos os métodos de estimação das componentes de variância, no contexto de um modelo linear misto, está fora do âmbito deste trabalho pelo facto de serem amplamente conhecidos. Por este motivo, em seguida são apenas enumerados esses métodos, remetendo-se o leitor para as referências bibliográficas apresentadas.

Alguns dos métodos mais antigos de estimação das componentes de variância, no contexto do modelo linear misto, não exigem que se assuma a hipóteses da normalidade. Neste grupo encontram-se o método de estimação análise de variância (ANOVA) que tem as suas origens com os trabalhos de Fisher (1922), e o método de estimação quadrática centrada de norma mínima[25] (MINQUE), proposto por C. R. Rao numa série de artigos (Rao, 1970, 1971, 1972). A ideia básica de ambos os métodos de estimação é originária do método dos momentos. Em particular, o método ANOVA baseia-se num procedimento que consiste em igualar somas de quadrados ao seu valor esperado, o que origina a possibilidade de existirem muitas formas de definir as somas de quadrados no caso de dados não balanceados. Os métodos de estimação I, II e III de C. R. Henderson (Henderson, 1953) são apenas três dos possíveis casos particulares do método ANOVA no contexto de dados não balanceados. Por sua vez, o método MINQUE procura determinar os estimadores centrados de variância mínima. Apesar destes métodos apresentarem a vantagem de não requererem a normalidade, eles apresentam, contudo, algumas desvantagens. Por um lado, o método ANOVA produz estimadores ineficientes para dados não balanceados e pode produzir estimativas não pertencentes ao espaço dos parâmetros (como por exemplo, estimativas negativas para a variância). Por outro lado, o método MINQUE depende dos valores iniciais das componentes de variância. Se o método MINQUE for executado iterativamente, utilizando em cada etapa os valores correntes para actualizar os valores iniciais, está-se perante um método denominado MINQUE iterativo. Este último método produz estimativas iguais às produzidas pelo método da máxima verosimilhança restrita (MVR) (Hocking e Kutner, 1975), e normalmente distribuídas sob amostras de grande dimensão (Brown, 1976).

No grupo dos métodos em que se assume a normalidade encontram-se o método da máxima verosimilhança (MV) e o método da MVR. Apesar do método da MV ter tido as suas origens com os trabalhos de Fisher no início do século passado (Fisher, 1922), em termos práticos este método foi pouco utilizado no contexto do modelo linear misto até 1967. A razão principal assentou no facto da estimação das componentes de variância pelo método da MV ser inicialmente difícil de implementar computacionalmente, ao contrário do método ANOVA. Para além desta dificuldade,

[25] Estimação Centrada Quadrática de Norma Mínima é a tradução para português de *Minimum Norm Quadratic Unbiased Estimation*. Decidiu usar-se nesta tese as siglas do método em língua inglesa (MINQUE), por assim ser sugerido no Glossário Estatístico Inglês-Português da SPE e ABE (2007).

existia também um problema relacionado com o comportamento assimptótico dos estimadores da MV, porque no contexto do modelo linear misto as observações estão normalmente correlacionadas, ao contrário do que se passa no caso tradicional em que se assume que as observações são independentes e identicamente distribuídas (*iid*). O método da MV passou a ser utilizado com maior frequência após Hartley e Rao (1967) terem abordado os problemas computacionais e assimptóticos do método. As propriedades assimptóticas dos estimadores da MV foram posteriormente estudadas por Anderson (1973) para o modelo marginal, bem como por Miller (1977) para um vasto conjunto de modelos. Miller (1977) mostrou que os estimadores da MV são consistentes e assimptoticamente normais. Os estimadores da MV das componentes de variância são eficientes, mas são, em geral, enviesados. Para além disso, o seu enviesamento não diminui com o aumento da dimensão amostral, se o número de efeitos fixos for proporcional à dimensão amostral (Jiang, 1996). De facto, neste último caso Neyman e Scott (1948) mostraram que os estimadores da MV são também inconsistentes.

Thompson (1962) propôs um método de estimação das componentes de variância que não exige a estimação simultânea dos efeitos fixos do modelo, os quais, em vários casos, não são considerados parâmetros de interesse, tal como acontece no âmbito da estimação em pequenos domínios. Este método, posteriormente estendido por Patterson e Thompson (1971), denomina-se por método da MVR. A ideia do método da MVR consiste em maximizar a parte da função de verosimilhança invariante aos efeitos fixos, ao contrário do método da MV que se baseia na maximização dessa função em relação a todos os parâmetros do modelo. Um vasto conjunto de autores tem vindo a argumentar que não existe perda de informação quando as componentes de variância são estimadas pelo método da MVR (por exemplo, Patterson e Thompson, 1971; Harville, 1977; Jiang, 1996). Várias derivações do método da MVR têm vindo a ser apresentadas por Harville (1974), Cooper e Thompson (1977), Barndorff-Nielsen (1983), Verbyla (1990), Heyde (1994) e Jiang (1996). Jiang (1996) mostrou que os estimadores da MVR são consistentes e assimptoticamente normais, mesmo que não se verifiquem as hipóteses de normalidade sob o modelo linear misto.

Existem vários artigos de revisão sobre estimação de componentes de variância nos quais se podem encontrar os métodos de verosimilhança (MV e MVR), nomeadamente de Harville (1977), Laird e Ware (1982), Jennrich e Schluchter (1986), Robinson

(1987), Cressie (1992) e Speed (1997). Nos trabalhos de Khuri e Sahai (1985) e Jiang (2007) de forma mais resumida, e na obra de Searle *et al.* (1992) de forma mais detalhada, são apresentados todos os métodos de estimação de componentes de variância no contexto de modelos lineares mistos. Na secção 4.2.3 é apresentada a estimação de uma componente de variância pelos métodos ANOVA, MV e MVR, sob o modelo de Fay-Herriot (Fay e Herriot, 1979), que é o mais conhecido modelo de estimação em pequenos domínios.

Os métodos de verosimilhança foram desenvolvidos sob a hipótese da normalidade, ou seja, sob o pressuposto de que os efeitos aleatórios e os resíduos estão normalmente distribuídos. Contudo, em inúmeras aplicações práticas o pressuposto da normalidade não é verificado. Por exemplo, Lange e Ryan (1989) apresentaram vários exemplos práticos que ilustram a não-normalidade dos efeitos aleatórios. Devido a esta situação, alguns investigadores têm vindo a seguir uma abordagem de quase-verosimilhança, a qual consiste em utilizar estimadores da MV e da MVR Gaussianos em situações de não-normalidade. Entre esses investigadores encontram-se Richardson e Welsh (1994), Heyde (1994, 1997) e Jiang (1996, 1997*b*), entre outros. Jiang (1996, 1997*b*) demonstrou a consistência e a normalidade assimptótica dos estimadores da MVR em situações de não-normalidade dos efeitos aleatórios, bem como apresentou condições necessárias e suficientes para idênticas propriedades assimptóticas dos estimadores da MV. Estes resultados garantem que a abordagem de quase-verosimilhança está bem justificada, pelo menos, do ponto de vista assimptótico.

Mas o problema da estimação das componentes de variância não se resume à determinação das suas estimativas pontuais. Deve também avaliar-se a variabilidade dos respectivos estimadores. Por conseguinte, a matriz de covariâncias assimptótica dos estimadores das componentes de variância é fundamental para várias inferências, nomeadamente a estimação por intervalos e o teste de hipóteses. Neste contexto, Jiang (1996, 1997*b*) deduziu a matriz de covariâncias assimptótica dos estimadores da MVR, bem como a matriz de covariâncias assimptótica dos estimadores da MV das componentes de variância. Segundo Jiang (1996, 1997*b*), estes dois resultados não exigem a verificação da hipótese da normalidade. Contudo, a matriz de covariâncias assimptótica numa situação de não normalidade envolve parâmetros adicionais para além das componentes de variância, como por exemplo os momentos de terceira e de

quarta ordem dos efeitos aleatórios, os quais não são estimados pelos procedimentos da MV e da MVR tradicionais. Para resolver esse problema da estimação da matriz de covariâncias assimptótica dos estimadores das componentes de variância segundo a abordagem quase-verosimilhança, Jiang (2005) propôs um método baseado na informação observada parcialmente, o qual é denominado por método da quase-informação observada parcialmente (*Partially Observed Quasi-Information Matrix*). Alternativamente, a referida matriz de covariâncias assimptótica também pode ser estimada através de um desenvolvimento do método *jackknife*[26], efectuado por Jiang *et al.* (2002) no contexto do modelo linear misto longitudinal aplicado à estimação em pequenos domínios.

3.2.4 Predição dos efeitos mistos

O problema da predição dos efeitos aleatórios[27], ou dos efeitos mistos num contexto mais geral, tem uma longa história que se iniciou em 1948 com o trabalho de C. R. Henderson no campo da reprodução animal (Jiang, 2007). A metodologia mais conhecida de predição dos efeitos mistos é a vulgarmente conhecida como melhor predição linear centrada (ou BLUP)[28], a qual foi publicada por Henderson em 1975 (Henderson, 1975).

Um efeito linear misto particular pode ser apresentado como $\eta = \mathbf{k}'\boldsymbol{\beta} + \mathbf{m}'\boldsymbol{\upsilon}$, onde $\mathbf{k}$ e $\mathbf{m}$ são vectores conhecidos de ordens $p\times 1$ e $h\times 1$ respectivamente, e $\boldsymbol{\beta}$ e $\boldsymbol{\upsilon}$ são os vectores dos efeitos fixos e aleatórios, respectivamente[29]. Se os efeitos fixos e as componentes de variância forem conhecidos, e admitindo a hipótese da normalidade,

[26] O método *jackknife* foi proposto por Quenouille (1949) e posteriormente desenvolvido por Tukey (1958).

[27] Quando se considera a existência de efeitos aleatórios presentes nos dados, eles são realizações de uma variável aleatória, mas não são observáveis. Todavia, utilizam-se os dados para propor valores numéricos, ou valores preditos para essas realizações. Esta é a razão pela qual se fala em predição (e não estimação) dos efeitos aleatórios.

[28] Neste contexto, a "melhor" significa que é a que tem EQM mínimo.

[29] No contexto da estimação em pequenos domínios, um efeito linear misto corresponde a um parâmetro de interesse de um determinado pequeno domínio.

então o melhor preditor[30] (BP) dos efeitos aleatórios, no sentido em que minimiza o EQM, é dado pelo valor esperado de $\boldsymbol{\upsilon}$ condicionado por **y**:

$$\breve{\boldsymbol{\upsilon}}_{BP} = E(\boldsymbol{\upsilon} \mid \mathbf{y}) = \mathbf{GZ'V^{-1}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.2.3}$$

Uma vez conhecido o melhor preditor dos efeitos aleatórios, tem-se que o melhor preditor dos efeitos mistos é dado por:

$$\breve{\eta}_{BP} = E(\eta \mid \mathbf{y}) = \mathbf{k'}\boldsymbol{\beta} + \mathbf{m'GZ'V^{-1}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.2.4}$$

Quando não se admite a hipótese da normalidade, (3.2.4) é ainda o melhor preditor linear de η no sentido em que minimiza o EQMP de um preditor, que é linear em **y** (Searle *et al.*, 1992).

Se os efeitos fixos forem desconhecidos, o que ocorre frequentemente nas aplicações práticas, mas se se continuar a admitir que as componentes de variância são conhecidas, então deve substituir-se $\boldsymbol{\beta}$ pelo seu melhor estimador linear centrado[31] (BLUE) em (3.2.4). O BLUE de $\boldsymbol{\beta}$, cuja derivação não exige a normalidade, é dado pelo estimador dos mínimos quadrados generalizados:

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}y}\,. \tag{3.2.5}$$

Desta forma, tem-se mesmo sem a exigência da hipótese da normalidade que o BLUP de η, é dado por:

$$\tilde{\eta}_{BLUP} = \tilde{\eta}^{H}(\boldsymbol{\psi}) = \mathbf{k'}\tilde{\boldsymbol{\beta}} + \mathbf{m'GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right), \tag{3.2.6}$$

onde $\tilde{\boldsymbol{\beta}}$ é o BLUE de $\boldsymbol{\beta}$ dado por (3.2.5). Neste contexto, o vector $\tilde{\boldsymbol{\upsilon}} = \mathbf{GZ'V^{-1}}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$ é também denominado por BLUP de $\boldsymbol{\upsilon}$. Frequentemente o BLUP de η é apresentado

[30] *Best Predictor* pode ser traduzido para português por melhor preditor. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (BP), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

[31] Melhor estimador linear centrado é a tradução para português de *Best Linear Unbiased Estimator* (SPE e ABE, 2007). Decidiu usar-se nesta tese as siglas da designação em língua inglesa (BLUE), por assim ser recomendado no Glossário Estatístico Inglês-Português da SPE e ABE (2007).

como $\tilde{\eta}_{BLUP} = \mathbf{k}'\tilde{\boldsymbol{\beta}} + \mathbf{b}'\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$, onde $\mathbf{b}' = \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$. O índice superior H em (3.2.6) refere-se a Henderson (Harville, 1990; Robinson, 1991).

Segundo Jiang (2007), a derivação original do BLUP foi apresentada por Henderson em 1950, na qual este autor propôs encontrar as "estimativas da máxima verosimilhança" dos efeitos fixos e dos efeitos aleatórios, considerando estes últimos como parâmetros fixos. Posteriormente, Henderson (1975) mostrou que depois de se substituir em (3.2.4) $\boldsymbol{\beta}$ pelo seu BLUE, o preditor resultante, $\tilde{\eta}$, é o melhor preditor linear centrado de η no sentido em que: (*i*) é linear em $\mathbf{y}$; (*ii*) o seu valor esperado é igual a η; e (*iii*) minimiza o EQM entre todos os preditores lineares centrados. Também neste caso o resultado não exige a verificação da hipótese da normalidade, ou seja, o BLUP está bem definido na classe dos modelos lineares mistos não-Gaussianos. Diferentes derivações do BLUP foram posteriormente apresentadas por Harville (1990) e por Jiang (1997*a*), entre outros. Veja-se também Robinson (1991), o qual apresenta um resumo alargado acerca do método BLUP, incluindo derivações alternativas, exemplos e aplicações.

A expressão (3.2.6) do BLUP envolve o vector das componentes de variância, $\boldsymbol{\psi}$, as quais são também normalmente desconhecidas nas aplicações práticas. Neste caso, deve substituir-se $\boldsymbol{\psi}$ por um estimador consistente, $\hat{\boldsymbol{\psi}}$, na expressão (3.2.6). O preditor resultante é normalmente denominado por BLUP empírico ou por EBLUP, e é representado por:

$$\hat{\eta}_{EBLUP} = \hat{\eta}^{H}\left(\hat{\boldsymbol{\psi}}\right) = \mathbf{k}'\hat{\boldsymbol{\beta}} + \mathbf{m}'\hat{\boldsymbol{\upsilon}}, \tag{3.2.7}$$

onde $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$ é o BLUE de $\boldsymbol{\beta}$, dado por (3.2.5), e $\hat{\boldsymbol{\upsilon}} = \hat{\boldsymbol{\upsilon}}(\hat{\boldsymbol{\psi}})$ é o BLUP de $\boldsymbol{\upsilon}$, dado por (3.2.3), nos quais $\boldsymbol{\psi}$ foi substituído por $\hat{\boldsymbol{\psi}}$. Kackar e Harville (1981) mostraram que se $\hat{\boldsymbol{\psi}}$ for um estimador ímpar e invariante a translações[32] e os dados forem normais, então o EBLUP permanece centrado. Segundo Jiang (2007), alguns dos mais conhecidos métodos de estimação de componentes de variância, incluindo o método ANOVA, MV e MVR, produzem estimadores ímpares e invariantes a translações.

[32] Um estimador $\hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\psi}}(\mathbf{y})$ é ímpar se $\hat{\boldsymbol{\psi}}(-\mathbf{y}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$ e é invariante a translações se $\hat{\boldsymbol{\psi}}(\mathbf{y} - \mathbf{X}\mathbf{h}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$, $\forall \mathbf{h} \in \Re^{p}$ e $\forall \mathbf{y}$.

Harville (1991) mostrou que o EBLUP é idêntico ao estimador de Bayes empírico[33] (EBP) quando considerado um modelo de efeitos aleatórios unidimensional. Este autor também observou que grande parte da investigação sobre estimadores EBP tem sido feita por estatísticos puros utilizando casos relativamente simples, como o modelo de efeitos aleatórios unidimensional. Por outro lado, grande parte da investigação sobre EBLUP tem sido feita por investigadores de outras áreas, mas que utilizam este tipo de ferramentas estatísticas. Uma das áreas onde a metodologia EBLUP tem sido extensivamente utilizada e estudada é a da estimação em pequenos domínios.

3.2.5 Medição da incerteza do *Empirical Best Linear Unbiased Predictor* (EBLUP)

Na secção anterior foi apresentado o melhor preditor linear centrado empírico dos efeitos mistos, o qual se obtém facilmente em qualquer modelo linear misto particular. Contudo, não basta conhecer o EBLUP. É necessário também conhecer a incerteza associada a esse EBLUP, uma vez que é indispensável saber qual é a precisão associada às estimativas dos parâmetros de interesse produzidas pelo EBLUP. Apesar do conhecimento do EQMP do EBLUP ter grande interesse nas aplicações práticas, em particular na área da estimação em pequenos domínios, a sua estimação origina um problema de difícil resolução na maioria dos casos práticos. Kackar e Harville (1984) mostraram que o EQMP do EBLUP, definido como $EQMP(\hat{\eta}) = E(\hat{\eta} - \eta)^2$, pode ser decomposto em três componentes:

$$EQMP(\hat{\eta}) = E(\tilde{\eta} - \eta)^2 + E(\hat{\eta} - \tilde{\eta})^2 + 2E[(\tilde{\eta} - \eta)(\hat{\eta} - \tilde{\eta})], \qquad (3.2.8)$$

onde $\tilde{\eta}$ é o BLUP de η dado por (3.2.6), $\hat{\eta}$ é o EBLUP de η dado por (3.2.7) e E representa o valor esperado relativo à distribuição conjunta de $\boldsymbol{y}$ e $\boldsymbol{\psi}$ subjacente ao modelo geral de estimação em pequenos domínios. Sob condições de normalidade de $\boldsymbol{\upsilon}$ e $\boldsymbol{\varepsilon}$, Kackar e Harville (1984) mostraram que $E[(\tilde{\eta} - \eta)(\hat{\eta} - \tilde{\eta})] = 0$, desde que as

[33] *Empirical Bayes Estimator* pode ser traduzido para português por estimador de Bayes empírico. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (EBP), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

componentes de variância sejam funções ímpares e invariantes a translações. Nesta situação,

$$\begin{aligned} EQMP(\hat{\eta}) &= E(\tilde{\eta}-\eta)^2 + E(\hat{\eta}-\tilde{\eta})^2 \\ &= EQMP(\tilde{\eta}) + E(\hat{\eta}-\tilde{\eta})^2 . \end{aligned} \tag{3.2.9}$$

Em (3.2.9) pode observar-se que o EQMP do BLUP é apenas uma parcela do EQMP do EBLUP, a qual pode ser decomposta em duas partes (Henderson, 1975) sem que para tal seja exigida a normalidade de $\boldsymbol{\upsilon}$ e de $\boldsymbol{\varepsilon}$:

$$EQMP(\tilde{\eta}) = E(\tilde{\eta}-\eta)^2 = g_1(\boldsymbol{\psi}) + g_2(\boldsymbol{\psi}), \tag{3.2.10}$$

onde as respectivas expressões analíticas são dadas por (Henderson, 1975):

$$g_1(\boldsymbol{\psi}) = \mathbf{m}'\left(\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\right)\mathbf{m}, \tag{3.2.11}$$

$$g_2(\boldsymbol{\psi}) = \left(\mathbf{k} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{k} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}\right). \tag{3.2.12}$$

A primeira parcela de (3.2.10), $g_1(\boldsymbol{\psi})$, avalia a variabilidade devida à estimação dos efeitos aleatórios e é de ordem $o(1)$, enquanto a segunda parcela, $g_2(\boldsymbol{\psi})$, avalia a variabilidade devida à estimação dos efeitos fixos do modelo e é de ordem $o(m^{-1})$. É importante salientar que as expressões de $g_1(\boldsymbol{\psi})$ e $g_2(\boldsymbol{\psi})$ não dependem do método utilizado para estimação das componentes de variância.

Um estimador simplista do EQMP do EBLUP, $eqmp_s(\hat{\eta})$, pode ser obtido através da substituição de $\boldsymbol{\psi}$ por $\hat{\boldsymbol{\psi}}$ na expressão do EQMP do BLUP, (3.2.10). Contudo, este estimador simplista subestima a verdadeira variabilidade do EBLUP por duas razões. Em primeiro lugar, porque não tem em consideração a variabilidade adicional devida à estimação das componentes de variância, representada pela segunda parcela do segundo membro de (3.2.9), a qual é de ordem $o(m^{-1})$. Em segundo lugar, porque o próprio estimador simplista do EQMP subestima a verdadeira variabilidade do EQMP do BLUP, sendo também essa subestimação de ordem $o(m^{-1})$. Segundo Prasad e Rao (1990) e Datta e Lahiri (2000), esta subestimação é mais acentuada nos casos em que $\boldsymbol{\psi}$ provoca uma variação significativa em $\tilde{\eta}$, assim como nos casos em que a variabilidade

de $\hat{\boldsymbol{\psi}}$ é elevada. Ainda de acordo com aqueles autores, o subenviesamento do estimador simplista do EQMP, de ordem $o\left(m^{-1}\right)$, é aproximadamente igual a duas vezes o estimador da variabilidade devida à estimação das componentes de variância.

Uma vez que o termo de (3.2.9) que contempla a variabilidade devida à estimação de $\boldsymbol{\psi}$ é geralmente intratável, torna-se fundamental obter uma aproximação para esse termo. Os métodos de estimação do EQMP do EBLUP que tomam em consideração essa variabilidade podem ser classificados em dois grupos: o método delta e os métodos por reamostragem. O método delta é baseado nos desenvolvimentos em série de Taylor, enquanto os métodos por reamostragem (métodos *jackknife* e *bootstrap*) são baseados nos princípios da amostragem repetida. Apesar dos métodos por reamostragem serem muito exigentes em termos computacionais, o que de resto é um problema cada vez menor devido ao progressivo avanço tecnológico a que se tem assistido, eles constituem uma solução quando algumas características da distribuição do estimador são exigidas, mas não estão disponíveis na sua forma explícita. Esta situação ocorre frequentemente quando o estimador não é linear nos valores da variável de interesse. Mesmo em alguns casos nos quais existem aproximações para grandes dimensões amostrais, os métodos *bootstrap* podem constituir alternativas mais precisas devido à sua correcção até à segunda ordem, o que não ocorre frequentemente nos métodos assimptóticos. Esta propriedade é referida por Efron e Tibshirani (1993) e provada por Hall (1992).

3.2.5.1 Método delta

O método delta foi proposto por Kackar e Harville (1984). Estes autores propuseram uma aproximação em série de Taylor para o EQMP do EBLUP, $EQMP\left(\hat{\eta}\right)$, para o caso do modelo linear misto Gaussiano (3.2.1), tendo em consideração a variabilidade introduzida pela estimação das componentes de variância, $\hat{\boldsymbol{\psi}}$. Apresentaram também um estimador do $EQMP\left(\hat{\eta}\right)$ baseada nessa aproximação. Kackar e Harville (1984) propuseram a seguinte aproximação heurística para o termo que contempla a variabilidade devida à estimação das componentes de variância:

$$E\left(\hat{\eta}-\tilde{\eta}\right)^2 \approx tr\left[\mathrm{K}\left(\boldsymbol{\psi}\right)E\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\right)\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\right)'\right], \qquad (3.2.13)$$

onde $\mathrm{K}(\boldsymbol{\psi})$ é a matriz de covariâncias de $\mathbf{d}(\boldsymbol{\psi})$, com $\mathbf{d}(\boldsymbol{\psi}) = \dfrac{\partial \tilde{\eta}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$. Contudo, esta aproximação tem sido considerada um pouco heurística por diversos investigadores, e para além disso não foi estudada a precisão da aproximação do EQMP do EBLUP nem do estimador do EQMP do EBLUP a ela associado. Num trabalho pioneiro, Prasad e Rao (1990) estudaram a precisão da aproximação de segunda ordem do $EQMP(\hat{\eta})$ para o caso do modelo linear misto longitudinal Gaussiano com componentes de variância estimadas pelo método ANOVA, o qual abrange o modelo de Fay-Herriot (Fay e Herriot, 1979), o modelo longitudinal de Rao-Yu (Rao e Yu, 1994) e os modelos de regressão com erros encaixados (utilizados, por exemplo, em Fuller e Battese (1973), Battese *et al.* (1988) e Datta e Ghosh (1991)). Na sua aproximação, foram desprezados os termos de ordem superior a $o(m^{-1})$, assumiu-se que $m \to \infty$ onde m é o número de pequenos domínios, e foram consideradas algumas condições de regularidade: (*i*) os elementos de $\mathbf{X}$ e $\mathbf{Z}$ são uniformemente limitados tal que $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = [O(m)]_{p\times p}$; (*ii*) n_i e b_i são finitos; (*iii*) os elementos de $\mathbf{G}$ e $\mathbf{R}$ são uniformemente limitados e diferenciáveis em relação a $\boldsymbol{\psi}$; e (*iv*) $\hat{\psi}_f = \mathbf{y}'\mathbf{C}_f\mathbf{y}$ é um estimador centrado e invariante a translações de ψ_f, onde $\mathbf{C}_f$ tem a forma $\mathbf{C}_f = diag_{1\le i\le m}[O(m^{-1})]_{n_i\times n_i} + [O(m^{-2})]_{m\times m}$. A aproximação de segunda ordem de Prasad-Rao para o termo que contempla a variabilidade devida à estimação das componentes de variância é dada por:

$$E(\hat{\eta} - \tilde{\eta})^2 \approx tr[\mathbf{L}(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}'(\boldsymbol{\psi})\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})] = g_3(\boldsymbol{\psi}), \tag{3.2.14}$$

onde $\mathbf{L}(\boldsymbol{\psi}) = \dfrac{\partial \mathbf{b}'}{\partial \boldsymbol{\psi}} = col_{1\le l\le q}\left(\dfrac{\partial \mathbf{b}'}{\partial \psi_l}\right)$, $\mathbf{b}' = \mathbf{b}'(\boldsymbol{\psi}) = \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$ e $\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})$ é a matriz de covariâncias assimptótica de $\hat{\boldsymbol{\psi}}$. Tem-se então que:

$$EQMP(\hat{\eta}) \approx g_1(\boldsymbol{\psi}) + g_2(\boldsymbol{\psi}) + g_3(\boldsymbol{\psi}), \tag{3.2.15}$$

onde $g_1(\boldsymbol{\psi})$ é dado por (3.2.11), $g_2(\boldsymbol{\psi})$ é dado por (3.2.12) e $g_3(\boldsymbol{\psi})$ é dado por (3.2.14). Para a mesma ordem da aproximação e utilizando estimativas das componentes de variância obtidas pelo método de Prasad-Rao (também conhecidas como estimativas ANOVA), Prasad e Rao (1990) propuseram o seguinte estimador para o $EQMP(\hat{\eta})$ sob um modelo linear misto com estrutura de covariância diagonal por blocos:

$$eqmp_{PR}(\hat{\eta}) = g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) + 2g_3(\hat{\boldsymbol{\psi}}), \qquad (3.2.16)$$

o qual é um estimador centrado até à segunda ordem[34]. No seguimento deste trabalho, Harville e Jeske (1992) propuseram também o estimador (3.2.16) para o $EQMP(\hat{\eta})$ sob um modelo linear misto geral (3.2.1). Posteriormente, Lahiri e Rao (1995) mostraram que o estimador do EQMP do EBLUP proposto por Prasad e Rao (1990) é robusto à violação da hipótese da normalidade dos efeitos aleatórios, quando aplicado ao modelo de Fay-Herriot e utilizando estimativas ANOVA das componentes de variância, desde que se verifique a simetria dos erros do modelo. No trabalho de Singh *et al.* (1998) foram utilizadas aproximações semelhantes às apresentadas por Prasad e Rao (1990), para deduzir o $EQMP(\hat{\eta})$ no âmbito do modelo de regressão com erros encaixados segundo as abordagens frequencista e Bayesiana. Estes autores concluíram que os métodos Bayesianos são inferiores aos métodos frequencistas no que se refere ao enviesamento e à variância das aproximações do $EQMP(\hat{\eta})$. Posteriormente, Datta e Lahiri (2000) voltaram a utilizar o mesmo modelo linear misto longitudinal Gaussiano com estrutura de covariância diagonal por blocos (tal como o modelo 3.2.2), mas apresentaram uma extensão dos resultados de Prasad-Rao de forma a cobrir um mais vasto conjunto de estimadores de componentes de variância, nomeadamente os estimadores da MV e da MVR. Na aproximação de Datta-Lahiri, foram igualmente desprezados os termos de ordem superior a $o(m^{-1})$, para $m \to \infty$, e foram assumidas as condições de regularidade (*i*)-(*ii*) de Prasad-Rao, adicionadas das seguintes condições: (*v*) $\mathbf{h} - \mathbf{X}'\mathbf{b} = [O(1)]_{p\times 1}$, $\forall \mathbf{h} \in \Re^p$; (*vi*) $\frac{\partial \mathbf{X}'\mathbf{b}}{\partial \psi_f} = [O(1)]_{p\times 1}$, f=1, …, q; (*vii*) $\mathbf{R}_i = \sum_{f=0}^{q} \psi_f \mathbf{D}_{if} \mathbf{D}_{if}'$ e $\mathbf{G}_i = \sum_{f=0}^{q} \psi_f \mathbf{F}_{if} \mathbf{F}_{if}'$, onde $\psi_0 = 1$, $\mathbf{D}_{if}$ e $\mathbf{F}_{if}$ são matrizes conhecidas de ordem $n_i \times n_i$ e $b_i \times b_i$, respectivamente, com elementos conhecidos uniformemente limitados tal que $\mathbf{R}_i$ e $\mathbf{G}_i$ sejam matrizes definidas positivas; e (*viii*) $\hat{\boldsymbol{\psi}}$ é o estimador de $\boldsymbol{\psi}$ que satisfaz as seguintes condições: $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} = o_p(m^{-1/2})$, $\hat{\boldsymbol{\psi}} - \tilde{\boldsymbol{\psi}} = o_p(m^{-1})$, $\hat{\boldsymbol{\psi}}(-\mathbf{y}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$, $\hat{\boldsymbol{\psi}}(\mathbf{y} - \mathbf{X}\mathbf{h}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$, $\forall \mathbf{h} \in \Re^p$ e $\forall \mathbf{y}$. Datta e Lahiri

[34] Diz-se que um estimador $eqmp$ é um estimador centrado até à segunda ordem do $EQMP$ do EBLUP quando $E(eqmp) = EQMP + o(m^{-1})$, ou seja, o enviesamento é de ordem $o(m^{-1})$.

(2000) notaram que quando as componentes de variância são estimadas pelos métodos de verosimilhança, então $\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}}_{MV}) \approx \overline{\mathbf{V}}(\hat{\boldsymbol{\psi}}_{MVR}) \approx m^{-1}\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})$, pelo que

$$g_3(\boldsymbol{\psi}) = m^{-1} tr\left[\mathbf{L}(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}'(\boldsymbol{\psi})\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})\right], \tag{3.2.17}$$

onde $\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})$ é a inversa da matriz de informação para $\boldsymbol{\psi}$. Datta e Lahiri (2000) propuseram o seguinte estimador, aproximadamente centrado até à segunda ordem, para o $EQMP(\hat{\eta})$:

$$eqmp_{DL}(\hat{\eta}) = g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) + 2g_3(\hat{\boldsymbol{\psi}}) - g_4(\hat{\boldsymbol{\psi}}), \tag{3.2.18}$$

onde $g_4(\hat{\boldsymbol{\psi}}) = \mathbf{c}'_{\hat{\boldsymbol{\psi}}}(\hat{\boldsymbol{\psi}})\nabla g_1(\hat{\boldsymbol{\psi}})$, sendo $\mathbf{c}_{\hat{\boldsymbol{\psi}}}(\hat{\boldsymbol{\psi}}) = E(\hat{\boldsymbol{\psi}}) - \boldsymbol{\psi}$ o enviesamento assimptótico de $\hat{\boldsymbol{\psi}}$ de ordem $o(m^{-1})$ e $\nabla g_1(\boldsymbol{\psi}) = col_{1\le l\le q}\left(\frac{\partial g_1(\boldsymbol{\psi})}{\partial \psi_l}\right)$ o gradiente de $g_1(\boldsymbol{\psi})$ em ψ_l, l=1, …, q. Datta e Lahiri (2000) mostraram ainda que $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MVR}}(\boldsymbol{\psi}) \approx \mathbf{0}$ e que $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MV}}(\boldsymbol{\psi}) \approx \frac{1}{2m}\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) col_{1\le l\le q}\left(tr\left\{\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta})\frac{\partial \mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})}{\partial \psi_l}\right\}\right)$, onde $\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ é a matriz de informação para $\boldsymbol{\beta}$. Note-se que quando as componentes de variância são estimadas pelos métodos de verosimilhança, tem-se assimptoticamente que $\overline{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{MVR}) \approx \overline{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{MV}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta})$. Verifica-se então que os estimadores (3.2.16) e (3.2.18) são iguais quando se utiliza o método da MV para estimar $\boldsymbol{\psi}$.

Das *et al.* (2004) também deduziram o estimador (3.2.18) para o $EQMP(\hat{\eta})$, tendo mostrado que é centrado até à segunda ordem, com componentes de variância estimadas pelos métodos de verosimilhança, mas para uma classe de modelos lineares mistos mais geral, cobrindo não só o modelo linear misto longitudinal (3.2.2), mas também o modelo ANOVA. Por exemplo, para modelos ANOVA mistos Gaussianos com estimação das componentes de variância pelos métodos de verosimilhança, Das *et al.* (2004) mostraram que:

$$g_3(\boldsymbol{\psi}) = tr\left[\mathbf{L}(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}'(\boldsymbol{\psi})\mathbf{H}^{-1}\right], \tag{3.2.19}$$

onde $\mathbf{H} = E\left(\frac{\partial^2 \mathbf{l}_R}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}\right)$ e $\mathbf{l}_R$ é o logaritmo da função de verosimilhança restrita. Adicionalmente, Das *et al.* (2004) também deduziram estimadores do $EQMP(\hat{\eta})$ baseados na sua aproximação para ambos os métodos de verosimilhança, os quais também são centrados até à segunda ordem. Os estimadores do EQMP do EBLUP do tipo (3.2.16) e (3.2.18) são normalmente denominados por estimadores ou aproximações analíticas. No capítulo quarto são apresentados estimadores analíticos do EQMP do EBLUP, no âmbito de todos os modelos de estimação em pequenos domínios revistos neste trabalho de investigação.

3.2.5.2 Método *jackknife*

Em alternativa ao método delta, Jiang *et al.* (1999, 2002) adaptaram o método *jackknife*, originalmente desenvolvido por Quenouille (1949) e posteriormente refinado por Tukey (1958), para o contexto da estimação em pequenos domínios, com o objectivo de estimar o EQMP do EBLUP. Neste contexto específico, o método *jackknife* consiste genericamente em retirar inicialmente a observação referente ao *e*-ésimo pequeno domínio, *e*=1, ..., *m*, em seguida estimar, para cada pequeno domínio excluído, os parâmetros do modelo com base no conjunto de dados remanescente e, por último, estimar as diferentes componentes do EQMP. O método *jackknife* adaptado por aqueles autores pode ser utilizado para uma classe geral de modelos mistos, incluindo os modelos lineares mistos e os modelos lineares mistos generalizados. Segundo Jiang *et al.* (2002), o método *jackknife* pode ser utilizado para estimar o $EQMP(\hat{\eta})$ no contexto dessa classe geral de modelos lineares mistos, utilizando estimadores–*M* gerais da componente de variância. Em particular, este método permite estimar o $EQMP(\hat{\eta})$ no contexto de modelos lineares mistos com componentes de variância estimadas pelos métodos ANOVA e de verosimilhança, mas também pode ser utilizado para estimar o EQMP do EBP sob modelos lineares mistos generalizados longitudinais com componentes de variância estimadas pelo método dos momentos.

Jiang *et al.* (2002) propuseram o seguinte estimador *jackknife* para o EQMP do EBLUP, aproximadamente não enviesado até à segunda ordem, para o caso de modelos lineares mistos longitudinais:

$$eqmp_{JLW}^{J}(\hat{\eta}) = g_1(\hat{\boldsymbol{\psi}}) - \frac{m-1}{m}\sum_{e=1}^{m}\left[g_1(\hat{\boldsymbol{\psi}}_{-e}) - g_1(\hat{\boldsymbol{\psi}})\right] + \frac{m-1}{m}\sum_{e=1}^{m}(\hat{\eta}_{-e} - \hat{\eta})^2 , \qquad (3.2.20)$$

onde $g_1(\boldsymbol{\psi})$ é dado por (3.2.11) e $\hat{\boldsymbol{\psi}}_{-e}$ e $\hat{\eta}_{-e}$ são estimados depois de eliminada a *e*-ésima observação do conjunto de dados. O método *jackknife* apresenta algumas vantagens, nomeadamente o facto de poder ser utilizado também em casos de não linearidade dos estimadores e de permitir construir intervalos de predição para além da estimação pontual do EQMP. Porém, este método é limitado apenas aos modelos paramétricos. O estimador (3.2.20) pode ainda ser decomposto em duas parcelas: $\hat{M}_1 = g_1(\hat{\boldsymbol{\psi}}) - \frac{m-1}{m}\sum_{e=1}^{m}\left[g_1(\hat{\boldsymbol{\psi}}_{-e}) - g_1(\hat{\boldsymbol{\psi}})\right]$ e $\hat{M}_2 = \frac{m-1}{m}\sum_{e=1}^{m}(\hat{\eta}_{-e} - \hat{\eta})^2$. A primeira parcela, $\hat{M}_1$, estima o EQMP do EBLUP quando os hiper-parâmetros do modelo, $\boldsymbol{\beta}$ e $\boldsymbol{\psi}$, são conhecidos. Defina-se o vector de hiper-parâmetros do modelo como $\boldsymbol{\delta} = \left[\boldsymbol{\beta}' \quad \boldsymbol{\psi}'\right]'$. Por sua vez, a segunda parcela, $\hat{M}_2$, estima a variabilidade adicional do EQMP devido à estimação desses hiper-parâmetros do modelo, $\boldsymbol{\delta}$.

Numa discussão do trabalho de Jiang-Lahiri-Wan, Bell (2001) observou que o estimador *jackknife* original apresenta ainda a desvantagem de poder assumir valores negativos. Contudo, Chen e Lahiri (2002, 2003) consideraram posteriormente que esse não é um problema grave, podendo ser facilmente resolvido. Também para o caso de modelos lineares mistos longitudinais, Chen e Lahiri (2002) propuseram o seguinte estimador *jackknife* ponderado para o EQMP do EBLUP:

$$\begin{aligned} eqmp_{CL}^{WJ}(\hat{\eta}) = {} & g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) - \frac{m-1}{m}\sum_{e=1}^{m}\omega_e\left[g_1(\hat{\boldsymbol{\psi}}_{-e}) + g_2(\hat{\boldsymbol{\psi}}_{-e}) - g_1(\hat{\boldsymbol{\psi}}) - g_2(\hat{\boldsymbol{\psi}})\right] + \\ & + \frac{m-1}{m}\sum_{e=1}^{m}\omega_e(\hat{\eta}_{-e} - \hat{\eta})^2 \end{aligned} , \qquad (3.2.21)$$

onde $g_1(\boldsymbol{\psi})$ e $g_2(\boldsymbol{\psi})$ são dados por (3.2.11) e (3.2.12) respectivamente, $\hat{\boldsymbol{\psi}}_{-e}$ e $\hat{\eta}_{-e}$ são estimados depois de eliminada a *e*-ésima observação do conjunto de dados e ω_e são pesos que devem satisfazer a seguinte condição $\omega_e = 1 + o(m^{-1})$. Chen e Lahiri (2002) observaram que existem várias possibilidades para a escolha dos pesos, tendo sugerido para o caso do modelo de Fay-Herriot que $\omega_e = 1 - \mathbf{x}_e'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_e$, onde $\mathbf{x}_e$ é um vector de ordem $p\times 1$ que contém a *e*-ésima observação. Para resolver o problema do estimador

(3.2.21) poder ainda assumir valores negativos, Chen e Lahiri (2003) propuseram a seguinte aproximação para a correcção de enviesamento:

$$\frac{m-1}{m}\sum_{e=1}^{m}\omega_e\left[g_1(\hat{\boldsymbol{\psi}}_{-e})+g_2(\hat{\boldsymbol{\psi}}_{-e})-g_1(\hat{\boldsymbol{\psi}})-g_2(\hat{\boldsymbol{\psi}})\right]\approx \hat{\mathbf{c}}'_{WJ}(\hat{\boldsymbol{\psi}})\nabla g_1(\hat{\boldsymbol{\psi}})- \\ -tr\left[\mathbf{L}(\hat{\boldsymbol{\psi}})\mathbf{V}(\hat{\boldsymbol{\psi}})\mathbf{L}'(\hat{\boldsymbol{\psi}})\hat{\boldsymbol{\upsilon}}_{WJ}\right], \tag{3.2.22}$$

onde $\hat{\mathbf{c}}_{WJ}=\sum_{e=1}^{m}\omega_e(\hat{\boldsymbol{\psi}}_{-e}-\hat{\boldsymbol{\psi}})$ é um estimador *jackknife* ponderado do enviesamento de $\hat{\boldsymbol{\psi}}$ e $\hat{\boldsymbol{\upsilon}}_{WJ}=\sum_{e=1}^{m}\omega_e(\hat{\boldsymbol{\psi}}_{-e}-\hat{\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}}_{-e}-\hat{\boldsymbol{\psi}})'$ é um estimador *jackknife* ponderado da matriz de covariâncias de $\hat{\boldsymbol{\psi}}$. Posteriormente, Chen e Lahiri (2005, 2008) apresentaram uma aproximação em série de Taylor do estimador *jackknife* ponderado do $EQMP(\hat{\eta})$ para o caso de modelos lineares mistos longitudinais. Segundo estes autores, essa aproximação apresenta ganhos ao nível computacional e tem uma expressão explícita dada por:

$$eqmp_{CL}^{AWJ}(\hat{\eta})=g_1(\hat{\boldsymbol{\psi}})+g_2(\hat{\boldsymbol{\psi}})-\hat{\mathbf{c}}'_{WJ}(\hat{\boldsymbol{\psi}})\nabla g_1(\hat{\boldsymbol{\psi}})+tr\left[\mathbf{L}(\hat{\boldsymbol{\psi}})\mathbf{V}(\hat{\boldsymbol{\psi}})\mathbf{L}'(\hat{\boldsymbol{\psi}})\hat{\boldsymbol{\upsilon}}_{WJ}\right]+ \\ +tr\left\{\mathbf{L}(\hat{\boldsymbol{\psi}})\left[\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right]\left[\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right]'\mathbf{L}'(\hat{\boldsymbol{\psi}})\hat{\boldsymbol{\upsilon}}_{WJ}\right\}. \tag{3.2.23}$$

Uma utilização do método *jackknife* pode ser encontrada em Maiti (2004). Este autor utilizou o estimador (3.2.20) para estimar o EQMP do EBLUP no contexto do modelo de Fay-Herriot. Na secção 4.2.5 são apresentados os estimadores (3.2.20) e (3.2.23), sob o modelo de Fay-Herriot (Fay e Herriot, 1979).

3.2.5.3 Método *bootstrap*

Outra metodologia por reamostragem que tem vindo a ser utilizada para estimar o EQMP do EBLUP é a técnica *bootstrap*, originalmente proposta por Efron (1979) e posteriormente desenvolvida por Efron e Tibshirani (1993). No contexto da estimação em pequenos domínios, o método *bootstrap* paramétrico consiste genericamente em gerar parametricamente um grande número de amostras *bootstrap* a partir do modelo ajustado aos dados originais, em seguida reestimar os parâmetros do modelo para cada amostra *bootstrap* e, por último, estimar as diferentes componentes do EQMP. A origem da utilização de métodos *bootstrap* para resolver o tipo de problemas em

epígrafe é atribuída a Laird e Louis (1987). Estes autores propuseram um método *bootstrap* para medir a incerteza de um EBP para um caso particular do modelo de Fay-Herriot (Fay e Herriot, 1979). O método proposto por Laird e Louis (1987) tem vindo a ser desenvolvido e adaptado para resolver um vasto conjunto de problemas. Veja-se, por exemplo, Arora *et al*. (1997), Booth e Hobert (1998) e Butar e Lahiri (2003). É, no entanto, de salientar que o principal desenvolvimento do método foi efectuado por Butar e Lahiri (2003), os quais propuseram um método *bootstrap* paramétrico suficientemente geral para estimar o EQMP do EBLUP no caso de modelos lineares mistos longitudinais. O método *bootstrap* desenvolvido por aqueles autores é aplicável, não só a diferentes metodologias de estimação de componentes de variância (ANOVA e métodos de verosimilhança), mas também a modelos Gaussianos e não Gaussianos (Lahiri e Maiti, 2002). Butar e Lahiri (2003), apresentaram também propriedades assimptóticas do método quando o número de domínios é elevado e a dimensão amostral de cada domínio é limitada. O estimador paramétrico *bootstrap* proposto por Butar e Lahiri (2003) é o seguinte:

$$eqmp_{BOOT}(\hat{\eta}) = g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) - E^*\left[g_1(\hat{\boldsymbol{\psi}}^*) + g_2(\hat{\boldsymbol{\psi}}^*) - g_1(\hat{\boldsymbol{\psi}}) - g_2(\hat{\boldsymbol{\psi}})\right] + E^*\left[\hat{\eta}(\hat{\boldsymbol{\psi}}^*) - \hat{\eta}(\hat{\boldsymbol{\psi}})\right]^2, \tag{3.2.24}$$

onde E^* é o valor esperado relativo ao modelo (3.2.25), o qual é uma réplica do modelo (3.2.1) Gaussiano, e o cálculo de $\hat{\boldsymbol{\psi}}^*$ é efectuado pelo mesmo método que o cálculo de $\hat{\boldsymbol{\psi}}$, com excepção do facto de ser baseado em $\mathbf{y}^*$ em vez de $\mathbf{y}$. O modelo utilizado é o seguinte:

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}^*, \tag{3.2.25}$$

onde $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$ e $\hat{\boldsymbol{\psi}}$ são estimados a partir dos dados iniciais, $\boldsymbol{\upsilon}^*$ e $\boldsymbol{\varepsilon}^*$ são gerados de forma independente a partir de $\boldsymbol{\upsilon}^* \overset{ind}{\sim} N\left(\mathbf{0}, \hat{\mathbf{G}}\right)$ e $\boldsymbol{\varepsilon}^* \overset{ind}{\sim} N\left(\mathbf{0}, \hat{\mathbf{R}}\right)$, respectivamente, com $\hat{\mathbf{G}} = \mathbf{G}(\hat{\boldsymbol{\psi}})$ e $\hat{\mathbf{R}} = \mathbf{R}(\hat{\boldsymbol{\psi}})$. De acordo com Butar e Lahiri (2003), o estimador (3.2.24) é centrado, sob determinadas condições de regularidade. Posteriormente, Pfeffermann e Glickman (2004), Pfeffermann e Tiller (2005) e Hall e Maiti (2006*a*, 2006*b*) apresentaram novos desenvolvimentos sobre este assunto. Pfeffermann e Glickman (2004) desenvolveram métodos *bootstrap* paramétricos e não paramétricos no contexto

do modelo de Fay-Herriot, que não exigem a geração de amostras a partir de uma determinada distribuição. Contudo, a hipótese da normalidade continua a ser implicitamente assumida nesses casos. Pfeffermann e Tiller (2005) também propuseram métodos *bootstrap* paramétricos e não paramétricos para predição do EQMP em modelos do tipo *state space*. Recentemente, Hall e Maiti (2006*a*, 2006*b*) introduziram algoritmos *bootstrap* duplos. Na secção 4.2.5 é apresentado o estimador (3.2.24), sob o modelo de Fay-Herriot (Fay e Herriot, 1979).

3.3 MODELO *STATE SPACE* LINEAR GERAL

3.3.1 Introdução

Os modelos do tipo *state space* são esporadicamente utilizados como suporte para inferência acerca de parâmetros de interesse relativos a pequenos domínios. Uma das suas principais utilizações neste contexto deve-se a Singh *et al*. (2005) que utilizaram um modelo espaciotemporal do tipo *state space*. Apesar da utilização de modelos do tipo *state space* estar fora dos objectivos deste trabalho, considera-se importante introduzir uma breve revisão sobre o modelo *state space* geral linear e filtro de Kalman por três razões: (*i*) por ser um caso particular do modelo linear misto; (*ii*) por ser menos conhecido do que o modelo linear misto geral; e (*iii*) pelo facto da literatura sobre este tema se encontrar um pouco dispersa.

3.3.2 Modelo *state space* linear e filtro de Kalman

Os modelos do tipo *state space* foram propostos por vários autores no âmbito das séries cronológicas (*e.g.* Harvey, 1989), mas têm as suas origens no controlo de engenharia. A ideia que está subjacente a este tipo de modelos de representação de sistemas lineares é a captação da dinâmica de um vector de variáveis observáveis em função de um vector de variáveis não observáveis, conhecido por vector de estado (*state vector*) do sistema. Supõe-se então que existe uma série de *m* observações, $\mathbf{y}_t = col_{1\le i\le m}(y_{ti})$, que podem ser definidas em termos de *r* variáveis de estado (*state variables*) não observáveis,

$\boldsymbol{\alpha}_t = col_{1\le j\le r}(\boldsymbol{\alpha}_{tj})$. Presume-se que estas variáveis devem descrever o estado corrente do sistema em questão. O modelo *state space* linear geral[35] é definido à custa de dois conjuntos de equações lineares. O primeiro conjunto de equações descreve a evolução do sistema, sendo denominado por "equação de transição":

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{A}\boldsymbol{\eta}_t, \tag{3.3.1}$$

onde $\mathbf{T}$ é a matriz $r \times r$ de transição (de constantes), $\mathbf{A}$ é uma matriz $r \times m$ de constantes e $\boldsymbol{\eta}_t = col_{1\le i\le m}(\eta_{ti})$ é um vector de erros *m*-dimensional. O segundo conjunto de equações descreve a relação entre o estado do sistema e as observações, sendo denominado por "equação de medição":

$$\mathbf{y}_t = \mathbf{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \tag{3.3.2}$$

onde $\mathbf{Z}$ é uma matriz $m \times r$ de coeficientes conhecida e $\boldsymbol{\varepsilon}_t = col_{1\le i\le m}(\varepsilon_{ti})$ é o vector dos erros de medida. Assume-se que $E(\boldsymbol{\varepsilon}_t \boldsymbol{\eta}_{t-j}) = \mathbf{0}$ para todo *t* e *j*, ou seja, os erros $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são não correlacionados contemporaneamente nem cronologicamente; $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}$, $E(\boldsymbol{\eta}_t) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \mathbf{R}_t$ e $E(\boldsymbol{\eta}_t \boldsymbol{\eta}_t') = \mathbf{Q}$.

Segundo Rao (2003), o modelo *state space* definido por (3.3.1) e (3.3.2) é um caso especial do modelo linear misto, permitindo essa sua forma a actualização das estimativas ao longo do tempo através das equações do filtro de Kalman apresentadas em seguida, e o alisamento de estimativas passadas à medida que novos dados ficam disponíveis. A estimação dos parâmetros do modelo *state space* pode ser efectuada através de um filtro de Kalman[36]. Os ingredientes do filtro de Kalman são as equações de predição (*predicting equations*) e as equações de actualização (*updating equations*). Em primeiro lugar são apresentadas as equações de predição.

[35] O modelo *state space* linear geral cobre os modelos de nível unidade e os modelos de nível área.

[36] O filtro de Kalman é um algoritmo recursivo criado por Kalman em 1960 que serve para produzir estimativas dos parâmetros de um modelo do tipo *state space* de séries cronológicas. Os modelos do tipo *state space* (lineares e não lineares) e o filtro de Kalman encontram-se mais desenvolvidos em Harvey (1989).

Seja $\tilde{\boldsymbol{\alpha}}_{t-1}$ o BLUP de $\boldsymbol{\alpha}_{t-1}$ baseado nos dados observados até ao período $t-1$. A melhor conjectura para $\boldsymbol{\alpha}_t$ baseada em toda a informação disponível até ao período $t-1$, é dada pelo BLUP de $\boldsymbol{\alpha}_t$ no período $t-1$:

$$\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathbf{T}\tilde{\boldsymbol{\alpha}}_{t-1}. \tag{3.3.3}$$

Da mesma forma, a melhor conjectura para $\mathbf{y}_t$ baseada em toda a informação disponível até ao período $t-1$, é dada por:

$$\tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}\tilde{\boldsymbol{\alpha}}_{t|t-1}. \tag{3.3.4}$$

A partir da expressão anterior, observa-se que a utilização do filtro de Kalman para estimar os parâmetros de um modelo *state space* permite fazer previsões para a variável de interesse (ou para os parâmetros de interesse) no período *t*, a partir dos valores observados da série até ao período $t-1$. A matriz de covariâncias dos erros de previsão $\left(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}\right)$, é igual a

$$\boldsymbol{\Sigma}_{t|t-1} = \mathbf{T}\boldsymbol{\Sigma}_{t-1}\mathbf{T}' + \mathbf{A}\mathbf{Q}\mathbf{A}', \tag{3.3.5}$$

onde $\boldsymbol{\Sigma}_{t-1} = E\left[\left(\boldsymbol{\alpha}_{t-1} - \tilde{\boldsymbol{\alpha}}_{t-1}\right)\left(\boldsymbol{\alpha}_{t-1} - \tilde{\boldsymbol{\alpha}}_{t-1}\right)'\right]$ é a matriz de covariâncias dos erros de previsão no período $t-1$, $\left(\boldsymbol{\alpha}_{t-1} - \tilde{\boldsymbol{\alpha}}_{t-1}\right)$. A matriz de covariâncias dos erros de previsão $\left(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}\right)$ é dada por

$$\mathbf{H}_t = \mathbf{R}_t + \mathbf{Z}\boldsymbol{\Sigma}_{t|t-1}\mathbf{Z}'. \tag{3.3.6}$$

No período *t*, a conjectura para $\boldsymbol{\alpha}_t$ e a matriz de covariâncias dos respectivos erros são actualizadas com base em nova informação, $\mathbf{y}_t$. A partir da definição do modelo *state space* e das equações de predição, obtém-se

$$\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}\left(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}\right) + \boldsymbol{\varepsilon}_t, \tag{3.3.7}$$

que é um caso especial do modelo linear misto (3.2.1), onde $\mathbf{y} = \mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}$ é o vector das "inovações", $\mathbf{X}\boldsymbol{\beta}$ está ausente, $\mathbf{Z} = \mathbf{Z}$, $\mathbf{v} = \boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{t|t-1}$, $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_t$; e $V(\boldsymbol{\varepsilon}) = \mathbf{R}_t$, $V(\mathbf{v}) = \boldsymbol{\Sigma}_{t|t-1}$ e

$V(\mathbf{y}) = \mathbf{H}_t$. Por conseguinte, o BLUP de $\mathbf{v}$ e a matriz de covariâncias dos erros de previsão $(\tilde{\mathbf{v}} - \mathbf{v})$ reduzem-se, respectivamente, às seguintes duas equações de actualização:

$$\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbf{Z}'\mathbf{H}_t^{-1}\left(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}\right), \tag{3.3.8}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{Z}'\mathbf{H}_t^{-1}\mathbf{Z}\boldsymbol{\Sigma}_{t|t-1}. \tag{3.3.9}$$

A expressão $\left(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}\right)$ da equação de actualização (3.3.8) contém a nova informação obtida através de $\mathbf{y}_t$, que é utilizada para actualizar as estimativas de $\boldsymbol{\alpha}_t$ produzidas com a informação disponível até ao período $t-1$, $\tilde{\boldsymbol{\alpha}}_{t|t-1}$. Por este motivo, o termo $\boldsymbol{\Sigma}_{t|t-1}\mathbf{Z}'\mathbf{H}_t^{-1}\left(\mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}\right)$ da equação (3.3.8) é denominado por "ganho" de Kalman. Para se implementar os cálculos recursivos (3.3.8) e (3.3.9), é necessário especificar-se[37] a média, $\tilde{\boldsymbol{\alpha}}_0$, e a matriz de covariâncias, $\boldsymbol{\Sigma}_0$, do vector de estado inicial, $\boldsymbol{\alpha}_0$. Quando a equação de transição, (3.3.1), não é estacionária, então pode fazer-se $\tilde{\boldsymbol{\alpha}}_0 = 0$ e $\boldsymbol{\Sigma}_0 = \kappa\mathbf{I}$, onde κ é uma constante positiva grande. Pelo contrário, quando essa equação é estacionária, então especifica-se $\tilde{\boldsymbol{\alpha}}_0$ e $\boldsymbol{\Sigma}_0$ como a média e a matriz de covariâncias incondicional de $\boldsymbol{\alpha}_t$, respectivamente.

Assumindo que as componentes de variância são conhecidas, o BLUP de $\mathbf{y}_t$, baseado em todas as estimativas até ao período *t* é dado por:

$$\tilde{\mathrm{y}}_t = \tilde{\mathrm{y}}_t^H(\boldsymbol{\psi}) = \mathbf{Z}\tilde{\boldsymbol{\alpha}}_t \tag{3.3.10}$$

onde $\tilde{\boldsymbol{\alpha}}_t$ é obtido através da equação do filtro de Kalman (3.3.8). O BLUP (3.3.10) envolve parâmetros desconhecidos, $\boldsymbol{\psi}$, que especificam as matrizes de covariâncias $\mathbf{R}_t$ e $\mathbf{Q}$ e eventualmente elementos desconhecidos na matriz de transição, que têm que ser estimados de forma consistente. Assumindo que os erros $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são normalmente distribuídos, esses parâmetros podem ser estimados através do método da MV. Quando

[37] Uma discussão detalhada sobre este tema pode ser encontrada em Harvey (1989).

se substituem em $\tilde{\boldsymbol{\alpha}}_t$ os parâmetros desconhecidos, $\boldsymbol{\psi}$, pelos seus estimadores consistentes, $\hat{\boldsymbol{\psi}}$, obtém-se o EBLUP de $\mathbf{y}_t$, dado por:

$$\hat{\mathrm{y}}_t = \hat{\mathrm{y}}_t^H(\hat{\boldsymbol{\psi}}) = \mathbf{Z}\hat{\boldsymbol{\alpha}}_t . \tag{3.3.11}$$

Segundo Rao (2003), ignorando a variabilidade associada a $\hat{\boldsymbol{\psi}}$, a matriz de covariâncias simplista dos erros de previsão no momento *t*, $(\boldsymbol{\alpha}_t - \hat{\boldsymbol{\alpha}}_t)$, é obtida quando se substitui $\boldsymbol{\psi}$ por $\hat{\boldsymbol{\psi}}$ na fórmula (3.3.9). Desta forma, um estimador simplista do EQMP do EBLUP é dado por:

$$eqmp_S(\hat{\mathrm{y}}_t) = \mathbf{Z}\hat{\boldsymbol{\Sigma}}_t\mathbf{Z}', \tag{3.3.12}$$

onde $\hat{\boldsymbol{\Sigma}}_t$ é um estimador de $\boldsymbol{\Sigma}_t$. Um estimador mais realista do EQMP do EBLUP, considerando a variabilidade associada a $\hat{\boldsymbol{\psi}}$, pode ser obtido a partir do trabalho de Ansley e Kohn (1986) sobre o desenvolvimento em série de Taylor do EBLUP em torno de $\boldsymbol{\psi}$ sob modelos *state space*. Neste caso, um estimador do EQMP do EBLUP é dado por:

$$eqmp(\hat{\mathrm{y}}_t) = \mathbf{Z}\hat{\boldsymbol{\Sigma}}_t\mathbf{Z}' + \left[\frac{\partial\tilde{\mathbf{y}}_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right]'_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \mathbf{I}^{-1}_{\hat{\boldsymbol{\psi}}}(\hat{\boldsymbol{\psi}})\left[\frac{\partial\tilde{\mathbf{y}}_t(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right]_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}, \tag{3.3.13}$$

onde $\mathbf{I}_{\hat{\boldsymbol{\psi}}}(\hat{\boldsymbol{\psi}})$ é a matriz de informação avaliada no ponto $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$.

3.4 MODELOS ESPACIAIS GERAIS

3.4.1 Introdução

A maior parte dos estimadores combinados para pequenos domínios não exploram a eventual correlação espacial existente entre domínios, em termos de semelhanças devidas à proximidade geográfica. Sendo evidente a influência que a localização espacial exerce sobre o comportamento da maior parte das variáveis de índole económico, as técnicas econométricas mais recentes começaram a tratar

convenientemente o espaço, ou seja a localização espacial das observações, enquanto factor fornecedor de informação. Desta forma, a informação fornecida pelas observações localizadas na vizinhança geográfica da observação que se pretende explicar é crucial para, em primeiro lugar, alcançar um modelo estatístico congruente e, em segundo lugar, explicar convenientemente a influência exercida pelos diversos factores sobre a variável explicada.

Com o objectivo de se ter em consideração na estimação a correlação existente entre efeitos aleatórios de domínios vizinhos, podem ser utilizados modelos espaciais no âmbito da estimação em pequenos domínios. Assim, neste subcapítulo é efectuada uma síntese dos modelos espaciais gerais, designadamente para dados referentes a áreas, que têm sido utilizados nesse âmbito. Para tal, é necessário fazer-se uma introdução sumária aos diferentes tipos de dados espaciais. Mais detalhes sobre este tema podem ser encontrados em Cliff e Ord (1981), em Cressie (1993) e em Carvalho e Natário (2008).

3.4.2 Dados espaciais e modelos espaciais

Seja $\mathbf{Y}(\mathbf{r})$ um vector aleatório cujo argumento, $\mathbf{r}$, representa uma localização genérica pertencente a um espaço Euclidiano *d*-dimensional, $\Re^d$ (normalmente d=2). O vector $\mathbf{r} = (r_1,...,r_d)'$ contém informação sobre a localização genérica dos dados, como por exemplo a longitude e a latitude, e nessa localização $\mathbf{r}$ o valor observado do vector aleatório é dado por $\mathbf{y}(\mathbf{r})$. As componentes básicas do processo são as localizações espaciais $\{\mathbf{r}_1,..., \mathbf{r}_n\}$ e os dados observados nessas localizações $\{\mathbf{y}(\mathbf{r}_1),..., \mathbf{y}(\mathbf{r}_n)\}$. O processo estocástico espacial geral pode ser apresentado como

$$\{\mathbf{Y}(\mathbf{r}) : \mathbf{r} \in D\}, \tag{3.4.1}$$

onde o conjunto índice $D \subseteq \Re^d$, que se assume ser aleatório, representa uma região espacial de área não nula (Cressie, 1993).

A flexibilidade oferecida por este processo espacial permite-lhe trabalhar com um vasto conjunto de problemas. Na realidade, este processo é adequado para trabalhar com dados discretos ou contínuos, com dados agregados espacialmente ou com observações

de um ponto no espaço, com localizações regulares ou irregulares, assim como com localizações provenientes de um espaço contínuo ou de um conjunto discreto.

Segundo Cressie (1993), os modelos espaciais podem ser diferenciados pelo tipo de dados que utilizam: (*i*) dados referentes a pontos (*geostatistical data*), (*ii*) dados referentes a processos pontuais (*point patterns*), ou (*iii*) dados referentes a áreas (*lattice data*).

Os dados referentes a pontos, ou dados contínuos no espaço, consistem em medidas de uma variável de interesse num conjunto finito de locais $\{\mathbf{r}_1,...,\mathbf{r}_n\}$, pertencentes a uma região de área não nula $D \subseteq \Re^d$. Alguns exemplos de dados referentes a pontos são os valores da temperatura observados em estações meteorológicas, o teor de um poluente no solo de uma região ou uma cota topográfica.

Os dados referentes a processos pontuais podem ser encarados como conjuntos de localizações aleatórias correspondentes às ocorrências de um certo acontecimento de interesse numa região fixa de $\Re^d$, designada por janela. Cada conjunto de localizações, denominada por configuração, serve de base ao processo pontual usado na modelação. Os dados referentes a processos pontuais ocorrem quando a variável de interesse é a localização de eventos, como por exemplo a concentração de árvores numa floresta.

Os dados referentes a áreas, ou dados agrupados em áreas, referem-se a observações (tipicamente somas ou médias) de uma variável de interesse em áreas (regulares ou irregulares) que constituem uma partição de uma região limitada *D*. Alguns exemplos de dados referentes a áreas são o número total de portadores de uma determinada doença num país, o número total de casos de morte por cancro do estômago numa determinada NUTSIII ou o preço médio de transacção da habitação num determinado concelho. As áreas regulares evocam a ideia de pontos regularmente espaçados em $\Re^d$, ligados aos seus vizinhos mais próximos, aos segundos vizinhos mais próximos, e assim sucessivamente. Um conjunto de dados referentes a áreas regulares é muito parecido a uma série temporal observada em períodos do tempo equidistantes. Pelo contrário, as áreas irregulares não seguem um padrão previsível de pontos espaciais. Para além disso, não são evidentes as ligações entre os pontos de $\Re^d$ a partir da geometria das áreas irregulares.

No âmbito da estimação em pequenos domínios, trabalha-se normalmente com dados referentes a áreas, os quais podem apresentar correlação espacial. A utilização de modelos de estimação em pequenos domínios que tomem em conta a correlação espacial entre células vizinhas e eventuais variáveis auxiliares, pode ser um instrumento precioso na estimação da característica de interesse para domínios com muito pequenas dimensões amostrais, ou mesmo nulas. Sendo o conceito de correlação ou associação espacial fundamental para a análise que se irá efectuar, este deve ser devidamente clarificado, o que se faz em seguida.

Diz-se que existe correlação espacial quando os valores de uma variável respeitantes a localizações mais próximas são mais semelhantes do que aqueles que dizem respeito a localizações mais distantes. Por exemplo, se um valor elevado para o preço médio de transacção da habitação se associa mais provavelmente a valores elevados (baixos) nos concelhos vizinhos, então diz-se que este fenómeno exibe correlação espacial positiva (negativa). A existência de correlação espacial nos dados poderá levar à existência de correlação espacial nos resíduos de uma regressão, uma vez que resíduos positivos/negativos tendem a ocorrer para observações geograficamente próximas. Consequentemente a hipótese da independência dos resíduos será violada quando existe correlação espacial[38]. Em termos formais, quando os resíduos estão espacialmente correlacionados, então a sua matriz de covariâncias não é uma matriz diagonal.

Com o objectivo de incorporar na estimação a correlação espacial existente entre domínios (áreas) vizinhos, são utilizados modelos espaciais no contexto da estimação em pequenos domínios (Cressie, 1993). A correlação espacial pode ser incorporada nos modelos de regressão linear de duas formas distintas (Anselin, 1992): (*i*) especificando um modelo de regressão com termos auto-regressivos; ou (*ii*) especificando um modelo de regressão com resíduos autocorrelacionados espacialmente.

Na estimação em pequenos domínios com dados referentes a áreas, são utilizados frequentemente os modelos de regressão linear com associação espacial incorporada na estrutura de erro. Essa associação espacial é definida na estrutura de covariâncias através de uma função da matriz de vizinhanças espaciais ou de pesos, **W**, e de um parâmetro de correlação ou associação espacial fixo e desconhecido. A matriz de

[38] Como se sabe, quando os resíduos não são independentes os estimadores de mínimos quadrados não são enviesados, mas são ineficientes.

vizinhanças espaciais é, portanto, uma ferramenta básica para se estimar a variabilidade espacial de dados referentes a áreas.

Dado um conjunto de localizações espaciais, $\{\mathbf{r}_1,...,\mathbf{r}_n\}$, define-se a matriz $\mathbf{W}=\{w_{ij}\}$, i,j=1, …, n, quadrada de ordem n e em geral simétrica, onde cada um dos seus elementos, w_{ij}, é uma função da distância entre a localização $\mathbf{r}_i$ e a localização $\mathbf{r}_j$. Usualmente, $w_{ii}=0$, i=1, …, n, mas para $i \neq j$, w_{ij}, pode ser definida de muitas maneiras diferentes. Por exemplo, esta função de distância pode ser calculada a partir de um dos seguintes critérios de escolha binária, os quais constituem as escolhas mais comuns:

(*i*) Vizinhança por adjacência: $w_{ij}=1$, se a fronteira de $\mathbf{r}_i$ partilha pelo menos um ponto comum com a fronteira de $\mathbf{r}_j$, caso contrário $w_{ij}=0$;

(*ii*) Vizinhança baseada na distância entre centróides: $w_{ij}=1$, se o centróide de $\mathbf{r}_i$ se encontra até uma determinada distância fixa, d, do centróide de $\mathbf{r}_j$, caso contrário $w_{ij}=0$.

Clayton e Bernardinelli (1996) mostraram que a especificação da matriz **W** com zeros e uns não é internamente consistente no caso em que o número de vizinhos de cada localização não é constante, o que é comum nas áreas irregulares. Nesta situação, estes autores recomendam que a matriz $\mathbf{W}=\{w_{ij}^{*}\}$ seja formada por pesos estandardizados por linhas, da forma $w_{ij}^{*}=w_{ij}/w_{i\bullet}$, onde $w_{i\bullet}$ representa o número total de áreas que partilham pelo menos um ponto comum com a fronteira de $\mathbf{r}_i$ (incluindo a área $\mathbf{r}_j$).

Outras escolhas mais informativas para a função de distância, w_{ij}, são baseadas em funções não binárias do inverso da distância dos centróides, uma vez que há situações em que é natural atribuir um valor mais elevado à associação entre blocos que se encontram mais próximos. A função de distância pode ainda ser atribuída com base na proporção de fronteira partilhada pelas áreas, nas áreas que estão num determinado raio da área de interesse, ou num índice de acessibilidade construído a partir da quantidade de vias de comunicação entre áreas.

Segundo Wall (2004), a estrutura espacial subjacente aos dados referentes a áreas pode ser modelada de duas formas diferentes. Ambas as formas são casos especiais do processo espacial geral, $\{\mathbf{Y}(\mathbf{r}) : \mathbf{r} \in D\}$, devendo-se as suas diferenças ao que se assume sobre o domínio *D*. Uma das formas de modelação da estrutura espacial consiste em tratar os dados referentes a áreas como se eles tivessem sido observados num subconjunto, *D*, contínuo de $\Re^n$, em vez de terem sido observados num subconjunto, *D*, contável de $\Re^n$. Quando se utiliza este tipo de modelação, assume-se que os dados referentes a áreas foram observados no centro ou no centróide de cada área, sendo as distâncias entre os centróides utilizadas para especificar a estrutura de covariância espacial através da função variograma. De acordo com Wall (2004), um dos problemas mais comuns neste método é a arbitrariedade presente na atribuição dos dados agrupados de toda uma região a um centróide. Mesmo que se utilize um ponto cuidadosamente escolhido para centróide, ainda persiste outro problema conceptual com a modelação de dados referentes a áreas, que consiste na impossibilidade das observações a serem modeladas se encontrarem continuamente na superfície da região como o modelo permitiria. Por outro lado, a vantagem inerente à modelação da estrutura espacial desta forma é que a função de covariância espacial é modelada directamente, o que permite uma interpretação fácil dessa estrutura. Outra das formas de modelação da estrutura espacial consiste, efectivamente, em tratar os dados referentes a áreas como observações de um subconjunto, *D*, contável de $\Re^n$. Nesta situação, a modelação da estrutura espacial é efectuada através da definição de uma estrutura de vizinhança baseada na forma do mapeamento em grade. Ao contrário do método anterior no qual são medidas as distâncias entre os centróides das regiões, neste método define-se, por exemplo, que duas regiões são vizinhas se as suas fronteiras se tocam em pelo menos um ponto (vizinhança por adjacência).

Dois conhecidos modelos que incorporam dados referentes a áreas como observações de um subconjunto, *D*, contável de $\Re^n$, são os modelos auto-regressivos simultâneos[39]

[39] A tradução para português de *Simultaneous Autoregressive Model* pode ser modelos auto-regressivos simultâneos. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (SAR), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

(SAR) e os modelos auto-regressivos condicionais[40] (CAR), os quais são modelos de regressão linear com associação espacial incorporada na estrutura de erro. O modelo SAR, apresentado por Whittle (1954), assim como o modelo CAR, proposto por Besag (1974), foram inicialmente desenvolvidos para modelar dados referentes a áreas regulares infinitas. Segundo Cressie (1993), quando estes dois modelos são utilizados para modelar dados de áreas regulares infinitas, apresentam características muito semelhantes ao modelo temporal auto-regressivo estacionário. Por sua vez, quando os modelos SAR e CAR são utilizados para modelar dados de áreas regulares finitas, Haining (1990) e Besag e Kooperberg (1995) mostraram que a estrutura de covariâncias subjacente a esses modelos é formada por variâncias heterogéneas para cada localização, assim como por covariâncias diferentes entre regiões que apresentam o mesmo número de regiões vizinhas. Na situação em que os modelos SAR e CAR são utilizados para modelar dados de áreas irregulares, Wall (2004) mostrou que a correlação espacial entre regiões, subjacente a esses modelos, não parece seguir um esquema intuitivo ou prático, ou seja, em geral não existe uma ligação intuitiva óbvia entre as estruturas de covariância e as correlações espaciais daí resultantes. Na situação em que existe interesse em perceber a estrutura espacial, Wall (2004) sugere que sejam utilizadas outras formas de modelação de dados referentes a áreas, como por exemplo os modelos geoestatísticos. É, contudo, de sublinhar que, segundo Wall (2004), se o objectivo principal da modelação de dados referentes a áreas consiste em obter bons preditores dos parâmetros de interesse a partir da regressão, em detrimento da compreensão da estrutura espacial subjacente, então os modelos SAR e CAR são uma boa escolha para essa modelação.

No âmbito dos processos espaciais é possível considerar um processo estocástico mais geral que, para além da informação espacial observada num espaço Euclidiano *d*-dimensional, inclui também a dimensão temporal, denotada pelo índice *t*. Seja $\mathbf{Y}(\mathbf{r},t)$ um vector aleatório referente à localização genérica, **r**, no momento *t*. O processo estocástico espaciotemporal geral pode ser apresentado como:

$$\{\mathbf{Y}(\mathbf{r},t):\mathbf{r}\in D(t),t\in T\}, \tag{3.4.2}$$

[40] A tradução para português de *Conditional Autoregressive Model* pode ser modelos auto-regressivos condicionais. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (CAR), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

onde se assume que os conjuntos $D(t) \subseteq \Re^d$ e $T \subseteq \Re$ são aleatórios (Cressie, 1993).

Frequentemente, $D(t) \equiv D$ e $T = \{1, 2, ...\}$, o que leva a que o processo estocástico (3.4.2) seja visto como uma série temporal de processos estocásticos, ocorrendo cada processo em períodos de tempo equidistantes. Naturalmente que o processo estocástico temporal pode ser considerado um caso particular do processo estocástico espaciotemporal geral (3.4.2), sendo apresentado como:

$$\{\mathbf{Y}(t) : t \in T\}, \tag{3.4.3}$$

onde $T \subseteq \Re$.

3.4.3 Modelos espaciais para dados referentes a áreas

Considere-se $\{\mathbf{Y}(\mathbf{r}) : \mathbf{r} \in D\}$ um processo estocástico Normal onde as localizações $\{\mathbf{r}_1, ..., \mathbf{r}_n\}$ constituem um conjunto finito de *n* pontos de *D*. Assume-se que as *n* áreas formam uma grade de *D* se $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ for uma partição de *D*, ou seja, se $\bigcup_{j=1}^{n} \mathbf{r}_i = D$ e $\mathbf{r}_i \cap \mathbf{r}_j = 0$, $\forall i \neq j$. Os valores observados nas localizações $\{\mathbf{r}_1, ..., \mathbf{r}_n\}$ são representados por $\mathbf{y} = col_{1 \le i \le n}[y(\mathbf{r}_i)]$.

3.4.3.1 Modelo auto-regressivo simultâneo (SAR)

O processo espacial $\{\mathbf{Y}(\mathbf{r}) : \mathbf{r} \in D\}$ pode ser modelado através de um modelo SAR devido a Whittle (1954), é especificado da seguinte forma:

$$y(\mathbf{r}_i) = \mu_i + \sum_{j=1}^{n} b_{ij} [y(\mathbf{r}_j) - \mu_j] + \varepsilon(\mathbf{r}_i), \tag{3.4.4}$$

onde $\varepsilon(\mathbf{r}_i)$ são erros *iid* com $\boldsymbol{\varepsilon} = col_{1 \le i \le n}[\varepsilon(\mathbf{r}_i)] \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ sendo $\boldsymbol{\Lambda}$ uma matriz diagonal, $E[y(\mathbf{r}_i)] = \mu_i$ e b_{ij} são constantes pertencentes a uma matriz $\mathbf{B} = [b_{ij}]$ de dimensão *n*×*n* com $b_{ii} = 0$, *i*=1,…,*n*. Este modelo é denominado "simultâneo" porque, em geral, os termos de erro, ε_i, estão correlacionados com $\{y(\mathbf{r}_j) : i \neq j\}$. Note-se que a

simultaneidade do modelo resulta do facto de se considerarem variáveis explicativas as áreas vizinhas contemporâneas, ou seja, a especificação dos modelos é baseada na forma como os dados de várias localizações interagem simultaneamente. O modelo (3.4.4) pode também ser especificado da seguinte forma:

$$(\mathbf{I}_n - \mathbf{B})(\mathbf{y} - \boldsymbol{\mu}) = \boldsymbol{\varepsilon}, \tag{3.4.5}$$

onde $\mathbf{y} = col_{1\le i\le n}[y(\mathbf{r}_i)]$ e $\boldsymbol{\mu} = col_{1\le i\le n}(\mu_i)$. Uma vez que $\boldsymbol{\varepsilon}$ segue uma distribuição Normal, então a distribuição conjunta de $\mathbf{y}$ é dada por:

$$\mathbf{y} \sim N\left[\boldsymbol{\mu}; (\mathbf{I}_n - \mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I}_n - \mathbf{B})^{-1}\right]. \tag{3.4.6}$$

Pode considerar-se a inclusão de parâmetros de regressão em tendência espacial no modelo SAR, bastando para tal considerar que a variação em larga escala possa ser modelada como $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, onde $\mathbf{X}$ é uma matriz de dimensão $n\times p$ cujas colunas são formadas por p variáveis explicativas e $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ é um vector de parâmetros de regressão. Ou seja, a variável de interesse na localização i, $Y(\mathbf{r}_i)$, está acompanhada de p variáveis explicativas $\{x_j(\mathbf{r}_i): j = 1,...,p\}$, tal que $E[y(\mathbf{r}_i)] = \sum_{j=1}^{p} x_j(\mathbf{r}_i)\beta_j$. Desta forma, o modelo SAR (3.4.4) pode ser especificado da seguinte forma:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}, \\ \boldsymbol{\delta} &= \mathbf{B}\boldsymbol{\delta} + \boldsymbol{\varepsilon}. \end{aligned} \tag{3.4.7}$$

Whittle (1954) mostrou que a estimação dos parâmetros em larga escala presentes em $\boldsymbol{\beta}$ e dos parâmetros em pequena escala presentes nas matrizes $\boldsymbol{\Lambda}$ e $\mathbf{B}$ pode ser efectuada através de um método de MV aproximado. Este trabalho pioneiro foi posteriormente melhorado por Guyon (1982). Whittle (1954) mostrou também que no âmbito do modelo SAR, o método dos mínimos quadrados e as equações de Yule-Walker não produzem estimadores consistentes para os parâmetros desse modelo. Por último, note-se que o modelo SAR é uma extensão para o caso *n*-dimensional do modelo auto-regressivo de primeira ordem [AR(1)], o qual foi originalmente desenvolvido para tratar séries temporais.

3.4.3.2 Modelo auto-regressivo condicional (CAR)

Outra forma de modelar o processo $\{\mathbf{Y}(\mathbf{r}): \mathbf{r} \in D\}$ é através do modelo CAR desenvolvido por Besag (1974). Este modelo define-se da seguinte forma:

$$E\left[y(\mathbf{r}_i) \mid \{y(\mathbf{r}_j), i \neq j\}\right] = \mu_i + \sum_{j=1}^{n} c_{ij}\left[y(\mathbf{r}_j) - \mu_j\right], \tag{3.4.8}$$

onde $E\left[y(\mathbf{r}_i)\right] = \mu_i$, $V\left[y(\mathbf{r}_i) \mid \{y(\mathbf{r}_j), i \neq j\}\right] = \tau_i^2$, c_{ij} são constantes pertencentes a uma matriz $\mathbf{C} = \left[c_{ij}\right]$ de dimensão $n \times n$ com $c_{ii} = 0$ e satisfazendo a seguinte condição de simetria $c_{ij}\tau_j^2 = c_{ji}\tau_i^2$, i=1,…,n. Pelo teorema de factorização (Besag, 1974), tem-se que a distribuição condicional conjunta é dada por:

$$\mathbf{y} \sim N\left[\boldsymbol{\mu}; (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{M}\right], \tag{3.4.9}$$

onde $\mathbf{M} = diag_{1 \leq i \leq n}\left(\tau_i^2\right)$. À semelhança do modelo SAR, quando se considera a inclusão de parâmetros de regressão em tendência espacial no modelo CAR, através de $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, o modelo (3.4.8) pode ser apresentado da seguinte forma:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}, \\ \boldsymbol{\delta} &\sim N\left[\mathbf{0}; (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{M}\right]. \end{aligned} \tag{3.4.10}$$

A estrutura das matrizes B e C, respectivamente nos modelos SAR e CAR, é normalmente especificada com base na forma do mapeamento em grade. Uma forma comum de definição dessas matrizes consiste na utilização de uma matriz que resulta do produto de uma matriz de distâncias, **W**, por um parâmetro que pondera essa estrutura de vizinhança. Por conseguinte, define-se habitualmente que $\mathbf{B} = \phi_S \mathbf{W}$ no modelo SAR e que $\mathbf{C} = \phi_C \mathbf{W}$ no modelo CAR, onde ϕ_S e ϕ_C são denominados por parâmetros de correlação ou dependência espacial.

3.4.3.3 Modelo auto-regressivo espaciotemporal simultâneo

O modelo SAR foi generalizado por Cliff e Ord (1975), de forma a acomodar também a dimensão temporal, para além da informação espacial n-dimensional. Uma dessas

generalizações deu origem ao conhecido modelo auto-regressivo espaciotemporal simultâneo[41] (STAR)[42]. O modelo STAR pode ser utilizado quando existe interesse e/ou necessidade de avaliar as interdependências espaciais e temporais presentes nos dados. Este modelo tem um vasto campo de aplicação, desde as experiências agrícolas, passando pelos estudos na área da saúde, até aos estudos económico-sociais (*vide e.g.* Haining, 1990). Para além disso, o modelo STAR também pode ser utilizado para modelar as dependências espaciotemporais presentes nos resíduos de um modelo de regressão. Segundo de Luna e Genton (2002), a modelação simultânea dos resíduos de um processo (por exemplo através de um modelo STAR), tem a vantagem de ser mais parcimoniosa do que a correspondente modelação condicional.

Seja $\{\mathbf{Y}(\mathbf{r},t):\mathbf{r}\in D(t), t\in T\}$ um processo estocástico Normal onde as localizações $\{\mathbf{r}_1,...,\mathbf{r}_n\}$ constituem um conjunto finito de *n* pontos de $D(t)$, no momento *t*, no qual $D(t)\subseteq\Re^d$ e $T\subseteq \mathbf{Z}$. Os valores observados nas localizações $\{\mathbf{r}_1,...,\mathbf{r}_n\}$ são representados por $y(t)=col_{1\le i\le n}[y(\mathbf{r}_i,t)]$, t=1, …, T_0. Este processo pode ser modelado através de um modelo STAR que, na sua forma mais geral, é especificado da seguinte forma:

$$\mathbf{y}(t)=\boldsymbol{\mu}+\sum_{k=0}^{p}\mathbf{B}_k[\mathbf{y}(t-k)-\boldsymbol{\mu}]+\boldsymbol{\varepsilon}(t), \qquad (3.4.11)$$

onde $E[\mathbf{y}(t)]=\boldsymbol{\mu}$, sendo que $\boldsymbol{\mu}=col_{1\le i\le n}[\mu_i(t)]$ é um vector que representa a variação em larga escala, e $\boldsymbol{\varepsilon}(t)=col_{1\le i\le n}[\varepsilon(\mathbf{r}_i,t)]$ é um vector de erros *iid* (normalmente Gaussianos) com média nula e variância σ^2 finita, $\mathbf{B}_k=\sum_{j=1}^{\lambda_k}\xi_{kj}\mathbf{W}_{kj}$, sendo $\mathbf{W}_{kj}$ uma matriz de pesos

[41] *Simultaneous Spatio-temporal Autoregressive Model* pode ser traduzido para português por modelo auto-regressivo espaciotemporal simultâneo. Decidiu usar-se nesta tese as siglas da designação em língua inglesa (STAR), por serem as utilizadas internacionalmente e por não existir uma tradução oficial para a língua portuguesa.

[42] Na realidade, Cliff e Ord (1975) apresentaram o modelo espaciotemporal auto-regressivo de médias móveis (*space-time autoregressive moving average*) (STARMA), como uma generalização do modelo temporal auto-regressivo de médias móveis e do modelo espacial especificado de forma simultânea. Neste documento só é apresentado o modelo STAR, o qual é um caso particular do modelo STARMA, porque no âmbito da estimação em pequenos domínios não se utilizam normalmente modelos de médias móveis.

conhecida, ξ_{kj} são parâmetros do modelo e λ_k é o grau do desfasamento espacial. À semelhança dos modelos SAR e CAR, pode considerar-se a inclusão de parâmetros de regressão em tendência espacial no modelo STAR através de $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

Quando no modelo (3.4.11) se admite o caso particular com *p*=0, então obtém-se para qualquer ponto no tempo, *t*, o modelo SAR (3.4.7), apresentado na secção 3.4.3.1. Quando se admite que $\lambda_k = 1$, então a matriz de dependência espacial é dada por $\mathbf{B}_k = \xi_{k1}\mathbf{W}_{k1}$, a qual depende apenas de um parâmetro desconhecido. Por sua vez, quando se admite que $\lambda_k = n$ e para uma particular escolha de $\mathbf{W}_{kj}$, *j*=1, …, *n*, então a matriz de dependência pode ser definida como $\mathbf{B}_k = diag_{1\le j\le n}(\xi_{kj})\mathbf{H}_k$, onde $\mathbf{H}_k$ é uma matriz quadrada de dimensão *n* conhecida.

A consideração da dimensão temporal no modelo SAR conduz a dificuldades adicionais na estimação do modelo, no qual é necessário ter um cuidado especial na estimação dos parâmetros só pelo facto de se considerar a informação espacial, tal como foi referido anteriormente. A estimação de processos Gaussianos com estas características, começou a ser desenvolvida por Ali (1979), que trabalhou ao nível da estimação pelo método da MV. Posteriormente, de Luna e Genton (2002) propuseram um método baseado em simulação, para estimação de processos estocásticos espaciotemporais simultâneos, destinados à modelação de dados referentes a áreas reticuladas regulares.

3.4.3.4 Padrão espacial

A análise de dados espaciais pode ser dividida em duas abordagens: uma abordagem baseada nos dados e uma abordagem baseada em modelos (Anselin, 1992). A abordagem baseada nos dados, frequentemente catalogada na categoria das análises exploratórias, caracteriza-se por dois aspectos essenciais. Em primeiro lugar, nesta abordagem assume-se a hipótese da aleatoriedade, isto é, qualquer valor observado pode ocorrer em qualquer localização com igual probabilidade. Em segundo lugar, o estudo da presença de um padrão espacial é efectuado exclusivamente com base nos dados, independentemente do modelo teórico subjacente aos dados. Por outro lado, a abordagem baseada em modelos é suportada pela especificação de um modelo teórico, o qual é ajustado aos dados. Na abordagem baseada nos dados são utilizadas estatísticas

globais de associação espacial, enquanto na abordagem baseada nos modelos são usados testes de diagnóstico de associação espacial específicos para cada modelo.

A. Estatísticas globais de associação espacial

As duas estatísticas globais mais conhecidas para medir a associação espacial de dados referentes a áreas[43] são a estatística *I* de Moran (Moran, 1948) e a estatística *c* de Geary (Geary, 1954). Estas duas estatísticas são definidas, respectivamente, da seguinte forma:

$$I = \frac{n\sum_i\sum_j w_{ij}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{\left(\sum_{i\neq j} w_{ij}\right)\sum_i\left(Y_i - \bar{Y}\right)^2}, \tag{3.4.12}$$

$$c = \frac{(n-1)\sum_i\sum_j w_{ij}\left(Y_i - Y_j\right)^2}{2\left(\sum_{i\neq j} w_{ij}\right)\sum_i\left(Y_i - \bar{Y}\right)^2}, \tag{3.4.13}$$

onde n é o número de observações indexadas a i e j, Y_i é a observação da variável de interesse na localização i e w_{ij} são os elementos de uma matriz de pesos, **W**. Apesar da distribuição da estatística *I* não ser estritamente suportada pelo intervalo [-1;1], ao contrário do que se verifica com o coeficiente de correlação de Pearson, pode no entanto concluir-se que a estatística é positiva quando os valores observados de uma variável respeitantes a localizações mais próximas tendem a ser mais semelhantes, é negativa quando esses valores tendem a ser dissemelhantes e é aproximadamente igual a zero quando os valores observados estão distribuídos aleatoriamente e independentemente no espaço (Cliff e Ord, 1981). Por sua vez, a estatística *c* encontra-se no intervalo [0;2], sendo $c \gg 1$ quando os valores observados de uma variável respeitantes a localizações mais próximas tendem a ser mais dissemelhantes (associação espacial negativa), e $c \ll 1$ quando esses valores tendem a ser semelhantes (associação espacial positiva) (Cliff e Ord, 1981).

Às estatísticas *I* de Moran e *c* de Geary podem associar-se testes de existência de associação espacial, para os quais não existem distribuições exactas, a não ser para

[43] As estatísticas *I* de Moran e *c* de Geary foram inicialmente desenvolvidas para medir a associação de dados referentes a pontos. Contudo, foram posteriormente generalizadas para a medição da associação de dados referentes a áreas.

dimensões amostrais muito pequenas. Cliff e Ord (1981) apresentaram as condições para as quais as distribuições assimptóticas de *I* e de *c* são normais. Carvalho e Natário (2008) referem outra alternativa para testar a existência de associação espacial, baseada na construção das distribuições de *I* e de *c* através da utilização de métodos de Monte Carlo. Neste caso, trata-se de calcular o valor-*p* empírico da hipótese nula de não existência de associação espacial, verificando a posição da estatística-teste na amostra ordenada dos valores da estatística calculados a partir de cada uma das réplicas, e compará-lo em seguida com o valor crítico definido.

B. Testes de diagnóstico de associação espacial na estrutura de erro

Os testes de diagnóstico de associação espacial incorporada na estrutura de erro em modelos espaciais dependem do tipo de estrutura espacial (SAR ou CAR). Em seguida são apresentados dois desses testes normalmente utilizados na detecção de estruturas espaciais auto-regressivas do tipo SAR, o qual é utilizado no estudo empírico deste trabalho.

Cliff e Ord (1973) sugeriram um teste com o objectivo de verificar a existência de associação espacial nos modelos SAR, à semelhança do teste de Durbin e Watson, utilizado para verificar a existência de autocorrelação de primeira ordem numa série temporal. Admitindo o modelo (3.4.7), $\boldsymbol{\Lambda} = \sigma^2 \mathbf{I}_n$ e $\mathbf{B} = \phi_S \mathbf{W}$, sendo $\mathbf{W}$ uma matriz de distâncias conhecida, esse teste tem as seguintes hipóteses subjacentes: $H_0 : \phi_S = 0$ versus $H_1 : \phi_S \neq 0$.

Com o objectivo de identificar a associação espacial incorporada na estrutura de erro, Cliff e Ord (1973) propuseram a seguinte estatística-teste, de acordo com o índice *I* de Moran (Moran, 1948):

$$T_{CO} = \frac{\hat{\mathbf{e}}'\mathbf{W}\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\hat{\mathbf{e}}} \overset{\circ}{\sim} N(\mu;\sigma), \tag{3.4.14}$$

onde $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ é um vector dos resíduos de uma regressão linear pelo método dos mínimos quadrados ordinários e $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ é o estimador dos mínimos quadrados de $\boldsymbol{\beta}$. Esta estatística teste segue uma distribuição normal assimptótica quando a

amostra for grande. Este teste tem a desvantagem de não conseguir fazer uma boa discriminação entre uma associação espacial incorporada na estrutura de erro e uma associação espacial substancial.

Burridge (1980) mostrou que a associação espacial incorporada na estrutura de erro pode ser identificada pelo princípio do Multiplicador de Lagrange, tendo demonstrado que sob $H_0 : \phi_S = 0$,

$$T_B = \frac{n\hat{\mathbf{e}}'\mathbf{W}\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\hat{\mathbf{e}}\sqrt{tr\left(\mathbf{W}^2 + \mathbf{W}'\mathbf{W}\right)}} \sim N(0;1), \qquad (3.4.15)$$

onde $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ é o estimador dos mínimos quadrados de $\boldsymbol{\beta}$.

3.4.3.5 Outros modelos espaciais

No contexto da modelação de dados espaciais contínuos referentes a áreas, podem ser utilizados outros modelos para além dos três modelos apresentados anteriormente, os quais ainda não têm sido utilizados no âmbito da estimação em pequenos domínios. Em seguida são enumerados alguns modelos que apresentam afinidades com os modelos SAR, CAR e STAR.

Um processo espacial $\{\mathbf{Y}(\mathbf{r}) : \mathbf{r} \in D\}$ também pode ser modelado através de um modelo espacial de médias móveis (*spatial-moving-average regression model*) ou de um modelo espacial auto-regressivo de médias móveis (*spatial autoregressive-moving-average regression model*). Por sua vez, um processo espacial $\{\mathbf{Y}(\mathbf{r},t) : \mathbf{r} \in D(t), t \in T\}$ também pode ser modelado através de um modelo espaciotemporal auto-regressivo de médias móveis, como uma generalização do modelo temporal auto-regressivo de médias móveis das séries temporais e do modelo espacial simultâneo.

Todos estes modelos podem ser consultados em Anselin (1992) e em Cressie (1993). Na obra de Anselin (1992) são apresentados os fundamentos dos efeitos espaciais na econometria, na qual são discutidos os métodos de estimação dos mínimos quadrados, da MV, das variáveis instrumentais e dos momentos, de forma a acomodar a correlação espacial no modelo de regressão linear.

4. MODELOS PARA ESTIMAÇÃO EM PEQUENOS DOMÍNIOS

4.1 INTRODUÇÃO

Na maior parte das situações, a estimação em pequenos domínios é assistida por casos particulares do modelo linear misto geral, os quais relacionam os valores de uma variável de interesse com os valores de variáveis auxiliares. Este tipo de modelos permite fazer a estimação de um parâmetro da variável de interesse num pequeno domínio, tendo em conta informação de outros pequenos domínios através de efeitos fixos e de efeitos aleatórios.

Fay e Herriot (1979) foram os primeiros autores a utilizar um modelo, considerado um caso particular do modelo linear misto, para produzir estimativas em pequenos domínios do rendimento médio *per capita* nos Estados Unidos da América, a partir de dados amostrais de natureza seccional. O modelo de nível área que foi utilizado em 1979 por Fay e Herriot, é hoje sobejamente conhecido na literatura sobre pequenos domínios como modelo de Fay-Herriot.

A partir dessa data, muitos outros autores utilizaram modelos lineares mistos para estimação em pequenos domínios, embora nem todos esses modelos sejam de nível área. Em 1981, Battese e Fuller utilizaram pela primeira vez um modelo de nível unidade, denominado modelo de regressão de erros encaixados (*one-fold nested error regression model*), para estimar o número médio de hectares de cereais em doze pequenos domínios (distritos) nos Estados Unidos da América. Para tal, utilizaram dados amostrais conjuntamente com informação obtida por satélite. Posteriormente, a especificação deste modelo foi publicada detalhadamente por Battese *et al*. (1988),

sendo este modelo frequentemente referido na literatura sobre pequenos domínios como modelo de Battese-Harter-Fuller. Algumas das mais importantes extensões do modelo de Battese-Harter-Fuller foram apresentadas por Fuller e Harter (1987), Arora e Lahiri (1997), Stukel e Rao (1999), Moura e Holt (1999), Coelho (2000), e Petrucci e Salvati (2004*a*). A apresentação e discussão de modelos de estimação em pequenos domínios de nível unidade estão fora do âmbito deste trabalho, pelo que se remete o leitor para as referências apresentadas acima.

Por outro lado, muitos outros autores têm vindo a apresentar extensões do modelo de nível área de Fay-Herriot (Fay e Herriot, 1979), de forma a ser possível contemplar erros de sondagem correlacionados, bem como trabalhar com dados seccionais, espaciais e/ou cronológicos, *etc*. Fay (1987) propôs um modelo de Fay-Herriot multivariado; Isaki *et al*. (2000) propuseram um modelo com erros de sondagem correlacionados; Choudhry e Rao (1989), Pfeffermann e Burk (1990), Rao e Yu (1992, 1994), Yu (1993) e Singh *et al*. (1994) propuseram modelos que utilizam informação seccional e cronológica; Cressie (1991) e Salvati (2004) propuseram modelos que utilizam informação espacial; e Singh *et al*. (2005) propuseram um modelo que utiliza simultaneamente informação seccional/espacial e cronológica.

Um dos objectivos deste capítulo consiste na apresentação detalhada dos principais modelos de nível área para estimação em pequenos domínios, presentes na literatura, adequados para tratar dados de natureza seccional, cronológica, espacial e espaciotemporal. Desta forma, no subcapítulo 4.2 é apresentado o modelo básico de Fay e Herriot (1979), o subcapítulo 4.3 é dedicado ao modelo longitudinal básico de Rao e Yu (1994) e no subcapítulo 4.4 são expostos modelos do tipo *state space*. Os modelos espaciais são introduzidos no subcapítulo 4.5 e no subcapítulo seguinte é apresentado um modelo espaciotemporal do tipo *state space*. Os comentários finais são apresentados no final do capítulo.

Outro dos objectivos deste capítulo consiste na proposta de novos métodos de estimação do EQMP do EBLUP temporal sob o modelo de Rao e Yu (1994). Estes novos desenvolvimentos são introduzidos nas secções 4.3.6 e 4.3.7.

4.2 MODELO BÁSICO DE NÍVEL ÁREA COM DADOS SECCIONAIS

4.2.1 Especificação do modelo de Fay-Herriot

O modelo de nível área básico que a seguir é exposto foi apresentado por Fay e Herriot (1979). Neste modelo, considera-se que existe informação auxiliar disponível para *p* variáveis auxiliares apenas ao nível de cada um dos *m* pequenos domínios de estudo, $\mathbf{x}_i = \left(x_{i1}, ..., x_{ip}\right)'$, *i*=1, …, *m*. Assume-se também que a média populacional da variável de interesse, μ_i (ou o total populacional, τ_i), ou qualquer função do tipo $\theta_i = g\left(\mu_i\right)$, para cada um dos *m* domínios da população, está ligada ao vector das variáveis explicativas, $\mathbf{x}_i$, através de um modelo linear com efeitos aleatórios associados aos domínios. O modelo de ligação proposto por Fay e Herriot (1979), foi definido como a soma de uma componente estrutural, $\mathbf{x}_i'\beta$, com uma componente de domínio, u_i, de natureza aleatória:

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \tag{4.2.1}$$

onde $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)'$ é um vector $p\times1$ dos coeficientes de regressão desconhecidos (efeitos fixos), e os u_i's representam efeitos aleatórios associados aos domínios, que se admite satisfazerem $u_i \overset{iid}{\sim} \left(0, \sigma_u^2\right)$. A hipótese da normalidade dos u_i's é também frequentemente utilizada, mas é possível fazer inferências "robustas" sem se ter em conta essa hipótese. Para se fazer inferência sobre um parâmetro populacional de uma variável de interesse em pequenos domínios, θ_i, sob o modelo (4.2.1), assume-se que está disponível o seu estimador directo *design-based*, aqui representado como y_i por conveniência de notação, sempre que $n_i \geq 1$:

$$y_i = g\left(\hat{\mu}_i\right) = \theta_i + \varepsilon_i \tag{4.2.2}$$

onde se assume que os erros da sondagem, ε_i, são independentes com média nula e variância conhecida, dados os θ_i: $E_d\left(\varepsilon_i \mid \theta_i\right) = 0$ e $V_d\left(\varepsilon_i \mid \theta_i\right) = \sigma_{\varepsilon,i}^2$.

Da combinação do modelo de ligação (4.2.1) com o modelo usado para incorporar os erros da sondagem na estimação, (4.2.2), obtém-se o modelo proposto por Fay e Herriot (1979), que se passa a denominar por modelo de Fay-Herriot:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i + \varepsilon_i, \tag{4.2.3}$$

no qual se assume que os u_i's são independentes dos ε_i's (i=1,..., m). O modelo (4.2.3) é um caso particular do modelo linear misto (3.2.1), com efeitos fixos, $\boldsymbol{\beta}$, e com efeitos aleatórios associados aos domínios, u_i. Segundo Ghosh e Rao (1994) e Rao (2000), o modelo combinado (4.2.3) compreende não só variáveis aleatórias *design-based*, ε_i, mas também variáveis aleatórias *model-based*, u_i. Note-se também que o modelo (4.2.3) compreende dois tipos de parâmetros – os parâmetros de elevada dimensão (*high dimensional parameters*), θ_i, e os parâmetros de baixa dimensão (*low dimensional parameters*), $\boldsymbol{\beta}$ e σ_u^2, normalmente denominados por hiper-parâmetros. No contexto da estimação em pequenos domínios, o principal objectivo consiste em estimar os parâmetros θ_i, o que exige a estimação de hiper-parâmetros desconhecidos.

Pode considerar-se um caso mais geral do modelo proposto por Fay e Herriot (1979), no qual se assume que os efeitos aleatórios específicos de domínio, u_i, têm um coeficiente positivo conhecido, z_i. Neste caso, o modelo (4.2.3) é definido da seguinte forma:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + z_i u_i + \varepsilon_i. \tag{4.2.4}$$

O modelo de Fay-Herriot é um caso particular deste novo modelo com $z_i = 1,\ \forall\, i$ (i=1, ..., m). O modelo combinado (4.2.4) pode ser apresentado em notação matricial da seguinte forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4.2.5}$$

onde $\mathbf{y} = col_{1\le i\le m}(y_i)$ é um vector $m\times 1$ de estimadores directos *design-based* para os m domínios, $\mathbf{X} = col_{1\le i\le m}(\mathbf{x}_i')$ é uma matriz $m\times p$ de valores conhecidos de p variáveis auxiliares para cada um dos m domínios, $\mathbf{Z} = diag_{1\le i\le m}(z_i)$ é uma matriz $m\times m$ de

constantes positivas conhecidas, $\mathbf{u} = col_{1\le i\le m}(u_i)$ é um vector $m\times 1$ de efeitos aleatórios associados aos *m* domínios e $\boldsymbol{\varepsilon} = col_{1\le i\le m}(\varepsilon_i)$ é um vector $m\times 1$ de erros da sondagem.

Assume-se igualmente que os efeitos aleatórios, $\mathbf{u} \overset{iid}{\sim} (\mathbf{0}; \mathbf{G})$, são independentes dos erros da sondagem, $\boldsymbol{\varepsilon} \overset{iid}{\sim} (\mathbf{0}; \mathbf{R})$, sendo $\mathbf{G} = \sigma_u^2 \mathbf{I}_m$ e $\mathbf{R} = diag_{1\le i\le m}(\sigma_{\varepsilon,i}^2)$ matrizes de dimensão $m\times m$. A matriz de covariâncias de $\mathbf{y}$, também de dimensão $m\times m$, é dada por $\mathbf{V} = diag_{1\le i\le m}(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2)$.

4.2.2 O EBLUP

Como o modelo (4.2.5) é um caso particular do modelo linear misto (3.2.1), então podem utilizar-se os resultados gerais deduzidos por Henderson (1975) na determinação do BLUP do parâmetro de interesse. Assumindo que a componente de variância σ_u^2 é conhecida, então o BLUP de θ_i, sob o modelo (4.2.4), é dado por (Ghosh e Rao, 1994):

$$\begin{aligned}\tilde{\theta}_i = \tilde{\theta}_i^H(\sigma_u^2) &= \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \gamma_i(y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}) \\ &= \gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\tilde{\boldsymbol{\beta}},\end{aligned} \tag{4.2.6}$$

onde $\gamma_i = \gamma_i(\sigma_u^2) = \dfrac{\sigma_u^2 z_i^2}{\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2}$ e $\tilde{\boldsymbol{\beta}}$ é o estimador dos mínimos quadrados generalizados de $\boldsymbol{\beta}$ com ponderadores $(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2)^{-1}$, dado por:

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_u^2) = \left[\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2)^{-1}\right]^{-1}\left[\sum_{i=1}^m \mathbf{x}_i y_i(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2)^{-1}\right]. \tag{4.2.7}$$

A partir da expressão (4.2.6), verifica-se que $\tilde{\theta}_i$ pode ser apresentado como uma média ponderada do estimador directo, y_i, e do estimador sintético pela regressão, $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}$, com ponderador γ_i $(0 \le \gamma_i \le 1)$, verificando, portanto, as condições de um estimador combinado. Este ponderador mede a incerteza na modelação dos parâmetros de interesse nos domínios, conhecida por variância do modelo, $\sigma_u^2 z_i^2$, relativamente à incerteza total, $(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2)$. Se a variância do modelo, $\sigma_u^2 z_i^2$, for relativamente

pequena, então é dado um peso maior ao estimador sintético pela regressão. Pelo contrário, é atribuído um peso superior ao estimador directo quando a variância no desenho, $\sigma^2_{\varepsilon,i}$, é relativamente pequena. Note-se, ainda, que para domínios com $n_i = 0$, o BLUP de θ_i é dado unicamente pelo estimador sintético pela regressão.

O estimador $\tilde{\theta}_i$ é válido para desenhos de sondagem gerais porque se está a modelar apenas os parâmetros de interesse nos domínios, θ_i's, e não os elementos individuais da população, como acontece nos modelos de nível unidade, e porque o estimador directo utiliza os pesos amostrais.

O BLUP (4.2.6) depende da componente de variância, σ^2_u, que é desconhecida nas aplicações práticas. Se se substituir em (4.2.6) σ^2_u por um estimador assimptoticamente consistente, $\hat{\sigma}^2_u$, obtém-se o seguinte preditor em dois passos:

$$\hat{\theta}_i = \hat{\theta}_i^H\left(\hat{\sigma}^2_u\right) = \hat{\gamma}_i y_i + \left(1-\hat{\gamma}_i\right)\mathbf{x}_i'\hat{\boldsymbol{\beta}} \tag{4.2.8}$$

onde $\hat{\gamma}_i$ e $\hat{\boldsymbol{\beta}}$ são os valores de $\gamma_i\left(\sigma^2_u\right)$ e $\tilde{\boldsymbol{\beta}}\left(\sigma^2_u\right)$, respectivamente, quando σ^2_u é substituído por $\hat{\sigma}^2_u$. O preditor $\hat{\theta}_i$ é também denominado por BLUP empírico ou EBLUP, por analogia ao EBP (Harville, 1991). A estimação da componente de variância é discutida na secção 4.2.3.

Com base nos resultados de Kackar e Harville (1981), é possível concluir que $\hat{\theta}_i$ é um preditor centrado de θ_i sob as seguintes condições de regularidade: (*i*) $E\left(\hat{\theta}_i\right)$ é finito; (*ii*) $\hat{\sigma}^2_u$ é qualquer estimador ímpar e invariante a translações para qualquer $\mathbf{y}$ e $\boldsymbol{\beta}$, ou seja, $\hat{\sigma}^2_u(\mathbf{y}) = \hat{\sigma}^2_u(-\mathbf{y})$ e $\hat{\sigma}^2_u(\mathbf{y} - \mathbf{X}\mathbf{h}) = \hat{\sigma}^2_u(\mathbf{y}), \forall \mathbf{h} \in \Re^p$; (*iii*) as distribuições de $\mathbf{u}$ e $\boldsymbol{\varepsilon}$ são simétricas (mas não necessariamente normais) (Ghosh e Rao, 1994).

4.2.3 Estimação da componente de variância

Na derivação do BLUP de θ_i assumiu-se que as matrizes de covariâncias $\mathbf{G}$ e $\mathbf{R}$ são conhecidas. Contudo, nas aplicações práticas a componente de variância σ^2_u é

geralmente desconhecida. Naturalmente que é necessário obter uma estimativa precisa e consistente de σ_u^2, de forma a se obterem estimadores EBLUP eficientes dos parâmetros de interesse. No âmbito da estimação em pequenos domínios, a estimação das componentes de variância tem sido alvo de muito interesse, apesar de constituir apenas um passo intermédio do processo.

Tem sido apresentada na literatura uma grande variedade de métodos de estimação de σ_u^2. Entre esses métodos encontra-se o método dos momentos iterativo de Fay e Herriot (1979), o método dos momentos explícito proposto por Prasad e Rao (1990) e os métodos da MV e da MVR utilizados, por exemplo, em Datta e Lahiri (2000) e Datta *et al*. (2005). Em seguida são sucintamente revistos todos esses métodos de estimação de σ_u^2, que sob certas condições de regularidade produzem estimadores assimptoticamente consistentes para $m \to \infty$.

4.2.3.1 Estimador pelo método dos momentos de Fay-Herriot

Fay e Herriot (1979) propuseram um método de estimação de σ_u^2 baseado numa solução iterativa de uma equação não linear. Este autores começaram por notar que a esperança matemática da soma dos quadrados dos resíduos ponderados é igual ao respectivo número de graus de liberdade, isto é:

$$E\left[\sum_{i=1}^{m}\frac{\left(y_i-\mathbf{x}_i'\tilde{\boldsymbol{\beta}}\right)^2}{\left(\sigma_u^2 z_i^2+\sigma_{\varepsilon,i}^2\right)}\right]=E\left[h\left(\sigma_u^2\right)\right]=m-p\,, \tag{4.2.9}$$

onde $\tilde{\boldsymbol{\beta}}$ é o estimador dos mínimos quadrados ponderados de $\boldsymbol{\beta}$ dado por (4.2.7). O estimador da componente de variância, $\hat{\sigma}_{u,FH}^2$, é obtido a partir da resolução iterativa de

$$h\left(\sigma_u^2\right)=m-p \tag{4.2.10}$$

até se encontrar uma solução única $\hat{\sigma}_{u,FH}^2 \geq 0$ que verifique (4.2.10) e $\tilde{\theta}_i=\mathbf{x}_i'\tilde{\boldsymbol{\beta}}$. Quando não for encontrada uma solução positiva, então faz-se $\hat{\sigma}_{u,FH}^2=0$. Fay e Herriot (1979) sugeriram o seguinte processo iterativo: começar com $\sigma_u^{2(0)}=0$ e definir

$$\sigma_u^{2(a+1)} = \sigma_u^{2(a)} + \frac{1}{h'_*\left(\sigma_u^{2(a)}\right)}\left[m - p - h\left(\sigma_u^{2(a)}\right)\right], \qquad (4.2.11)$$

forçando a que $\sigma_u^{2(a+1)} \geq 0$, onde o índice superior (a) se refere aos valores de σ_u^2 na a-ésima iteração (a=0, 1, 2, ...) e onde $h'_*\left(\sigma_u^2\right) = -\sum_{i=1}^{m} \frac{z_i^2\left(y_i - \mathbf{x}'_i\tilde{\boldsymbol{\beta}}\right)^2}{\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^2}$ é uma aproximação à derivada de $h\left(\sigma_u^2\right)$. Segundo Fay e Herriot (1979), a convergência do processo iterativo (4.2.11) é rápida, requerendo geralmente menos do que dez iterações. Vários autores têm notado que o estimador de Fay-Herriot é ímpar, invariante a translações e assimptoticamente consistente sob certas condições de regularidade (Ghosh e Rao, 1994; Datta *et al*., 2005).

Só recentemente foi deduzida por Datta *et al.* (2005) a variância assimptótica (quando $m \to \infty$) do estimador da componente de variância produzido pelo método dos momentos de Fay-Herriot, $\hat{\sigma}_{u,FH}^2$. Essa variância é dada por:

$$\overline{V}\left(\hat{\sigma}_{u,FH}^2\right) \approx 2m\left[\sum_{i=1}^{m} z_i^2\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-1}\right]^{-2}. \qquad (4.2.12)$$

Datta *et al*. (2005) também mostraram, através de um estudo por simulação, que o estimador da componente de variância de Fay-Herriot é o melhor estimador em termos de enviesamento relativo do EQMP do EBLUP.

4.2.3.2 Estimador pelo método dos momentos de Prasad-Rao

O método de estimação de Prasad-Rao é baseado no bem conhecido método III dos momentos de C.R. Henderson (Henderson, 1953). Prasad e Rao (1990) propuseram o estimador $\hat{\sigma}_{u,PR}^2 = \max\left\{0; \tilde{\sigma}_{u,PR}^2\right\}$, onde $\tilde{\sigma}_{u,PR}^2$ é um estimador quadrático não enviesado de σ_u^2, dado por:

$$\tilde{\sigma}_{u,PR}^2 = \frac{1}{m-p}\left\{\sum_{i=1}^{m} z_i^{-2}\left(y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}^*\right)^2 - \sum_{i=1}^{m}\frac{\sigma_{\varepsilon,i}^2}{z_i^2}\left[1 - \mathbf{x}'_i\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}'_i\right)^{-1}\mathbf{x}_i\right]\right\}, \qquad (4.2.13)$$

onde $\hat{\boldsymbol{\beta}}^{*}=\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{x}_i y_i\right)$ é o estimador dos mínimos quadrados ordinários de $\boldsymbol{\beta}$.

O estimador (4.2.13) é ímpar, invariante a translações e, sob certas condições de regularidade, é assimptoticamente consistente quando $m\to\infty$. A variância assimptótica do estimador $\hat{\sigma}^2_{u,PR}$, sob condições de normalidade de u_i e de ε_i, é dada por (Prasad e Rao, 1990):

$$\overline{V}\left(\hat{\sigma}^2_{u,PR}\right)=\overline{V}\left(\tilde{\sigma}^2_{u,PR}\right)\approx 2m^{-2}\sum_{i=1}^{m}\left(\sigma_u^2+\sigma^2_{\varepsilon,i}z_i^{-2}\right)^2. \qquad (4.2.14)$$

Note-se que nenhum dos dois estimadores de σ_u^2 apresentados acima exige a normalidade de u_i e de ε_i.

4.2.3.3 Estimador da máxima verosimilhança

O logaritmo da função de verosimilhança sob o modelo de Fay-Herriot (4.2.5) tem a forma:

$$l_{MV}\left(\boldsymbol{\beta},\sigma_u^2\right)=c-\frac{1}{2}\left[\log\left(|\mathbf{V}|\right)+\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right] \qquad (4.2.15)$$

onde *c* é uma constante. Da diferenciação de (4.2.15) em ordem a $\boldsymbol{\beta}$ e a σ_u^2 obtém-se:

$$\frac{\partial l_{MV}\left(\boldsymbol{\beta},\sigma_u^2\right)}{\partial\boldsymbol{\beta}}=\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \qquad (4.2.16)$$

$$\frac{\partial l_{MV}\left(\boldsymbol{\beta},\sigma_u^2\right)}{\partial\sigma_u^2}=\frac{1}{2}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\mathbf{B}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)-tr\left(\mathbf{V}^{-1}\mathbf{B}\right)\right] \qquad (4.2.17)$$

onde $\mathbf{B}=diag_{1\le i\le m}\left(z_i^2\right)$. Usando (4.2.17) e fazendo $\frac{\partial l_{MV}\left(\boldsymbol{\beta},\sigma_u^2\right)}{\partial\sigma_u^2}=0$, obtém-se $\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}\left(\sigma_u^2\right)=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. E substituindo-se em (4.2.15) $\boldsymbol{\beta}$ por $\tilde{\boldsymbol{\beta}}$, obtém-se a seguinte expressão depois de alguma manipulação algébrica

$$l_{MV}\left(\sigma_u^2\right)=c-\frac{1}{2}\left[\log\left(|\mathbf{V}|\right)+\mathbf{y}'\mathbf{P}\mathbf{y}\right]. \qquad (4.2.18)$$

A primeira derivada de (4.2.18) é dada por:

$$\frac{\partial l_{MV}\left(\sigma_u^2\right)}{\partial \sigma_u^2} = \frac{1}{2}\left[\mathbf{y}'\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{y} - tr\left(\mathbf{V}^{-1}\mathbf{B}\right)\right]. \tag{4.2.19}$$

onde $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}$ e $\mathbf{B} = diag_{1\le i\le m}\left(z_i^2\right)$. Finalmente, o estimador da MV de σ_u^2 é dado por $\hat{\sigma}_{u,MV}^2 = \max\left\{0; \tilde{\sigma}_{u,MV}^2\right\}$, onde $\tilde{\sigma}_{u,MV}^2$ é a solução de $\frac{\partial l_{MV}\left(\sigma_u^2\right)}{\partial \sigma_u^2} = 0$. Note-se que, ao contrário do que ocorre nos métodos dos momentos, os estimadores obtidos por métodos de verosimilhança não têm normalmente uma forma "fechada". De acordo com os resultados de Miller (1977), o estimador $\hat{\sigma}_{u,MV}^2$ é consistente e assimptoticamente normal. Datta *et al*. (2005) verificaram que, no âmbito do modelo de Fay-Herriot, o referido estimador é ímpar e invariante a translações.

4.2.3.4 Estimador da máxima verosimilhança restrita

De forma semelhante, o logaritmo da função de verosimilhança restrita sob o modelo de Fay-Herriot (4.2.5) tem a forma

$$l_{MVR}\left(\sigma_u^2\right) = c - \frac{1}{2}\left[\log\left(\left|\mathbf{F}'\mathbf{V}\mathbf{F}\right|\right) + \mathbf{y}'\mathbf{P}\mathbf{y}\right], \tag{4.2.20}$$

onde c é uma constante, $\mathbf{F}$ é qualquer matriz $m\times(m\text{-}p)$ tal que $r(\mathbf{F})=m\text{-}p$ e $\mathbf{F}'\mathbf{X} = \mathbf{0}$, e $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}$. A primeira derivada de (4.2.20) em ordem a σ_u^2 é dada por:

$$\frac{\partial l_{MVR}\left(\sigma_u^2\right)}{\partial \sigma_u^2} = \frac{1}{2}\left[\mathbf{y}'\mathbf{P}\mathbf{B}\mathbf{P}\mathbf{y} - tr\left(\mathbf{P}\mathbf{B}\right)\right], \tag{4.2.21}$$

onde $\mathbf{B} = diag_{1\le i\le m}\left(z_i^2\right)$. Por último, o estimador da MVR de σ_u^2 é dado por $\hat{\sigma}_{u,MVR}^2 = \max\left\{0; \tilde{\sigma}_{u,MVR}^2\right\}$, onde $\tilde{\sigma}_{u,MVR}^2$ é a solução de $\frac{\partial l_{MVR}\left(\sigma_u^2\right)}{\partial \sigma_u^2} = 0$. De acordo com os resultados de Jiang (1996), o estimador da MVR é consistente e assimptoticamente normal. Datta *et al*. (2005) observaram que, no âmbito do modelo de Fay-Herriot, o referido estimador também é ímpar e invariante a translações.

Datta e Lahiri (2000) deduziram as variâncias assimptóticas dos estimadores da MV e da MVR de σ_u^2 apresentados anteriormente, e mostraram que são iguais e dadas por:

$$\overline{V}\left(\hat{\sigma}_{u,MV}^2\right)=\overline{V}\left(\hat{\sigma}_{u,MVR}^2\right)\approx\left[I\left(\sigma_u^2\right)\right]^{-1}=2\left[\sum_{i=1}^{m}z_i^4\left(\sigma_u^2z_i^2+\sigma_{\varepsilon,i}^2\right)^{-2}\right]^{-1}. \qquad (4.2.22)$$

Segundo Rao (2003), o EBLUP $\hat{\theta}_i$ obtido a partir de um estimador de σ_u^2 calculado por um dos quatro métodos expostos acima, permanece não enviesado no modelo se os efeitos aleatórios, u_i, e os erros da sondagem, ε_i, estiverem distribuídos de forma simétrica à volta do zero. Em particular, o EBLUP será também não enviesado no modelo se os u_i's e os ε_i's forem normalmente distribuídos.

Por último, é ainda de salientar que os quatro métodos de estimação de componentes de variância apresentados podem produzir estimativas negativas, especialmente quando o número de domínios é pequeno. Uma estimativa da componente de variância nula (após a truncagem a zero) tem como consequência a redução do EBLUP de θ_i a um estimador sintético pela regressão (*vide* a discussão de Morris em Jiang e Lahiri (2006)). Para além disso, uma estimativa nula da componente de variância gera problemas ao nível da predição *bootstrap* paramétrica por intervalos de confiança. Para obviar esta situação, Li (2007) propôs um novo estimador consistente e estritamente positivo de σ_u^2 baseado num método de densidade ajustado.

4.2.4 Aproximação analítica do Erro Quadrático Médio de Predição (EQMP) do EBLUP

Uma medida de incerteza associada ao EBLUP, $\hat{\theta}_i$, é dada pelo seu EQMP, que pode ser decomposto em três componentes, tal como foi apresentado na secção 3.2.5 (com base nos trabalhos de Kackar e Harville (1984) e utilizando os resultados gerais de Henderson (1975)):

$$EQMP\left(\hat{\theta}_i\right)=g_{1i}\left(\sigma_u^2\right)+g_{2i}\left(\sigma_u^2\right)+E\left(\hat{\theta}_i-\tilde{\theta}_i\right)^2, \qquad (4.2.23)$$

onde E representa o valor esperado relativo ao modelo (4.2.4).

No contexto do modelo de Fay-Herriot (4.2.4), as primeiras duas parcelas são dadas por (Prasad e Rao, 1990; Ghosh e Rao, 1994):

$$g_{1i}\left(\sigma_u^2\right)=\frac{\sigma_u^2 z_i^2 \sigma_{\varepsilon,i}^2}{\sigma_u^2 z_i^2+\sigma_{\varepsilon,i}^2}=\gamma_i \sigma_{\varepsilon,i}^2, \tag{4.2.24}$$

$$g_{2i}\left(\sigma_u^2\right)=\left(1-\gamma_i\right)^2 \mathbf{x}_i'\left[\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i'\left(\sigma_u^2 z_i^2+\sigma_{\varepsilon,i}^2\right)^{-1}\right]^{-1}\mathbf{x}_i. \tag{4.2.25}$$

Contudo, geralmente não existe uma expressão "fechada" para a terceira parcela de (4.2.23). Prasad e Rao (1990) e Datta e Lahiri (2000) deduziram uma aproximação de segunda ordem para essa parcela pelo método delta, ignorando todos os termos de ordem superior a $o\left(m^{-1}\right)$, quando $m\to\infty$. Supondo a normalidade de u_i e de ε_i, e assumindo certas condições de regularidade, uma aproximação de segunda ordem para essa parcela é dada por (Prasad e Rao, 1990; Datta e Lahiri, 2000):

$$g_{3i}\left(\sigma_u^2\right)=\left[\sigma_{\varepsilon,i}^4 z_i^4\left(\sigma_u^2 z_i^2+\sigma_{\varepsilon,i}^2\right)^{-3}\right]\overline{V}\left(\hat{\sigma}_u^2\right), \tag{4.2.26}$$

onde $\overline{V}\left(\hat{\sigma}_u^2\right)$ é a variância assimptótica de $\hat{\sigma}_u^2$. Se σ_u^2 for estimada pelo método dos momentos de Fay-Herriot, então $\overline{V}\left(\hat{\sigma}_u^2\right)=\overline{V}\left(\hat{\sigma}_{u,FH}^2\right)$ é dada por (4.2.12). Se for usado o estimador de Prasad-Rao $\hat{\sigma}_{u,PR}^2$, então $\overline{V}\left(\hat{\sigma}_u^2\right)=\overline{V}\left(\hat{\sigma}_{u,PR}^2\right)$ é dada por (4.2.14). Tem-se ainda que $\overline{V}\left(\hat{\sigma}_u^2\right)=\overline{V}\left(\hat{\sigma}_{u,MV}^2\right)=\overline{V}\left(\hat{\sigma}_{u,MVR}^2\right)$ é dada por (4.2.22), quando se estima σ_u^2 através de métodos de verosimilhança. Com base nos resultados (4.2.12), (4.2.14) e (4.2.22) é fácil observar que $g_{3i,MV}\left(\sigma_u^2\right)=g_{3i,MVR}\left(\sigma_u^2\right)\le g_{3i,FH}\left(\sigma_u^2\right)\le g_{3i,PR}\left(\sigma_u^2\right)$.

4.2.5 Estimação do EQMP do EBLUP

Se a estimação de σ_u^2 for efectuada pelo método da MVR ou pelo método dos momentos de Prasad-Rao, sob certas condições de regularidade e assumindo a normalidade de u_i e de ε_i, então um estimador analítico não enviesado até à segunda ordem do EQMP do EBLUP é dado por (Prasad e Rao, 1990; Datta e Lahiri, 2000):

$$eqmp\left(\hat{\theta}_i\right)=g_{1i}\left(\hat{\sigma}_u^2\right)+g_{2i}\left(\hat{\sigma}_u^2\right)+2g_{3i}\left(\hat{\sigma}_u^2\right). \tag{4.2.27}$$

Se a estimação de σ_u^2 for efectuada pelo método da MV (Datta e Lahiri, 2000) ou pelo método dos momentos de Fay-Herriot (Datta *et al.*, 2005), então um estimador analítico não enviesado até à segunda ordem do EQMP do EBLUP é dado por:

$$eqmp^*\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_u^2\right) + g_{2i}\left(\hat{\sigma}_u^2\right) + 2g_{3i}\left(\hat{\sigma}_u^2\right) - g_{4i}\left(\hat{\sigma}_u^2\right). \tag{4.2.28}$$

onde $g_{4i}\left(\hat{\sigma}_u^2\right) = c_{\hat{\sigma}_u^2}\left(\hat{\sigma}_u^2\right)\nabla g_{1i}\left(\hat{\sigma}_u^2\right)$ com $\nabla g_{1i}\left(\hat{\sigma}_u^2\right) = z_i^2\left(1-\hat{\gamma}_i\right)^2$. Uma aproximação de segunda ordem do termo de enviesamento, $c_{\hat{\sigma}_u^2}\left(\hat{\sigma}_u^2\right)$, no caso da componente de variância ser estimada pelo método da MV, é dada por (Datta e Lahiri, 2000):

$$c_{\hat{\sigma}_u^2=\hat{\sigma}_{u,MV}^2}\left(\hat{\sigma}_u^2\right) \approx -\left[2I\left(\sigma_u^2\right)\right]^{-1} tr\left[\left\{\sum_{i=1}^{m}\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-1}\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\left\{\sum_{i=1}^{m} z_i^2\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-2}\mathbf{x}_i\mathbf{x}_i'\right\}\right], \tag{4.2.29}$$

onde $I\left(\sigma_u^2\right) = \frac{1}{2}\sum_{i=1}^{m}\frac{z_i^4}{\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^2}$. A partir de (4.2.29) conclui-se que o termo $-g_{4i}\left(\hat{\sigma}_u^2\right)$ da expressão (4.2.28) é positivo. Desta forma, se este termo for ignorado e for utilizada a expressão (4.2.27) na estimação do EQMP do EBLUP, com $\hat{\sigma}_u^2 = \hat{\sigma}_{u,MV}^2$, então verifica-se uma subestimação do EQMP.

Segundo Datta *et al.* (2005), uma aproximação de segunda ordem do termo de enviesamento, $c_{\hat{\sigma}_u^2}\left(\hat{\sigma}_u^2\right)$, no caso da componente de variância ser estimada pelo método dos momentos de Fay-Herriot, é dada por:

$$c_{\hat{\sigma}_u^2=\hat{\sigma}_{u,FH}^2}\left(\hat{\sigma}_u^2\right) \approx 2\left[m\sum_{i=1}^{m}\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-2} - \left\{\sum_{i=1}^{m}\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-1}\right\}^2\right]\left[\sum_{i=1}^{m}\left(\sigma_u^2 z_i^2 + \sigma_{\varepsilon,i}^2\right)^{-1}\right]^{-3} \tag{4.2.30}$$

Segundo aqueles autores, neste caso o termo $-g_{4i}\left(\hat{\sigma}_u^2\right)$ da expressão (4.2.28) é negativo. Desta forma, se este termo for ignorado e for utilizada a expressão (4.2.27) na estimação do EQMP do EBLUP, com $\hat{\sigma}_u^2 = \hat{\sigma}_{u,FH}^2$, então verifica-se uma sobreestimação do EQMP.

É de salientar que nenhum dos estimadores do EQMP do EBLUP depende das estimativas directas do parâmetro de interesse. Note-se, também, que na derivação dos

estimadores do EQMP do EBLUP se assumiu não só a linearidade do modelo, mas também a normalidade dos efeitos aleatórios, u_i. Contudo, Lahiri e Rao (1995) apresentaram uma versão mais robusta do método de Prasad e Rao (1990), mostrando que o estimador do EQMP do EBLUP (4.2.27), sob o modelo de Fay-Herriot, continua a ser válido mesmo na situação em que os efeitos aleatórios não sejam normais, com $E|u_i|^{8+\delta} < \infty$ para $0 < \delta < 1$. Este resultado foi demonstrado assumindo apenas a hipótese da normalidade dos erros de sondagem, ε_i, sendo a componente de variância estimada pelo método dos momentos de Prasad-Rao.

Para além dos estimadores analíticos do EQMP do EBLUP, ainda existem os estimadores baseados em métodos por reamostragem, tal como apresentado na secção 3.2.5. Quando é utilizado o método *jackknife* introduzido por Jiang *et al.* (2002) para estimar o EQMP do EBLUP no contexto do modelo de Fay-Herriot, qualquer que seja o método de estimação de σ_u^2, o estimador corrigido até à segunda ordem é dado por:

$$eqmp^{JLW}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_u^2\right) - \frac{m-1}{m}\sum_{e=1}^{m}\left[g_{1i}\left(\hat{\sigma}_{u,-e}^2\right) - g_{1i}\left(\hat{\sigma}_u^2\right)\right] + \frac{m-1}{m}\sum_{e=1}^{m}\left(\hat{\theta}_{i,-e} - \hat{\theta}_i\right)^2, \quad (4.2.31)$$

onde $g_{1i}\left(\sigma_u^2\right)$ é dado por (4.2.24), $g_{1i}\left(\hat{\sigma}_{u,-e}^2\right) = \hat{\gamma}_{i,-e}\sigma_{\varepsilon,i}^2$ e $\hat{\theta}_{i,-e} = \hat{\theta}_{i,-e}\left(\hat{\sigma}_{u,-e}^2\right) = \hat{\gamma}_{i,-e}y_i + \left(1-\hat{\gamma}_{i,-e}\right)\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{-e}$, com $\hat{\gamma}_{i,-e} = \dfrac{\hat{\sigma}_{u,-e}^2 z_i^2}{\hat{\sigma}_{u,-e}^2 z_i^2 + \sigma_{\varepsilon,i}^2}$. Neste caso, $\hat{\sigma}_{u,-e}^2$ e $\hat{\boldsymbol{\beta}}_{-e}$ são estimados com os dados de todos os domínios, com excepção do *e*-ésimo domínio (*e*=1, ..., *m*).

Quando é utilizada a aproximação em série de Taylor do estimador *jackknife* ponderado introduzida por Chen e Lahiri (2005, 2008), então o estimador corrigido até à segunda ordem do EQMP do EBLUP assume diferentes formas, consoante o método utilizado para estimar σ_u^2. Se for utilizado o método ANOVA ou o método da MVR, então tem-se o seguinte estimador:

$$eqmp^{CL1}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_u^2\right) + g_{2i}\left(\hat{\sigma}_u^2\right) + \frac{\sigma_{\varepsilon,i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{\varepsilon,i}^2\right)^3}\hat{v}_{WJ}\left(\hat{\sigma}_u^2\right) + \frac{\sigma_{\varepsilon,i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{\varepsilon,i}^2\right)^4}\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)^2\hat{v}_{WJ}\left(\hat{\sigma}_u^2\right), \quad (4.2.32)$$

onde $g_{1i}\left(\sigma_u^2\right)$ é dado por (4.2.24), $g_{2i}\left(\sigma_u^2\right)$ é dado por (4.2.25), e $\hat{v}_{WJ} = \sum_{e=1}^{m} w_e\left(\hat{\sigma}_{u,-e}^2 - \hat{\sigma}_u^2\right)^2$ é um estimador *jackknife* ponderado da variância de $\hat{\sigma}_u^2$.

Se σ_u^2 for estimado pelo método de Fay-Herriot ou pelo método da MV, então tem-se o seguinte estimador (Chen e Lahiri, 2005, 2008):

$$eqmp^{CL2}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_u^2\right) + g_{2i}\left(\hat{\sigma}_u^2\right) - \frac{\sigma_{\varepsilon,i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{\varepsilon,i}^2\right)^2}\hat{c}_{WJ}\left(\hat{\sigma}_u^2\right) + \frac{\sigma_{\varepsilon,i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{\varepsilon,i}^2\right)^3}\hat{v}_{WJ}\left(\hat{\sigma}_u^2\right) + \\ + \frac{\sigma_{\varepsilon,i}^4}{\left(\hat{\sigma}_u^2 + \sigma_{\varepsilon,i}^2\right)^4}\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\hat{v}_{WJ}\left(\hat{\sigma}_u^2\right), \qquad (4.2.33)$$

onde $\hat{c}_{WJ}\left(\hat{\sigma}_u^2\right) = \sum_{e=1}^{m} w_e\left(\hat{\sigma}_{u,-e}^2 - \hat{\sigma}_u^2\right)$ é um estimador *jackknife* ponderado do enviesamento de $\hat{\sigma}_u^2$, *i.e.*, de $E\left(\hat{\sigma}_u^2\right) - \sigma_u^2$.

Por último, a estimação do EQMP do EBLUP sob o modelo de Fay-Herriot pode ainda ser efectuada pelo método *bootstrap* paramétrico introduzido também na secção 3.2.5. Neste caso, é necessário gerar *B* conjuntos de dados *bootstrap* $y_i^{(b)}$, b=1, …, B para todos os pequenos domínios, i=1, …, m, sob o modelo:

$$y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + u_i + \varepsilon_i, \qquad (4.2.34)$$

onde $\hat{\boldsymbol{\beta}}^{(b)}$ é dado por (4.2.7) e os erros do modelo são independentes satisfazendo $u_i \overset{iid}{\sim}\left(0, \hat{\sigma}_u^2\right)$ e $\varepsilon_i \overset{iid}{\sim}\left(0, \sigma_{\varepsilon,i}^2\right)$, i=1, …, m. Para cada b-ésimo conjunto de dados, é estimada a componente de variância, $\hat{\sigma}_u^{2(b)}$, substituindo y_i por $y_i^{(b)}$ no respectivo estimador de σ_u^2; são estimados os efeitos fixos, $\hat{\boldsymbol{\beta}}^{(b)}$, substituindo σ_u^2 por $\hat{\sigma}_u^{2(b)}$ em (4.2.7); e posteriormente estimados $\hat{\theta}_i\left(\hat{\sigma}_u^{2(b)}\right)$, $g_{1i}\left(\hat{\sigma}_u^{2(b)}\right)$ e $g_{2i}\left(\hat{\sigma}_u^{2(b)}\right)$. Tem-se finalmente que o estimador *bootstrap* paramétrico é dado por (Butar e Lahiri, 2003):

$$eqmp^{BOOT}\left(\hat{\theta}_i\right) = g_{1i}\left(\hat{\sigma}_u^2\right) + g_{2i}\left(\hat{\sigma}_u^2\right) - \frac{1}{B}\sum_{b=1}^{B}\left[g_{1i}\left(\hat{\sigma}_u^{2(b)}\right) + g_{2i}\left(\hat{\sigma}_u^{2(b)}\right) - g_{1i}\left(\hat{\sigma}_u^2\right) - g_{2i}\left(\hat{\sigma}_u^2\right)\right] + \\ + \frac{1}{B}\sum_{b=1}^{B}\left[\hat{\theta}_i\left(\hat{\sigma}_u^{2(b)}\right) - \hat{\theta}_i\left(\hat{\sigma}_u^2\right)\right]^2, \qquad (4.2.35)$$

onde $g_{1i}\left(\sigma_u^2\right)$ é dado por (4.2.24) e $g_{2i}\left(\sigma_u^2\right)$ é dado por (4.2.25).

4.3 MODELO BÁSICO DE NÍVEL ÁREA COM DADOS SECCIONAIS E CRONOLÓGICOS

4.3.1 Introdução

O estimador EBLUP assistido pelo modelo de Fay-Herriot utiliza unicamente dados amostrais da variável de interesse referentes a um único período de tempo. Desta forma, não explora a informação referente a outros períodos de tempo, disponível na situação de uma sondagem repetida no tempo. Segundo Rao (2003), na situação de inquéritos repetidos no tempo, podem ser obtidos ganhos de eficiência significativos com a utilização de informação de outros pequenos domínios e de outros períodos de tempo.

Apesar da maioria da investigação efectuada sobre estimação em pequenos domínios, no âmbito dos modelos de nível área, ter estado inicialmente centrada na utilização de dados seccionais, actualmente é possível encontrar na literatura várias aplicações de modelos que combinam dados de natureza seccional e cronológica, dado que é cada vez mais frequente a realização de inquéritos repetidos no tempo. Neste contexto, existindo informação relativa à variável de interesse e a variáveis auxiliares em diversos momentos no tempo, a estimação de relações entre essas variáveis com o objectivo de melhorar as propriedades dos estimadores em pequenos domínios parece muito interessante em várias áreas do conhecimento. Choudhry e Rao (1989), Pfeffermann e Burck (1990), Rao e Yu (1992, 1994), Yu (1993), Ghosh e Nangia (1993), Singh *et al.* (1994), Ghosh *et al.* (1996), Datta *et al.* (1997, 1999, 2002), You *et al.* (2001) e Saei e Chambers (2003*a*) são alguns dos autores que utilizaram modelos com dados de natureza seccional e cronológica para fazerem estimação em pequenos domínios. Estes modelos podem ser classificados em dois grandes grupos: os modelos do tipo do modelo de Rao-Yu e os modelos do tipo *state space*.

Nas secções seguintes apresenta-se uma breve revisão da literatura dos principais modelos que utilizam dados seccionais e cronológicos, e que se resumem a casos

particulares do modelo linear misto. No âmbito do modelo de estimação em pequenos domínios de Rao-Yu, são também propostos dois métodos por reamostragem para estimação do EQMP do EBLUP temporal.

4.3.2 Especificação do modelo de Rao-Yu

Nesta secção é apresentada uma extensão do modelo de Fay-Herriot para dados seccionais e cronológicos, contemplando efeitos aleatórios de domínio-tempo modelados através de um processo auto-regressivo de primeira ordem [AR(1)], proposta por Rao e Yu (1992, 1994). Sejam θ_{it} e y_{it}, respectivamente, um parâmetro populacional da variável de interesse e o seu estimador directo não enviesado no desenho associados ao *i*-ésimo domínio no período *t* (*i*=1,..., *m*; *t*=1,..., *T*), e $\mathbf{x}_{it} = \left(x_{it1}, ..., x_{itp}\right)'$ um vector $p\times 1$ de variáveis explicativas referenciado à mesma unidade. O modelo de erro da sondagem é dado por:

$$y_{it} = \theta_{it} + \varepsilon_{it}, \tag{4.3.1}$$

onde ε_{it}' são os erros da sondagem satisfazendo $\varepsilon_{it} \mid \theta_{it} \overset{iid}{\sim} N\left(0, \sigma_{it}^2\right)$. Rao e Yu (1994) propuseram que o parâmetro de interesse esteja ligado ao vector de variáveis explicativas através do seguinte modelo de ligação:

$$\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_i + u_{it}, \tag{4.3.2}$$

onde $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)'$ é um vector $p\times 1$ de parâmetros de regressão, v_i são os efeitos aleatórios específicos de domínio satisfazendo $v_i \overset{iid}{\sim} N\left(0; \sigma_v^2\right)$ e os u_{it}'s são os efeitos aleatórios específicos de domínio-tempo, os quais incorporam a estrutura de dependência temporal do processo. Neste caso, a inclusão de correlações temporais entre estes últimos efeitos aleatórios é efectuada através do seguinte processo AR(1):

$$u_{it} = \rho u_{i,t-1} + \xi_{it}, \quad |\rho| < 1 \tag{4.3.3}$$

onde ξ_{it}'s são os erros do processo a satisfazer $\xi_{it} \overset{iid}{\sim} N(0;\sigma^2)$ e ρ é uma medida do nível de autocorrelação temporal. O modelo (4.3.2) especifica que os θ_{it}'s dependem de efeitos aleatórios específicos de domínio, v_i, e de efeitos aleatórios associados ao domínio e ao período de tempo, u_{it}, os quais estão correlacionados ao longo do tempo para cada domínio. Assume-se também que os $\{\xi_{it}\}$, $\{\varepsilon_{it}\}$ e $\{v_i\}$ são independentes uns dos outros. O modelo combinado baseado em (4.3.1)-(4.3.3) é dado por:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + v_i + u_{it} + \varepsilon_{it} \\ u_{it} &= \rho u_{i,t-1} + \xi_{it}, \quad |\rho| < 1. \end{aligned} \tag{4.3.4}$$

Note-se que o conhecido modelo de Fay-Herriot (4.2.3) pode ser obtido a partir do modelo (4.3.4) fazendo T=1, ρ=0 e $\sigma^2 = 0$. Rao e Yu (1994), procederam à ordenação das estimativas directas do parâmetro da variável de interesse, $\mathbf{y} = col_{1\le i\le m}(\mathbf{y}_i)$ e $\mathbf{y}_i = col_{1\le t\le T}(y_{it})$, de forma a que o modelo apresentado em (4.3.4) pudesse ser escrito na forma compacta como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \tag{4.3.5}$$

com $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$, $\mathbf{X}_i = col_{1\le t\le T}(\mathbf{x}_{it}')$, $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{I}_{mT}]$, $\mathbf{Z}_1 = \mathbf{I}_m \otimes \mathbf{1}_T$, $\boldsymbol{\upsilon} = [\mathbf{v}' \quad \mathbf{u}']'$, $\mathbf{v} = col_{1\le i\le m}(v_i)$, $\mathbf{u} = col_{1\le i\le m}(\mathbf{u}_i)$, $\mathbf{u}_i = col_{1\le t\le T}(u_{it})$, $\boldsymbol{\varepsilon} = col_{1\le i\le m}(\boldsymbol{\varepsilon}_i)$ e $\boldsymbol{\varepsilon}_i = col_{1\le t\le T}(\varepsilon_{it})$. Assume-se também que $\mathbf{v}$, $\mathbf{u}$ e $\boldsymbol{\varepsilon}$ são mutuamente independentes, com $\mathbf{v} \sim N(\mathbf{0};\sigma_v^2\mathbf{I}_m)$, $\mathbf{u} \sim N(\mathbf{0};\sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma})$ e $\boldsymbol{\varepsilon} \sim N(\mathbf{0};\mathbf{R})$, onde $\boldsymbol{\Gamma} = \{\gamma_{rs}\}$ é uma matriz $T \times T$ com elementos $\gamma_{rs} = \rho^{|r-s|}/(1-\rho^2)$, r, s=1, ..., T. Tem-se ainda que $\mathbf{R} = diag_{1\le i\le m}(\mathbf{R}_i)$, onde $\mathbf{R}_i = diag_{1\le t\le T}(\sigma_{it}^2)$. É fácil verificar que o modelo (4.3.5) é um caso particular do modelo linear misto (3.2.1) com uma matriz de covariâncias de $\mathbf{y}$ diagonal por blocos com a seguinte estrutura $\mathbf{V} = diag_{1\le i\le m}(\mathbf{V}_i)$, onde $\mathbf{V}_i = \mathbf{R}_i + \sigma^2\boldsymbol{\Gamma} + \sigma_v^2\mathbf{J}_T$.

4.3.3 O EBLUP

Rao e Yu (1994) obtiveram a expressão do BLUP de θ_{it} a partir dos resultados gerais obtidos por Henderson (1975), uma vez que o modelo (4.3.5) é um caso especial do

modelo linear misto. Assumindo que o vector de componentes de variância, $\boldsymbol{\psi} = \left(\sigma^2, \sigma_v^2, \rho\right)'$, é conhecido, então o BLUP temporal de θ_{it} é dado por:

$$\tilde{\theta}_{it} = \tilde{\theta}_{it}^{H}\left(\boldsymbol{\psi}\right) = \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}} + \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right)' \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}\right), \tag{4.3.6}$$

onde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\boldsymbol{\psi}\right) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ é o estimador dos mínimos quadrados generalizados de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}_t$ é a *t*-ésima linha da matriz $\boldsymbol{\Gamma}$. De acordo com Rao e Yu (1994), o BLUP de θ_{it} pode também ser escrito como um estimador combinado, sendo uma soma ponderada do estimador directo, y_{it}, do estimador sintético, $\mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}$, e dos resíduos $\left(y_{is} - \mathbf{x}_{is}'\tilde{\boldsymbol{\beta}}\right)$, *s*=1, 2, ..., *T*-1:

$$\tilde{\theta}_{it} = w_{it}^{*} y_{it} + \left(1 - w_{it}^{*}\right)\mathbf{x}_{it}'\tilde{\boldsymbol{\beta}} + \sum_{s=1}^{T-1} w_{it}^{*}\left(y_{is} - \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}\right), \tag{4.3.7}$$

onde $\left(w_{i1}^{*}, ..., w_{iT}^{*}\right) = \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right)' \mathbf{V}_i^{-1}$.

O BLUP temporal de θ_{it} depende das componentes de variância, $\boldsymbol{\psi}$, as quais são geralmente desconhecidas na prática. Assumindo que ρ é conhecido no modelo AR(1), Rao e Yu (1994) obtiveram o preditor em dois passos de θ_{it} fazendo a substituição em (4.3.6) de σ^2 e σ_v^2 por estimadores assimptoticamente consistentes, $\hat{\sigma}^2(\rho)$ e $\hat{\sigma}_v^2(\rho)$, respectivamente. O EBLUP temporal resultante é então dado por:

$$\hat{\theta}_{it} = \hat{\theta}_{it}^{H}(\rho) = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}(\rho) + \left[\hat{\sigma}_v^2(\rho)\mathbf{1}_T + \hat{\sigma}^2(\rho)\boldsymbol{\gamma}_t\right]' \hat{\mathbf{V}}_i^{-1}(\rho)\left[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\rho)\right], \tag{4.3.8}$$

onde $\hat{\boldsymbol{\beta}}(\rho)$ e $\hat{\mathbf{V}}_i^{-1}(\rho)$ são os estimadores de $\boldsymbol{\beta}(\rho)$ e de $\mathbf{V}_i^{-1}(\rho)$, respectivamente, quando σ^2 e σ_v^2 são substituídos por $\hat{\sigma}^2(\rho)$ e $\hat{\sigma}_v^2(\rho)$, respectivamente.

4.3.4 Estimação das componentes de variância

Rao e Yu (1994) propuseram uma extensão do método III de Henderson (1953) para a estimação das componentes de variância, σ^2 e σ_v^2, no contexto de um modelo, (4.3.4), com erros autocorrelacionados, u_{it}, erros da sondagem independentes, ε_{it}, mas

assumindo que ρ é conhecido. Por esta razão, a partir deste ponto, define-se o vector de componentes de variância como $\boldsymbol{\psi} = \left[\sigma^2(\rho), \sigma_v^2(\rho)\right]'$ no contexto do modelo de Rao-Yu.

Os estimadores das componentes de variância são baseados em regressões pelo método dos mínimos quadrados ordinários efectuadas sobre o modelo de Rao-Yu transformado. Em primeiro lugar, transforma-se o modelo (4.3.5) no modelo reduzido $\mathbf{z}^{(1)} = \mathbf{H}^{(1)}\boldsymbol{\beta} + \mathbf{e}^{(1)}$ e denota-se por $\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}$ a soma dos quadrados dos resíduos obtidos pela regressão de $\mathbf{z}^{(1)}$ sobre $\mathbf{H}^{(1)}$, onde $\mathbf{z}^{(1)} = col_{1\le i\le m}\left(\mathbf{z}_i^{(1)}\right)$, $\mathbf{H}^{(1)} = col_{1\le i\le m}\left(\mathbf{H}_i^{(1)}\right)$, $\mathbf{z}_i^{(1)} = (\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{y}_i$, $\mathbf{H}_i^{(1)} = (\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{X}_i$, $\mathbf{D} = (\mathbf{f}\mathbf{f}')/c$, $\mathbf{f} = col_{1\le t\le T}(f_t)$ com $f_1 = \left(1-\rho^2\right)^{1/2}$ e $f_t = 1-\rho$ para $2 \le t \le T$, $c = \mathbf{f}'\mathbf{f}$ e $\mathbf{P}$ ($T\times T$) tem a seguinte forma: $p_{11} = \left(1-\rho^2\right)^{1/2}$, $p_{t,t'} = 1, \forall t = t'$ para $t, t' = 2, ..., T$, $p_{t+1,t} = -\rho$ para $t = 1, ..., T-1$ e os restantes elementos são $p_{t,t'} = 0$. Em segundo lugar, transforma-se o modelo (4.3.4) no seguinte modelo $c^{-1/2}\mathbf{f}'\mathbf{z}_i = c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{X}_i\boldsymbol{\beta} + c^{1/2}v_i + u_i^* + c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{e}_i$ e denota-se por $\hat{\mathbf{e}}^{(2)\prime}\hat{\mathbf{e}}^{(2)}$ a soma dos quadrados dos resíduos obtidos pela regressão de $\mathbf{z}^{(2)}$ sobre $\mathbf{H}^{(2)}$, onde $\mathbf{z}^{(2)} = col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{y}_i\right)$, $\mathbf{H}^{(2)} = col_{1\le i\le m}(c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{X}_i)$ e $u_i^* = c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{u}_i$. Rao e Yu (1994) obtiveram então os seguintes estimadores não enviesados de σ^2 e σ_v^2, respectivamente:

$$\tilde{\sigma}^2(\rho) = \left\{\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} - tr\left[\left(diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D}) - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}\right)diag_{1\le i\le m}(\mathbf{P}\mathbf{R}_i\mathbf{P}')\right]\right\}\times$$
$$\times\left[m(T-1) - r\left(\mathbf{H}^{(1)}\right)\right]^{-1}, \qquad (4.3.9)$$

$$\tilde{\sigma}_v^2(\rho) = \left\{\hat{\mathbf{e}}^{(2)\prime}\hat{\mathbf{e}}^{(2)} - tr\left[\left(\mathbf{I}_m - \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)\prime}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)\prime}\right)diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]\right\}\times$$
$$\times c^{-1}\left[m - r\left(\mathbf{H}^{(2)}\right)\right]^{-1} - c^{-1}\tilde{\sigma}^2(\rho)^{-1}, \qquad (4.3.10)$$

onde $\mathbf{A}^{-}$ representa a inversa generalizada de Moore-Penrose[44] de uma matriz $\mathbf{A}$. Como os estimadores (4.3.9) e (4.3.10) podem assumir valores negativos, Rao e Yu

[44] A matriz inversa generalizada é muitas vezes denominada por matriz inversa condicional, matriz pseudo-inversa ou matriz *g*-inversa (Peterson e Pederson, 2008).

(1994) truncaram esses estimadores a zero: $\hat{\sigma}^2(\rho) = \max\{0; \tilde{\sigma}^2(\rho)\}$ e $\hat{\sigma}_v^2(\rho) = \max\{0; \tilde{\sigma}_v^2(\rho)\}$. Segundo estes autores, os estimadores truncados, $\hat{\sigma}^2(\rho)$ e $\hat{\sigma}_v^2(\rho)$, passam a ser enviesados, mas continuam a ser assimptoticamente consistentes quando $m \to \infty$.

Por último, Rao e Yu (1994) observaram que o EBLUP temporal (4.3.8) é um estimador não enviesado no modelo pelo facto de $\hat{\sigma}^2(\rho)$ e $\hat{\sigma}_v^2(\rho)$ serem funções ímpares em **y** e invariantes a translações.

4.3.5 Aproximação analítica do EQMP do EBLUP

Sob a hipótese da normalidade dos efeitos aleatórios e dos erros da sondagem, então o EQMP do EBLUP temporal pode ser decomposto como (com base nos trabalhos de Kackar e Harville (1984) e utilizando os resultados gerais de Henderson (1975)):

$$EQMP\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\boldsymbol{\psi}) + g_{2it}(\boldsymbol{\psi}) + E\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_{it}(\boldsymbol{\psi})\right]^2, \qquad (4.3.11)$$

onde *E* representa o valor esperado relativo ao modelo (4.3.4), $g_{1it}(\boldsymbol{\psi})$ representa a incerteza presente no EBLUP temporal devida à estimação dos efeitos aleatórios e é de ordem $o(1)$, $g_{2it}(\boldsymbol{\psi})$ mede a incerteza devida à estimação dos efeitos fixos e é de ordem $o(m^{-1})$, e o último termo mede a incerteza relativa à estimação das componentes de variância. Enquanto os dois primeiros termos podem ser analiticamente avaliados a partir das seguintes expressões "fechadas" sem exigirem a normalidade dos erros do modelo (Rao e Yu, 1994),

$$g_{1it}(\boldsymbol{\psi}) = \sigma_v^2 + \frac{\sigma^2}{1-\rho^2} - \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right)' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right), \qquad (4.3.12)$$

$$g_{2it}(\boldsymbol{\psi}) = \left[\mathbf{x}_{it} - \mathbf{X}_i' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right)\right]' \left(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \left[\mathbf{x}_{it} - \mathbf{X}_i' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t\right)\right], \qquad (4.3.13)$$

o último termo do membro direito da expressão (4.3.11) não pode ser analiticamente avaliado, sendo por este motivo necessária uma aproximação. Com base na aproximação em séries de Taylor de Kackar e Harvile (1984) e nos desenvolvimentos

do método delta apresentados por Prasad e Rao (1990), Rao e Yu (1994) obtiveram uma aproximação analítica desse termo de ordem $o\left(m^{-1}\right)$, sob determinadas condições de regularidade e assumindo que ρ é conhecido, dada por:

$$E\left[\hat{\theta}_{it}(\hat{\psi})-\tilde{\theta}_{it}(\psi)\right]^2 \approx tr\left(\mathbf{A}_{it}\boldsymbol{\Sigma}^*\right)= g_{3it}(\psi), \tag{4.3.14}$$

onde $\boldsymbol{\Sigma}^*$ é uma matriz 2×2 de covariâncias dos estimadores não enviesados das componentes de variância, $\tilde{\sigma}^2(\rho)$ e $\tilde{\sigma}_v^2(\rho)$, e $\mathbf{A}_{it}=\{a_{kl}\}$ é uma matriz simétrica da mesma dimensão com elementos principais $a_{11}=\left[\boldsymbol{\gamma}_t-\boldsymbol{\Gamma}\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]'\mathbf{V}_i^{-1}\left[\boldsymbol{\gamma}_t-\boldsymbol{\Gamma}\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]$ e $a_{22}=\left[\mathbf{1}_t-\mathbf{J}_T\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]'\mathbf{V}_i^{-1}\left[\mathbf{1}_t-\mathbf{J}_T\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]$, e com elementos opostos $a_{12}=a_{21}=\left[\boldsymbol{\gamma}_t-\boldsymbol{\Gamma}\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]'\mathbf{V}_i^{-1}\left[\mathbf{1}_t-\mathbf{J}_T\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_t\right)\right]$.

Segundo Rao e Yu (1994), a avaliação dos elementos de $\boldsymbol{\Sigma}^*$, $V\left[\tilde{\sigma}^2(\rho)\right]$, $V\left[\tilde{\sigma}_v^2(\rho)\right]$ e $Cov\left[\tilde{\sigma}^2(\rho);\tilde{\sigma}_v^2(\rho)\right]$, é baseada na avaliação da covariância de duas formas quadráticas de variáveis normalmente distribuídas. Se as componentes de variância forem estimadas pelo método dos momentos descrito na secção 4.3.4, então podem ser reescritas como formas quadráticas da seguinte forma (Rao e Yu, 1994):

$$\tilde{\sigma}^2(\rho)=\left\{(m-1)T-r\left(\mathbf{H}^{(1)}\right)\right\}^{-1}\mathbf{a}'\mathbf{C}_1\mathbf{a}+const\,, \tag{4.3.15}$$

$$\tilde{\sigma}_v^2(\rho)=c^{-1}\left\{m-r\left(\mathbf{H}^{(2)}\right)\right\}^{-1}\mathbf{a}'\mathbf{C}_2\mathbf{a}-c^{-1}\left\{(m-1)T-r\left(\mathbf{H}^{(1)}\right)\right\}^{-1}\mathbf{a}'\mathbf{C}_1\mathbf{a}+const\,, \tag{4.3.16}$$

onde $\mathbf{a}=\mathbf{Z}\mathbf{v}+\mathbf{u}+\mathbf{e}\sim N(\mathbf{0};\mathbf{V})$, $\mathbf{C}_1=\mathbf{C}'\left\{\mathbf{I}_{mT}-\mathbf{C}\mathbf{X}\left(\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}'\right\}\mathbf{C}$ com $\mathbf{C}=diag_{1\le i\le m}\left[(\mathbf{I}_T-\mathbf{D})\mathbf{P}\right]$, e $\mathbf{C}_2=\mathbf{C}^{*\prime}\left\{\mathbf{I}_m-\mathbf{C}^*\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{*\prime}\mathbf{C}^*\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{*\prime}\right\}\mathbf{C}^*$ com $\mathbf{C}^*=diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\right)$.

Nesta situação, Rao e Yu (1994) propuseram um estimador analítico do EQMP do EBLUP temporal, aproximadamente não enviesado até à ordem $o\left(m^{-1}\right)$, sob a hipótese de um pequeno número de períodos de tempo e um grande número de pequenos domínios, dado por:

$$eqmp^{RY}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) + 2g_{3it}(\hat{\boldsymbol{\psi}}). \qquad (4.3.17)$$

4.3.6 Aproximação *bootstrap* do EQMP do EBLUP

Nesta secção é proposta uma metodologia para obter uma aproximação do EQMP do EBLUP temporal através de um procedimento *bootstrap*. Relembre-se que a expressão do EQMP, dada em (4.3.11), envolve expressões exactas para as parcelas g_{1it} e g_{2it}. Contudo, a parcela g_{3it}, a qual representa a variabilidade adicional presente no EBLUP temporal devida à estimação das componentes de variância, não pode ser calculada analiticamente de forma exacta, sendo necessário utilizar aproximações. A metodologia aqui proposta é baseada nos trabalhos de Butar e Lahiri (2003) ao nível da utilização de um método *bootstrap* robusto (Wu, 1986) no contexto de estimação em pequenos domínios com populações finitas. Assumindo que a estimação é assistida pelo modelo temporal de nível área estacionário devido a Rao e Yu (1994), que está disponível um conjunto de dados iniciais provenientes de uma amostra aleatória, que as componentes de variância são estimadas pelo método dos momentos apresentado na secção 4.3.4, e que ρ é conhecido, propõe-se o seguinte procedimento *bootstrap*:

1. Calcular as estimativas das componentes de variância pelo método dos momentos, $\hat{\sigma}_v^2$ e $\hat{\sigma}^2$, com base nos dados iniciais, $\mathbf{y}$, e ajustar o modelo (4.3.5) de forma a determinar as estimativas dos efeitos fixos $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}; \hat{\boldsymbol{\psi}})$, com $\hat{\boldsymbol{\psi}} = \left(\hat{\sigma}_v^2, \hat{\sigma}^2\right)'$.

2. Calcular estimativas EBLUP temporais de θ_{it}, $\hat{\theta}_{it} = \hat{\theta}_{it}(\mathbf{y}; \hat{\boldsymbol{\psi}})$, e dos dois primeiros termos do seu EQMP, $g_{1it}(\hat{\boldsymbol{\psi}})$ e $g_{2it}(\hat{\boldsymbol{\psi}})$.

3. Gerar *m* cópias independentes da variável $\mathbf{v}^*$, com $\mathbf{v}^* \sim N\left(\mathbf{0}; \hat{\sigma}_v^2 \mathbf{I}_m\right)$.

4. Gerar *mT* cópias independentes da variável $\boldsymbol{\xi}^*$, com $\boldsymbol{\xi}^* \sim N\left(\mathbf{0}; \hat{\sigma}^2 \mathbf{I}_{mT}\right)$, independentes de $\mathbf{v}^*$. Com base nestes valores gerar o vector aleatório $\mathbf{u}^*$, admitindo ρ conhecido.

5. Gerar *mT* cópias independentes da variável $\boldsymbol{\varepsilon}^*$, com $\boldsymbol{\varepsilon}^* \sim N(\mathbf{0};\mathbf{R})$, independentes de $\mathbf{v}^*$ e $\boldsymbol{\xi}^*$.

6. Construir o conjunto de dados *bootstrap* $\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}^*$, onde $\boldsymbol{\upsilon}^* = \left(\mathbf{v}^{*\prime} \quad \mathbf{u}^{*\prime}\right)'$.

7. Calcular estimativas *bootstrap* das componentes de variância, $\hat{\sigma}_v^{2*}$ e $\hat{\sigma}^{2*}$, com base nos dados *bootstrap,* $\mathbf{y}^*$, e ajustar o modelo (4.3.5) de forma a obter as estimativas *bootstrap* dos efeitos fixos $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}\left(\mathbf{y};\hat{\boldsymbol{\psi}}^*\right)$, com $\hat{\boldsymbol{\psi}}^* = \left(\hat{\sigma}_v^{2*}, \hat{\sigma}^{2*}\right)'$.

8. Calcular estimativas *bootstrap* do EBLUP temporal, bem como das duas primeiras componentes do seu EQMP, utilizando as estimativas *bootstrap* das componentes de variância, $\hat{\boldsymbol{\psi}}^*$:

$$\hat{\theta}_{it}^* = \hat{\theta}_{it}^*(\mathbf{y};\hat{\boldsymbol{\beta}}^*;\hat{\boldsymbol{\psi}}^*) = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^* + (\hat{\sigma}_v^{2*}\mathbf{1}_T + \hat{\sigma}^{2*}\boldsymbol{\gamma}_t)'\left[\hat{\mathbf{V}}_i\left(\hat{\boldsymbol{\psi}}^*\right)\right]^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^*),$$

$$g_{1it}^* = g_{1it}^*\left(\mathbf{y};\hat{\boldsymbol{\beta}}^*;\hat{\boldsymbol{\psi}}^*\right) = \hat{\sigma}_v^{2*} + \frac{\hat{\sigma}^{2*}}{1-\rho^2} - \left(\hat{\sigma}_v^{2*}\mathbf{1}_T + \hat{\sigma}^{2*}\boldsymbol{\gamma}_t\right)'\left(\hat{\mathbf{V}}_i^*\right)^{-1}\left(\hat{\sigma}_v^{2*}\mathbf{1}_T + \hat{\sigma}^{2*}\boldsymbol{\gamma}_t\right),$$

$$g_{2it}^* = g_{2it}^*\left(\mathbf{y};\hat{\boldsymbol{\beta}}^*;\hat{\boldsymbol{\psi}}^*\right) = \left[\mathbf{x}_{it} - \mathbf{X}_i'\left(\hat{\mathbf{V}}_i^*\right)^{-1}\left(\hat{\sigma}_v^{2*}\mathbf{1}_T + \hat{\sigma}^{2*}\boldsymbol{\gamma}_t\right)\right]'\left[\mathbf{X}'\left(\hat{\mathbf{V}}^*\right)^{-1}\mathbf{X}\right]^{-1} \times \left[\mathbf{x}_{it} - \mathbf{X}_i'\left(\hat{\mathbf{V}}_i^*\right)^{-1}\left(\hat{\sigma}_v^{2*}\mathbf{1}_T + \hat{\sigma}^{2*}\boldsymbol{\gamma}_t\right)\right].$$

9. Repetir as etapas 3)-8) *B* vezes. Defina-se $\hat{\sigma}_v^{2*(b)}$ e $\hat{\sigma}^{2*(b)}$ como estimativas *bootstrap* dos parâmetros de variância obtidas na *b*-ésima réplica *bootstrap*, $\hat{\boldsymbol{\psi}}^{*(b)} = \left(\hat{\sigma}_v^{2*(b)}, \hat{\sigma}^{2*(b)}\right)'$; e $\hat{\boldsymbol{\beta}}^{*(b)}$, $\hat{\theta}_{it}^{*(b)}$, $g_{1it}^{*(b)}$ e $g_{2it}^{*(b)}$ como estimativas *bootstrap* de $\boldsymbol{\beta}$, θ_{it}, g_{1it} e g_{2it}, respectivamente, obtidas na *b*-ésima réplica *bootstrap*, *b*=1, …, *B*.

10. Calcular uma estimativa *bootstrap* de g_{3it}, usando a seguinte aproximação de Monte Carlo: $g_{3it}^* = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{it}^{*(b)} - \hat{\theta}_{it}\right)^2$.

Uma vez obtidas as estimativas *bootstrap* g_{3it}^*, e considerando-se que $g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}})$ é um estimador enviesado de $g_{1it}(\boldsymbol{\psi}) + g_{2it}(\boldsymbol{\psi})$ (Prasad e Rao, 1990), então propõe-se o

seguinte estimador *bootstrap* do EQMP do EBLUP temporal com correcção de enviesamento:

$$eqmp^{B}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right]=2\left[g_{1it}(\hat{\boldsymbol{\psi}})+g_{2it}(\hat{\boldsymbol{\psi}})\right]-B^{-1}\sum_{b=1}^{B}\left[g_{1it}^{*(b)}+g_{2it}^{*(b)}\right]+g_{3it}^{*}. \qquad (4.3.18)$$

4.3.7 Aproximação *jackknife* do EQMP do EBLUP

Nesta secção é proposta uma metodologia alternativa que permite obter uma aproximação do estimador do EQMP do EBLUP temporal através de um procedimento *jackknife* ponderado. A metodologia proposta, baseada no trabalho geral de Jiang *et al.* (2002) e nos desenvolvimentos em série de Taylor de Chen e Lahiri (2008), está desenhada para determinar estimativas para a parcela g_{3it} do EQMP, que, como tem vindo a ser referido, não pode ser calculada analiticamente de forma exacta. É então deduzida uma aproximação *jackknife* ponderada para o estimador do EQMP do EBLUP temporal sob o modelo de Rao e Yu (1994).

Antes de se avançar para a apresentação da referida metodologia *jackknife*, é conveniente referir que o procedimento habitual que consiste em retirar repetidamente uma única observação do conjunto de dados, baseada na hipótese de independência entre as observações, é inconsistente no âmbito do modelo de Rao-Yu porque as observações pertencentes a um determinado domínio estão correlacionadas. Desta forma, decidiu retirar-se repetidamente o conjunto de *T* observações pertencente a cada domínio, uma vez que se admite que as observações pertencentes a domínios diferentes são independentes, fazendo com que este procedimento não provoque alterações na estrutura de correlação intra-domínio.

Considerando-se o modelo de Rao-Yu, a estimação das componentes de variância pelo método dos momentos e ρ conhecido, uma aproximação *jackknife* ponderada para o estimador do EQMP do EBLUP temporal é dada por:

$$\begin{aligned} eqmp^{J}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] &= g_{1it}(\hat{\boldsymbol{\psi}})+g_{2it}(\hat{\boldsymbol{\psi}})-\hat{\mathbf{c}}'_{WJ,t}(\hat{\boldsymbol{\psi}})\nabla g_{1it}(\hat{\boldsymbol{\psi}})+tr\left[\mathbf{A}_{it}\hat{\mathbf{v}}_{WJ,t}\right]+ \\ &\quad +tr\left\{\mathbf{L}_{it}(\hat{\boldsymbol{\psi}})\left[\mathbf{y}_{i}-\mathbf{X}_{i}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]\left[\mathbf{y}_{i}-\mathbf{X}_{i}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]'\mathbf{L}'_{it}(\hat{\boldsymbol{\psi}})\hat{\mathbf{v}}_{WJ,t}\right\} \end{aligned}, \qquad (4.3.19)$$

onde $g_{1it}(\boldsymbol{\psi})$ é dado por (4.3.12); $g_{2it}(\boldsymbol{\psi})$ é dado por (4.3.13) e $\mathbf{A}_{it}$ é uma matriz 2×2 simétrica tal como descrito acima. Para além disso, $\nabla g_{1it}(\boldsymbol{\psi}) = \left(\frac{\partial g_{1it}}{\partial \sigma^2}, \frac{\partial g_{1it}}{\partial \sigma_v^2} \right)'$ é o gradiente de $g_{1it}(\boldsymbol{\psi})$ em σ^2 e σ_v^2, o qual é um vector de dimensão 2×1 com elementos:

$$\begin{aligned}
\frac{\partial g_{1it}}{\partial \sigma^2} &= \frac{1}{1-\rho^2} - \boldsymbol{\gamma}_t' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \left(\sigma_v^2 \mathbf{1}_T' + \sigma^2 \boldsymbol{\gamma}_t' \right) \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \sigma^2} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) - \\
&\quad - \left(\sigma_v^2 \mathbf{1}_T' + \sigma^2 \boldsymbol{\gamma}_t' \right) \mathbf{V}_i^{-1} \boldsymbol{\gamma}_t \\
&= \frac{1}{1-\rho^2} - 2\boldsymbol{\gamma}_t' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right)' \mathbf{V}_i^{-1} \boldsymbol{\Gamma} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) \\
&= \frac{1}{1-\rho^2} + \left[\left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right)' \mathbf{V}_i^{-1} \boldsymbol{\Gamma} - 2\boldsymbol{\gamma}_t' \right] \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) \qquad (4.3.20)
\end{aligned}$$

e

$$\begin{aligned}
\frac{\partial g_{1it}}{\partial \sigma_v^2} &= 1 - \mathbf{1}_T' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \left(\sigma_v^2 \mathbf{1}_T' + \sigma^2 \boldsymbol{\gamma}_t' \right) \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \sigma_v^2} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) - \\
&\quad - \left(\sigma_v^2 \mathbf{1}_T' + \sigma^2 \boldsymbol{\gamma}_t' \right) \mathbf{V}_i^{-1} \mathbf{1}_t \\
&= 1 - 2\mathbf{1}_T' \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right)' \mathbf{V}_i^{-1} \mathbf{J}_T \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) \\
&= 1 + \left[\left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right)' \mathbf{V}_i^{-1} \mathbf{J}_T - 2\mathbf{1}_T' \right] \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right). \qquad (4.3.21)
\end{aligned}$$

Tem-se ainda que $\mathbf{L}_{it}(\boldsymbol{\psi}) = \left(\frac{\partial \mathbf{b}_{it}}{\partial \sigma^2}, \frac{\partial \mathbf{b}_{it}}{\partial \sigma_v^2} \right)'$ é uma matriz por blocos de ordem $2T$×1, onde $\mathbf{b}_{it} = \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right)$, e com blocos de ordem T×1 dados por:

$$\begin{aligned}
\frac{\partial \mathbf{b}_{it}}{\partial \sigma^2} &= -\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \sigma^2} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \mathbf{V}_i^{-1} \boldsymbol{\gamma}_t \\
&= -\mathbf{V}_i^{-1} \boldsymbol{\Gamma} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \mathbf{V}_i^{-1} \boldsymbol{\gamma}_t \\
&= \mathbf{V}_i^{-1} \left[\boldsymbol{\gamma}_t - \boldsymbol{\Gamma} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) \right] \qquad (4.3.22)
\end{aligned}$$

e

$$\frac{\partial \mathbf{b}_{it}}{\partial \sigma_v^2} = -\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \sigma_v^2} \mathbf{V}_i^{-1} \left(\sigma_v^2 \mathbf{1}_T + \sigma^2 \boldsymbol{\gamma}_t \right) + \mathbf{V}_i^{-1} \mathbf{1}_T$$

$$= -\mathbf{V}_i^{-1}\mathbf{J}_T\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_t\right) + \mathbf{V}_i^{-1}\mathbf{1}_T$$

$$= \mathbf{V}_i^{-1}\left[\mathbf{1}_T - \mathbf{J}_T\mathbf{V}_i^{-1}\left(\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_t\right)\right] . \qquad (4.3.23)$$

Por último, $\hat{\mathbf{c}}_{WJ,t} = \sum_{e=1}^{m} w_{et}\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)$ é um estimador *jackknife* ponderado do enviesamento de $\hat{\boldsymbol{\psi}}$; e $\hat{\mathbf{v}}_{WJ,t} = \sum_{e=1}^{m} w_{et}\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)'$ é um estimador *jackknife* ponderado da matriz de covariâncias de $\hat{\boldsymbol{\psi}}$, ambos corrigidos até à ordem $o\left(m^{-1}\right)$, e nos quais $\hat{\boldsymbol{\psi}}_{-e}$ é um estimador de $\boldsymbol{\psi}$ depois de eliminar as *T* observações referentes ao *e*-ésimo pequeno domínio e w_{et} são ponderadores a satisfazerem a seguinte condição $w_{et} = 1 + o\left(m^{-1}\right)$.

Como existem diferentes possibilidades de escolha dos ponderadores, decidiu utilizar-se duas possibilidades utilizadas por Chen e Lahiri (2008), no contexto do modelo de Fay-Herriot: $w_{1et} = \dfrac{m-1}{m}$ e $w_{2et} = 1 - \mathbf{x}_{et}'\left(\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}\mathbf{x}_{et}$. Naturalmente que o uso de diferentes ponderadores resulta em diferentes estimadores *jackknife* do EQMP do EBLUP temporal.

4.3.8 Comentários finais

Na situação mais realista em que o parâmetro de correlação, ρ, do processo AR(1) também é desconhecido, Rao e Yu (1994) sugeriram três métodos para a sua estimação: (*i*) um estimador em dois passos baseado em conjecturas *a priori* sobre o valor de ρ; (*ii*) um estimador consistente deduzido pelo método dos momentos que não ignora os erros da sondagem, ε_{dt}, mas que frequentemente toma valores fora do intervalo admissível $[-1;1]$; e (*iii*) um estimador simplista, também deduzido pelo método dos momentos, mas que ignora os erros da sondagem. Este último estimador, proposto por Pantula e Pollock (1985), é dado por:

$$\hat{\rho}_S = \frac{\sum_{i=1}^{m}\sum_{t=1}^{T-2}\hat{\phi}_{it}\left(\hat{\phi}_{i,t+1}-\hat{\phi}_{i,t+2}\right)}{\sum_{i=1}^{m}\sum_{t=1}^{T-2}\hat{\phi}_{it}\left(\hat{\phi}_{it}-\hat{\phi}_{i,t+1}\right)}, \quad T>2, \tag{4.3.24}$$

onde $\hat{\phi}_{it} = y_{it} - \mathbf{x}'_{it}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ é o *it*-ésimo resíduo dos mínimos quadrados ordinários. Segundo Rao e Yu (1994), o estimador $\hat{\rho}_S$ é inconsistente e geralmente subestima ρ. Apesar disso, o EBLUP de θ_{it} resultante, $\hat{\theta}_{it}\left(\hat{\rho}_S\right)$, permanece não enviesado. A substituição de $\hat{\rho}_S$ por ρ na expressão (4.3.17) permite fazer a estimação do EQMP de $\hat{\theta}_{it}\left(\hat{\rho}_S\right)$. Contudo, esse estimador do EQMP não tem os termos corrigidos à ordem $o\left(m^{-1}\right)$. Por esta razão, todo o trabalho de Rao e Yu (1994) foi baseado na hipótese de ρ ser conhecido. Para ser possível comparar directamente o estimador analítico do EQMP do EBLUP proposto por Rao e Yu (1994) com os estimadores baseados em métodos de reamostragem propostos neste trabalho, decidiu também assumir-se que ρ é conhecido.

Em 1989, num trabalho pioneiro de estimação em pequenos domínios com modelos temporais de nível área, Choudhry e Rao (1989) utilizaram um caso especial do modelo de Rao-Yu para produzirem estimativas EBLUP para o desemprego mensal nas *census divisions* canadianas, utilizando dados do *Canadian Labour Force Survey*[45]. Estes autores trataram os erros compósitos, $a_{it} = u_{it} + \varepsilon_{it}$, como um processo AR(1) e assumiram que $y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \upsilon_i$ (i=1, ..., m; t=1,..., T). Neste caso, o modelo combinado foi apresentado como:

$$\begin{aligned} y_{it} &= \mathbf{x}'_{it}\boldsymbol{\beta} + \upsilon_i + a_{it} \\ a_{it} &= \rho a_{i,t-1} + \xi_{it}, \quad |\rho| < 1 \end{aligned} \tag{4.3.25}$$

onde $\upsilon_i \overset{iid}{\sim} N\left(0;\sigma_\upsilon^2\right)$ e $\xi_{it} \overset{iid}{\sim} N\left(0;\sigma^2\right)$. Choudhry e Rao (1989) também utilizaram um estimador do EQMP do EBLUP temporal, embora na sua versão simplista. As

[45] O *Canadian Labour Force Survey* é um inquérito que fornece dados para a produção de estimativas do emprego e do desemprego no Canadá. Este inquérito repetido no tempo cobre a população civil, não institucionalizada, com 15 ou mais anos de idade. Os dados são recolhidos através de entrevistas assistidas por computador, efectuadas por entrevistadores treinados (Rao, 2003).

componentes de variância foram estimadas através de uma extensão do método III de Henderson (1953) e o parâmetro de correlação foi estimado através de um método dos momentos que ignora os erros da sondagem.

É possível encontrar na literatura outros trabalhos que utilizaram o modelo de Rao-Yu modificado, mas que seguiram uma abordagem de estimação pelo método EBP ou pelo método *Hierarchical Bayes Prediction*. Segundo Rao (2003), os estimadores EBP são idênticos aos EBLUP no caso em que se assume a normalidade no modelo linear misto. Datta *et al.* (1997, 2002) e You na sua investigação de doutoramento (Rao, 2003), obtiveram os estimadores EBLUP (estimadores EBP) e os estimadores do EQMP associados, corrigidos até à segunda ordem, para o modelo de Rao-Yu supondo um passeio aleatório sobre os u_{it}. Datta *et al.* (2002) usaram os métodos da MV e da MVR para estimar as componentes de variância, enquanto You utilizou o método dos momentos. Datta *et al.* (2002) utilizaram o EBLUP para estimar o rendimento mediano das famílias americanas com quatro pessoas, nos cinquenta estados americanos e no distrito de Columbia, com dados resultantes do *Current Population Survey*[46].

4.4 MODELOS DO TIPO *STATE SPACE*

4.4.1 Introdução

Pfeffermann e Burck (1990) e Singh *et al.* (1994) também propuseram generalizações do modelo de Fay-Herriot, mas nas quais os efeitos fixos, $\boldsymbol{\beta}$, foram substituídos por efeitos aleatórios, $\boldsymbol{\beta}_{it}$, obedecendo a um processo auto-regressivo. A generalização do modelo de Fay-Herriot proposta por Singh *et al.* (1994) que incorpora informação histórica no processo de estimação, e na qual os parâmetros de regressão, incluindo os

[46] O *Current Population Survey* é um inquérito mensal realizado aos agregados familiares dos Estados Unidos da América. Este inquérito fornece dados sobre a força de trabalho e sobre o emprego e o desemprego e cobre a população civil, não institucionalizada, com 16 ou mais anos de idade. A amostra é composta por 60.000 agregados familiares e os dados são recolhidos através de entrevistas pessoais e telefónicas (Rao, 2003).

efeitos aleatórios de domínio, evoluem de acordo com determinados modelos temporais, enquadra-se no modelo geral proposto por Pfeffermann e Burck (1990).

4.4.2 Modelo de Pfeffermann-Burck

A classe geral de modelos com coeficientes seccionais referenciados aos pequenos domínios e variantes no tempo, proposta por Pfeffermann e Burck (1990), enquadra-se no contexto dos modelos de nível unidade. Contudo, essa classe de modelos pode ser adaptada para o caso de modelos de nível área, tal como é apresentado em Pfeffermann (2002) e Rao (2003). Nesta subsecção apresenta-se essa classe de modelos *state space* de nível área, designada por modelo de Pfeffermann-Burck. Sejam θ_{it} e y_{it}, respectivamente, um parâmetro populacional da variável de interesse e o seu estimador directo, e $\mathbf{x}_{dt} = \left(1, x_{dt1}, ..., x_{dtp}\right)'$ um vector (p+1)-dimensional de variáveis explicativas associadas ao i-ésimo pequeno domínio no período t (i=1,..., m; t=1,..., T). O modelo de Pfeffermann-Burck é especificado da seguinte forma:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_{it} + \varepsilon_{it} \tag{4.4.1}$$

onde os coeficientes, $\boldsymbol{\beta}_{it} = \left(\beta_{it0}, \beta_{it1}, ..., \beta_{itp}\right)'$, podem variar seccionalmente e cronologicamente, os erros da sondagem, ε_{it}, são não correlacionados cronologicamente para cada i, e têm $E\left(\varepsilon_{it}\right) = 0$ e $V\left(\varepsilon_{it}\right) = \sigma_{it}^2$. A variação de β_{it} ao longo do tempo é especificada pela seguinte "equação de transição":

$$\begin{bmatrix} \beta_{itj} \\ \beta_{ij} \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \beta_{i,t-1,j} \\ \beta_{ij} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{itj}, \quad j=0, 1, ..., p. \tag{4.4.2}$$

Os β_{ij} são coeficientes fixos, $\mathbf{T}_j$ define uma matriz 2×2 conhecida com $(0,1)$ na segunda linha, e os erros do modelo, η_{itj}, satisfazem $E\left(\eta_{itj}\right) = 0$ e $E\left(\eta_{ijt}\eta_{ilt}\right) = \sigma_{\eta,jl}$; j, l=0, 1, ..., p. Isto significa que, para o mesmo momento t, os erros de diferentes coeficientes podem estar correlacionados, mas são não correlacionados cronologicamente e seccional-cronologicamente. Pfeffermann e Burck (1990) consideraram ainda a possibilidade de existência de correlação contemporânea de um

parâmetro entre dois domínios, formulada como $E(\eta_{ijt}\eta_{i'jt}) = \sigma_{\eta,jj}\rho_j$ e $E(\eta_{ijt}\eta_{i'lt}) = 0$, $j \neq l$. Contudo, Coelho (2000) considera que a existência de correlação contemporânea de um parâmetro entre dois domínios constante e independente dos domínios se pode apresentar inadequada para um grande número de situações, uma vez que não contempla as possíveis "dissemelhanças" entre pares de domínios. A especificação (4.4.2) da autoria de Pfeffermann e Burck (1990) cobre alguns modelos de evolução dos coeficientes regressão muito utilizados[47], tais como o modelo com coeficientes de regressão aleatórios, $\beta_{itj} = \beta_{ij} + \eta_{itj}$; o modelo de passeio aleatório, $\beta_{itj} = \beta_{i,t-1,j} + \eta_{itj}$; o modelo auto-regressivo de primeira ordem, $(\beta_{itj} - \beta_{ij}) = \rho(\beta_{i,t-1,j} - \beta_{ij}) + \eta_{itj}$; entre outros.

O modelo de Pfeffermann-Burck definido por (4.4.1) e (4.4.2), pode ser apresentado de forma compacta obedecendo à formulação clássica do modelo *state space* linear apresentado no subcapítulo 3.3:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \varepsilon_t \\ \boldsymbol{\alpha}_t &= \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{A}\boldsymbol{\eta}_t \end{aligned} \qquad (4.4.3)$$

onde $\mathbf{y}_t = col_{1\le i\le m}(y_{it})$, $\mathbf{Z}_t = diag_{1\le i\le m}(\mathbf{Z}_{it})$, $\mathbf{Z}_{it} = (1, 0, x_{it1}, 0, .., x_{itp}, 0)$, $\boldsymbol{\alpha}_t = col_{1\le i\le m}(\boldsymbol{\alpha}_{it})$, $\boldsymbol{\alpha}'_{it} = (\beta_{it0}, \beta_{i0}, \beta_{it1}, \beta_{i1}, ..., \beta_{itp}, \beta_{ip})$, $\boldsymbol{\varepsilon}_t = col_{1\le i\le m}(\varepsilon_{it})$, $\boldsymbol{\eta}_t = col_{1\le i\le m}(\boldsymbol{\eta}_{it})$, $\boldsymbol{\eta}_{it} = (\eta_{it0}, \eta_{it1}, ..., \eta_{itp})$, $\mathbf{T} = \mathbf{I}_m \otimes \tilde{\mathbf{T}}$, $\tilde{\mathbf{T}} = diag_{1\le j\le p}(\mathrm{T}_j)$, $\mathbf{A} = \mathbf{I}_m \otimes \tilde{\mathbf{A}}$ e $\tilde{\mathbf{A}} = \mathbf{I}_{p+1} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Assume-se que $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são não correlacionados contemporaneamente, nem cronologicamente, têm média nula e matrizes de covariâncias dadas, respectivamente, por $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t) = \mathbf{R}_t = diag_{1\le i\le m}(\sigma^2_{it})$ e $E(\boldsymbol{\eta}_t\boldsymbol{\eta}'_t) = \mathbf{Q}$, onde $\mathbf{Q} = \{\mathbf{Q}_{ii'}\}$, i,i'=1,..., m, com $\mathbf{Q}_{ii'} = \begin{cases} E(\boldsymbol{\eta}_{it}\boldsymbol{\eta}'_{it}) & , i = i' \\ diag_{1\le j\le p}(\sigma_{\eta,jj}\rho_j) & , i \neq i' \end{cases}$.

Pfeffermann e Burck (1990) propuseram estimar os coeficientes de regressão do modelo *state space* (4.4.3) através de um filtro de Kalman. Na estimação dos coeficientes de regressão, estes autores assumiram que as matrizes de covariâncias $\mathbf{R}_t$ e $\mathbf{Q}$ são

[47] Consoante os elementos das matrizes $\mathbf{T}_j$, j=0, 1, ..., p.

conhecidas. No caso particular do modelo de Pfeffermann-Burck com $\mathbf{T}=\mathbf{I}$ e $\mathbf{A}=\mathbf{I}$, ou seja, no caso em que os coeficientes de regressão seguem um passeio aleatório, aqueles autores deduziram que o BLUP de θ_{it}, $\tilde{\theta}_{it}$, pode ser escrito como um estimador combinado, sendo uma soma ponderada do estimador directo, y_{it}, do estimador sintético, $\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_{it|t-1}$, e dos "factores de ajustamento" $\left(y_{i't}-\mathbf{x}'_{i't}\tilde{\boldsymbol{\beta}}_{i't|t-1}\right)$, $i{\neq}i'$ (Pfeffermann e Burk, 1990):

$$\tilde{\theta}_{it}=\tilde{\theta}_{it}^{H}(\boldsymbol{\psi})=\left(1-\frac{\sigma_{it}^2}{\iota_i^2}\right)y_{it}+\frac{\sigma_{it}^2}{\iota_t^2}\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_{it|t-1}+\frac{\sigma_{it}^2}{\iota_t^2}\sum_{\substack{i'=1\\ i'\neq i}}^{m}\gamma_{ii'}\left(y_{i't}-\mathbf{x}'_{i't}\tilde{\boldsymbol{\beta}}_{i't|t-1}\right), \qquad (4.4.4)$$

onde $\tilde{\boldsymbol{\beta}}_{it|t-1}$ é o estimador de $\boldsymbol{\beta}_{it}$ com base em toda a informação disponível até ao período $t-1$, $\gamma_{ii'}$ são os coeficientes de regressão parciais obtidos na regressão de $e_{it}=\left(y_{it}-\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_{it|t-1}\right)$ sobre os erros de predição dos outros pequenos domínios $e_{i't}=\left(y_{i't}-\mathbf{x}'_{i't}\tilde{\boldsymbol{\beta}}_{i't|t-1}\right)$, e ι_i^2 representa a variância dos resíduos desta regressão. Quando os parâmetros desconhecidos das matrizes de covariâncias, $\boldsymbol{\psi}$, são substituídos pelos seus estimadores, $\hat{\boldsymbol{\psi}}$, obtém-se o EBLUP $\hat{\theta}_{it}$.

De acordo com o trabalho desenvolvido por Kackar e Harville (1984), e supondo a normalidade dos erros $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$, o EQMP do EBLUP é dado por:

$$EQMP\left(\hat{\theta}_{it}\right)=EQMP\left(\tilde{\theta}_{it}\right)+E\left(\hat{\theta}_{it}-\tilde{\theta}_{it}\right)^2, \qquad (4.4.5)$$

onde $EQMP\left(\tilde{\theta}_{it}\right)=\mathbf{x}'_{it}\boldsymbol{\Sigma}_t\mathbf{x}_{it}$. A partir do trabalho de Ansley e Kohn (1986), sobre o desenvolvimento em série de Taylor do EBLUP em torno de $\boldsymbol{\psi}$ sob modelos *state space*, Rao (2003) sugere o seguinte estimador do EQMP do EBLUP[48]:

$$eqmp\left(\hat{\theta}_{it}\right)\approx\mathbf{x}'_{it}\boldsymbol{\Sigma}_t(\hat{\boldsymbol{\psi}})\mathbf{x}_{it}+\left[\frac{\partial\tilde{\theta}_{it}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right]'_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\mathbf{I}^{-1}_{\hat{\boldsymbol{\psi}}}(\hat{\boldsymbol{\psi}})\left[\frac{\partial\tilde{\theta}_{it}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right]_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}, \qquad (4.4.6)$$

[48] O estimador do EQMP (4.4.6) não é corrigido até à segunda ordem.

onde $\hat{\boldsymbol{\psi}}$ é o estimador da MV de $\boldsymbol{\psi}$ e $\mathbf{I}(\hat{\boldsymbol{\psi}})$ é a correspondente matriz de informação avaliada no ponto $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}$. Posteriormente, Pfeffermann e Tiller (2005) propuseram um estimador do EQMP do EBLUP baseado num método *bootstrap* paramétrico.

Pfeffermann e Burck (1990) usaram este modelo do tipo *state space* no contexto de estimação dos índices de preços de habitação. É, ainda, de salientar que, segundo Rao (2003), apesar do modelo *state space* (4.4.3) ser bastante geral, a hipótese dos erros da sondagem, ε_{it}, serem não correlacionados cronologicamente é restritiva no contexto de sondagens repetidas no tempo com sobreposição.

A generalização do modelo de Fay-Herriot proposta por Singh *et al.* (1994), que incorpora informação temporal no processo de estimação, e na qual os parâmetros de regressão, incluindo os efeitos aleatórios de domínio, evoluem de acordo com determinados modelos cronológicos, enquadra-se na classe geral de modelos definidos por Pfeffermann e Burck (1990). Estes autores propuseram estimadores EBLUP que são casos particulares de um modelo estrutural para as estimativas directas baseado em séries temporais, especificado pelo seguinte modelo *state space*, e portanto passível de ser estimado através de um filtro de Kalman:

$$\mathbf{y}_t = \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t \tag{4.4.7}$$

$$\boldsymbol{\theta}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \mathbf{Z}_t\mathbf{u}_t \equiv \mathbf{U}_t\boldsymbol{\alpha}_t \tag{4.4.8}$$

onde $\boldsymbol{\alpha}_t = (\boldsymbol{\beta}_t', \mathbf{u}_t')'$ e $\mathbf{U}_t = (\mathbf{X}_t, \mathbf{Z}_t)$, t=1,..., T. Combinando as equações (4.4.7) e (4.4.8), tem-se a seguinte "equação de medição":

$$\mathbf{y}_t = \mathbf{U}_t\boldsymbol{\alpha}_t + \varepsilon_t . \tag{4.4.9}$$

Permite-se então que os parâmetros da regressão, bem como os efeitos aleatórios dos domínios sintetizados em $\boldsymbol{\alpha}_t$ (vector de estado), evoluem no tempo de acordo com a "equação de transição" $\boldsymbol{\alpha}_t = \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\varsigma}_t$, onde $\mathbf{T}_t = diag_{1\le k\le 2}\left(\mathbf{T}_t^{(k)}\right)$ e $\boldsymbol{\varsigma}_t = (\boldsymbol{\xi}_t', \boldsymbol{\eta}_t')'$. Assume-se ainda que os $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\varsigma}_t$ são não correlacionados, $\boldsymbol{\varsigma}_t$ é não correlacionado com os α_s para $s < t$, e que $\boldsymbol{\varepsilon}_t \sim (\mathbf{0}; \mathbf{R}_t)$, $\boldsymbol{\varsigma}_t \sim (\mathbf{0}; \mathbf{Q}_t)$, onde $\mathbf{Q}_t = diag\{\mathbf{Q}_{1t}, \mathbf{Q}_{2t}\}$, tal que $\boldsymbol{\xi}_t \sim (\mathbf{0}; \mathbf{Q}_{1t})$ e $\boldsymbol{\eta}_t \sim (\mathbf{0}; \mathbf{Q}_{2t})$. Se $\mathbf{T}_t^{(1)} = \mathbf{I}$ e $\mathbf{T}_t^{(2)} = \mathbf{I}$, então $\boldsymbol{\beta}_t$ e $\mathbf{u}_t$ evoluem de acordo

com um processo do tipo passeio aleatório. Assume-se, também, que as matrizes $\mathbf{R}_t$, $\mathbf{Q}_{1t}$ e $\mathbf{Q}_{2t}$ são diagonais. Segundo Singh *et al.* (1994), a evidência empírica parece sugerir que os ganhos de eficiência desta última generalização (de Pfeffermann-Burk), relativamente ao modelo de Rao-Yu, são provavelmente pequenos.

Na literatura sobre esta matéria, encontram-se outros trabalhos onde foram utilizados modelos do tipo *state space*. Ghosh e Nangia (1993) e Ghosh *et al.* (1996) também propuseram modelos do tipo *state space* para estimar, segundo uma abordagem *bayesiana*, o rendimento mediano das famílias americanas com quatro pessoas, nos cinquenta estados americanos e no distrito de Columbia. Posteriormente, Pfeffermann *et al*. (1998) aplicaram um modelo deste tipo aos dados da força de trabalho na Austrália.

4.5 MODELOS DE NÍVEL ÁREA COM DADOS ESPACIAIS

4.5.1 Introdução

Recorde-se que os modelos de estimação em pequenos domínios são normalmente baseados em modelos de ligação explícita, os quais envolvem efeitos aleatórios específicos de domínio de forma a acomodar a variabilidade existente entre esses domínios, para além do poder explicativo das variáveis auxiliares incluídas na parte fixa do modelo. Apesar de normalmente se assumir que os efeitos aleatórios específicos de domínio são independentes, o que se verifica nas aplicações práticas é que as fronteiras dos pequenos domínios são arbitrárias. Por este motivo, não parece plausível que as unidades populacionais de um lado da fronteira não estejam ligadas, de alguma forma, às unidades populacionais do outro lado da fronteira. Desta forma, é razoável assumir que os efeitos aleatórios de domínios vizinhos (uma vizinhança pode ser definida, por exemplo, através de um critério de contiguidade ou de distância) estejam associados e, por vezes, apresentem uma associação decrescente para zero com o aumento da distância.

Os estimadores para pequenos domínios apresentados nas secções anteriores, não exploram a eventual associação espacial existente entre domínios, em termos de

semelhanças devidas à proximidade geográfica. Com o objectivo de se ter em consideração, na estimação, a associação existente entre efeitos aleatórios de domínios vizinhos, podem ser utilizados modelos espaciais no âmbito da estimação em pequenos domínios. A primeira generalização do modelo de Fay-Herriot de forma a contemplar essa associação espacial foi efectuada por Cressie (1991).

Tal como foi referido no subcapítulo 3.4, no âmbito dos modelos de regressão linear, a associação espacial pode ser incorporada nos modelos de duas formas alternativas: pode ser especificado um modelo de regressão com termos auto-regressivos ou com erros autocorrelacionados espacialmente (Anselin, 1992). De acordo com a motivação exposta para utilização de modelos espaciais no contexto da estimação em pequenos domínios, naturalmente que os modelos de regressão linear com dependência espacial na estrutura de erro, ou seja, nos efeitos aleatórios sejam os preferidos.

Na estimação em pequenos domínios, os dados relativos a um dado domínio representam a informação resumida relativa a essa subregião, pelo que a informação disponível pode classificar-se como dados referentes a áreas irregulares. Neste contexto particular, assume-se frequentemente processos SAR ou CAR para modelar a estrutura de erro (efeitos aleatórios específicos de domínio). A escolha de modelos SAR ou CAR deve-se ao facto destes modelos constituírem, segundo Wall (2004), a melhor escolha para a modelação quando o objectivo principal consiste em obter bons preditores dos parâmetros de interesse a partir da regressão, em detrimento da compreensão da estrutura espacial subjacente. Note-se que, no âmbito da estimação em pequenos domínios, o objectivo principal consiste em obter estimativas dos parâmetros de interesse com a melhor precisão possível, menosprezando-se a compreensão da estrutura espacial subjacente. A generalização do modelo de Fay-Herriot contemplando efeitos aleatórios de domínio modelados através de um processo SAR tem sido vastamente utilizada no último lustro, designadamente por Salvati (2004), Pratesi e Salvati (2004, 2005, 2008), Petrucci e Salvati (2004*b*, 2006), Petrucci *et al*. (2005), Singh *et al*. (2005) e Chandra *et al*. (2007*a*, 2007*b*). Por sua vez, a generalização do modelo de Fay-Herriot que contempla associação espacial modelada através de um processo CAR foi utilizada inicialmente por Cressie (1991) e posteriormente desenvolvida por Salvati (2004, 2005).

Nas secções seguintes apresenta-se uma breve revisão da literatura sobre modelos de estimação em pequenos domínios que utilizam dados espaciais, os quais se resumem a casos particulares do modelo linear misto (3.2.1).

4.5.2 Modelo espacial do tipo SAR

Nesta secção é apresentada a extensão do modelo de Fay-Herriot contemplando efeitos aleatórios de domínio modelados através de um processo SAR, proposta por Salvati (2004) e utilizada por Singh *et al.* (2005). Sejam $\boldsymbol{\theta} = col_{1\le i\le m}(\theta_i)$ e $\mathbf{y} = col_{1\le i\le m}(y_i)$ vectores $m\times 1$, respectivamente, dos parâmetros populacionais da variável de interesse e dos seus estimadores directos não enviesados no desenho (i=1, ..., m). O modelo de erro da sondagem é dado por:

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{4.5.1}$$

onde $\boldsymbol{\varepsilon} = col_{1\le i\le m}(\varepsilon_i)$ é um vector $m\times 1$ dos erros da sondagem com $\boldsymbol{\varepsilon} \overset{ind}{\sim} N(\mathbf{0};\mathbf{R})$, sendo $\mathbf{R} = diag_{1\le i\le m}(\sigma_i^2)$ uma matriz $m\times m$ com variâncias amostrais conhecidas. A introdução da dependência espacial entre pequenos domínios é efectuada através da especificação do seguinte modelo de ligação com efeitos aleatórios correlacionados espacialmente:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \tag{4.5.2}$$

onde $\mathbf{X} = col_{1\le i\le m}(\mathbf{x}_i')$ é uma matriz $m\times p$ conhecida onde $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ é um vector p-dimensional de variáveis explicativas associadas ao i-ésimo pequeno domínio (i=1, ..., m), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ é um vector $p\times 1$ de coeficientes de regressão, $\mathbf{Z}$ é uma matriz $m\times m$ de constantes positivas conhecidas e $\mathbf{v} = col_{1\le i\le m}(v_i)$ é um vector $m\times 1$ que representa a variação espacial de segunda ordem, a qual reflecte a estrutura de dependência espacial do processo. Neste caso, a inclusão de correlações espaciais entre os efeitos aleatórios específicos de domínio é efectuada através do seguinte processo auto-regressivo simultâneo (Whittle, 1954):

$$\mathbf{v} = \phi\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{u}, \tag{4.5.3}$$

onde ϕ é o coeficiente de associação espacial, $\mathbf{W}$ é uma matriz quadrada de ordem *m* de pesos espaciais conhecidos com elementos principais nulos e $\mathbf{u} = col_{1\le i\le m}(u_i)$ é um vector $m\times 1$ de erros com $\mathbf{u} \overset{iid}{\sim} N(\mathbf{0}; \sigma_u^2 \mathbf{I})$. O modelo combinado com erros correlacionados espacialmente, baseado em (4.5.1) - (4.5.3), é dado por[49]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4.5.4}$$

no qual se assume que os termos de erro $\mathbf{v}$ e $\boldsymbol{\varepsilon}$ são mutuamente independentes, com $\mathbf{G} = \sigma_u^2\left[(\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})\right]^{-1}$. O modelo de Fay e Herriot (1979) pode ser obtido a partir do modelo (4.5.4), bastando para tal fazer $\phi = 0$. O modelo (4.5.4) é um caso particular do modelo linear misto (3.2.1) com a seguinte matriz de covariâncias de $\mathbf{y}$:

$$\mathbf{V} = \mathbf{R} + \sigma_u^2 \mathbf{Z}\left[(\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})\right]^{-1}\mathbf{Z}'. \tag{4.5.5}$$

Sob o modelo apresentado, Salvati (2004) deduziu o BLUP de $\boldsymbol{\theta}$ e o respectivo EQMP a partir dos resultados gerais obtidos por Henderson (1975). Assumindo que o vector de parâmetros $\boldsymbol{\psi} = (\phi, \sigma_u^2)'$ é conhecido, então o BLUP espacial de $\boldsymbol{\theta}$ é dado por:

$$\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}^H(\boldsymbol{\psi}) = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\Lambda}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{4.5.6}$$

onde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ é o estimador dos mínimos quadrados generalizados de $\boldsymbol{\beta}$ e $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\psi}) = \sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}$ com $\mathbf{B} = (\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})$. De acordo com Salvati (2004), o BLUP espacial do parâmetro de interesse relativo ao *i*-ésimo domínio, θ_i, pode ser escrito da seguinte forma:

$$\tilde{\theta}_i = \tilde{\theta}_i^H(\boldsymbol{\psi}) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \mathbf{m}_i'\boldsymbol{\Lambda}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{4.5.7}$$

onde $\mathbf{m}_i' = (0, 0, ..., 0, 1, 0, ..., 0, 0)$ é um vector $1\times m$ com um na *i*-ésima posição e zeros nas restantes. O EMQP do BLUP (4.5.7) é dado por (Salvati, 2004):

[49] Singh *et al.* (2005) apresentaram o modelo combinado da seguinte forma: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$, onde $\mathbf{Z} = (\mathbf{I}_m - \phi\mathbf{W})^{-1}$. Este modelo é equivalente ao modelo (4.5.4) quando se considera $\mathbf{Z} = \mathbf{I}_m$.

$$EQMP\left(\tilde{\theta}_i\right) = g_{1i}(\boldsymbol{\psi}) + g_{2i}(\boldsymbol{\psi}), \tag{4.5.8}$$

onde

$$g_{1i}(\boldsymbol{\psi}) = \mathbf{m}_i'\left\{\sigma_u^2\mathbf{B}^{-1} - \sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{B}^{-1}\right\}\mathbf{m}_i, \tag{4.5.9}$$

$$g_{2i}(\boldsymbol{\psi}) = \left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{m}_i\right]'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{m}_i\right]. \tag{4.5.10}$$

O BLUP espacial de $\boldsymbol{\theta}$ depende de parâmetros, $\boldsymbol{\psi}$, que são geralmente desconhecidos na prática. Quando se substitui essas componentes de variância pelos seus estimadores assimptoticamente consistentes, $\hat{\boldsymbol{\psi}}$, então obtém-se o seguinte EBLUP espacial:

$$\hat{\theta}_i = \hat{\theta}_i^H(\hat{\boldsymbol{\psi}}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{m}_i'\hat{\boldsymbol{\Lambda}}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{4.5.11}$$

onde $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\psi}})$ e $\hat{\boldsymbol{\Lambda}} = \hat{\sigma}_u^2\hat{\mathbf{B}}^{-1}(\hat{\boldsymbol{\psi}})\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\psi}})$. Os algoritmos de estimação de ϕ e de σ_u^2 pelos métodos da MV e da MVR podem ser encontrados em Salvati (2004). Recorde-se que sob condições de normalidade de $\mathbf{v}$ e $\boldsymbol{\varepsilon}$, o EQMP do EBLUP é dado por (Kackar e Harville, 1984):

$$EQMP\left(\hat{\theta}_i\right) = g_{1i}(\boldsymbol{\psi}) + g_{2i}(\boldsymbol{\psi}) + E\left[\hat{\theta}_i(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_i(\boldsymbol{\psi})\right]^2. \tag{4.5.12}$$

As parcelas $g_{1i}(\boldsymbol{\psi})$ e $g_{2i}(\boldsymbol{\psi})$ são dadas por (4.5.9) e (4.5.10), respectivamente. Pelo facto do último termo da expressão (4.5.12) ser intratável, Salvati (2004) propôs uma aproximação em séries de Taylor de ordem $o\left(m^{-1}\right)$ para esse termo baseada nos resultados de Prasad e Rao (1990), e introduziu também estimadores do EQMP do EBLUP espacial envolvendo aquela aproximação. Contudo, Salvati (2004) não estudou se os estimadores do EQMP do EBLUP espacial por si propostos são centrados até à ordem $o\left(m^{-1}\right)$, tal como se verifica para a esmagadora maioria dos estimadores do EQMP dos EBLUP apresentados ao longo deste texto. Porém, Singh *et al.* (2005) fizeram essa investigação, a qual se apresenta de seguida. Sob o modelo (4.5.4), uma aproximação para o terceiro termo da expressão (4.5.12) é dada por (Salvati, 2004; Singh *et al.*, 2005):

$$E\left[\hat{\theta}_i(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_i(\boldsymbol{\psi})\right]^2 \approx tr\left[\mathbf{L}_i(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}_i'(\boldsymbol{\psi})\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})\right] = g_{3i}(\boldsymbol{\psi}), \tag{4.5.13}$$

onde $\mathbf{L}_i(\boldsymbol{\psi})=\begin{bmatrix}\mathbf{m}_i'\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}+\sigma_u^2\mathbf{m}_i'\mathbf{B}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\\ \mathbf{m}_i'\mathbf{E}^{-1}\mathbf{Z}'\mathbf{V}^{-1}+\sigma_u^2\mathbf{m}_i'\mathbf{B}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{E}\mathbf{Z}'\mathbf{V}^{-1}\right)\end{bmatrix}$ é uma matriz por blocos de dimensão $2\times m$, $\mathbf{E}=\sigma_u^2\left[-\mathbf{B}^{-1}\left(2\phi\mathbf{W}\mathbf{W}'-2\mathbf{W}\right)\mathbf{B}^{-1}\right]$ e $\overline{\mathbf{V}}\left(\hat{\boldsymbol{\psi}}_{MVR}\right)\approx\overline{\mathbf{V}}\left(\hat{\boldsymbol{\psi}}_{MV}\right)\approx m^{-1}\mathbf{I}_{\boldsymbol{\psi}}^{-1}\left(\boldsymbol{\psi}\right)$. A matriz de informação, de ordem 2×2, é dada por $\mathbf{I}_{\boldsymbol{\psi}}\left(\boldsymbol{\psi}\right)=\left\{I_{ef}\left(\boldsymbol{\psi}\right)\right\}$, cujos elementos são $I_{ef}\left(\boldsymbol{\psi}\right)=\frac{1}{2}tr\left[\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\psi_e}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\psi_f}\right]$, e,f=1, 2, com $\frac{\partial\mathbf{V}}{\partial\psi_e}=\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}'$ e $\frac{\partial\mathbf{V}}{\partial\psi_f}=\mathbf{Z}\mathbf{E}\mathbf{Z}'$.

Singh *et al.* (2005) propuseram um estimador analítico aproximadamente não enviesado[50] do EQMP do EBLUP espacial, para o caso em que as componentes de variância são estimadas pelo método da MV ou pelo método da MVR. Esse estimador é dado por:

$$eqmp\left[\hat{\theta}_i\left(\hat{\boldsymbol{\psi}}\right)\right]=g_{1i}\left(\hat{\boldsymbol{\psi}}\right)+g_{2i}\left(\hat{\boldsymbol{\psi}}\right)+2g_{3i}\left(\hat{\boldsymbol{\psi}}\right)-g_{4i}\left(\hat{\boldsymbol{\psi}}\right)-g_{5i}\left(\hat{\boldsymbol{\psi}}\right),\qquad(4.5.14)$$

onde $g_{3i}\left(\hat{\boldsymbol{\psi}}\right)-g_{4i}\left(\hat{\boldsymbol{\psi}}\right)-g_{5i}\left(\hat{\boldsymbol{\psi}}\right)$ é o contributo para a variabilidade devido à estimação do vector dos parâmetros de variância desconhecidos, $\boldsymbol{\psi}$. As componentes $g_{4i}\left(\boldsymbol{\psi}\right)$ e $g_{5i}\left(\boldsymbol{\psi}\right)$, também de ordem inferior a $o\left(m^{-1}\right)$, são dadas por:

$$g_{4i}\left(\boldsymbol{\psi}\right)=\mathbf{c}_{\hat{\boldsymbol{\psi}}}'\left(\boldsymbol{\psi}\right)\nabla g_{1i}\left(\hat{\boldsymbol{\psi}}\right),\qquad(4.5.15)$$

$$g_{5i}\left(\boldsymbol{\psi}\right)=\mathbf{m}_i'\frac{1}{2}\underset{m}{tr}\left\{\left[\mathbf{I}_2\otimes\left(\mathbf{R}\mathbf{V}^{-1}\right)\right]\frac{\partial^2\mathbf{V}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\left[\mathbf{I}_{\boldsymbol{\psi}}^{-1}\left(\boldsymbol{\psi}\right)\otimes\left(\mathbf{V}^{-1}\mathbf{R}\right)\right]\right\}\mathbf{m}_i,\qquad(4.5.16)$$

onde $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MV}}\left(\boldsymbol{\psi}\right)\approx\frac{1}{2m}\left\{\mathbf{I}_{\boldsymbol{\psi}}^{-1}\left(\boldsymbol{\psi}\right)\begin{bmatrix}tr\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{X}\right]\\ tr\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{E}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{X}\right]\end{bmatrix}\right\}$ e $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MVR}}\left(\boldsymbol{\psi}\right)\approx\mathbf{0}$ são vectores de dimensão 2×1, e $\nabla g_{1i}\left(\boldsymbol{\psi}\right)=\left(\frac{\partial g_{1i}}{\partial\sigma_u^2},\frac{\partial g_{1i}}{\partial\phi}\right)'$ é o gradiente de $g_{1i}\left(\boldsymbol{\psi}\right)$ em σ_u^2 e ϕ, o qual é um vector de dimensão 2×1 com elementos:

$$\frac{\partial g_{1i}}{\partial\sigma_u^2}=\mathbf{m}_i'\left\{\mathbf{B}^{-1}-\left[\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{B}^{-1}+\sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{Z}\sigma_u^2\mathbf{B}^{-1}+\sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right]\right\}\mathbf{m}_i\qquad(4.5.17)$$

[50] O enviesamento é de ordem inferior ou igual a $o\left(m^{-1}\right)$.

e

$$\frac{\partial g_{1i}}{\partial \phi} = \mathbf{m}_i' \left\{ \mathbf{E} - \left[\mathbf{E}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_u^2\mathbf{B}^{-1} + \sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{E}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{Z}\sigma_u^2\mathbf{B}^{-1} + \sigma_u^2\mathbf{B}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{E} \right] \right\} \mathbf{m}_i \,. \tag{4.5.18}$$

Ademais, $\frac{\partial^2 \mathbf{V}}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}$ é uma matriz por blocos de dimensão $2m \times 2m$ e $\underset{m}{tr}(\boldsymbol{\Omega}) = \sum_{e=1}^{E} \boldsymbol{\Omega}_{ee}$, onde $\boldsymbol{\Omega}$ é uma qualquer matriz por blocos quadrada, sendo todos os blocos também matrizes quadradas da mesma dimensão.

4.5.3 Modelo espacial do tipo CAR

Nesta secção é apresentada uma extensão do modelo de Fay-Herriot contemplando efeitos aleatórios de domínio modelados através de um processo CAR, proposta por Salvati (2004). Relembre-se que, ao contrário do modelo SAR, no modelo CAR se assume que a probabilidade de se observar um valor particular da variável de interesse num determinado domínio, é condicionada pelos valores da variável de interesse nos domínios vizinhos. Tal como anteriormente, sejam $\boldsymbol{\theta} = col_{1 \le i \le m}(\theta_i)$ e $\mathbf{y} = col_{1 \le i \le m}(y_i)$ vectores $m \times 1$, respectivamente, dos parâmetros populacionais da variável de interesse e dos seus estimadores directos não enviesados no desenho (i=1, ..., m). O modelo de erro da sondagem é dado por:

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{4.5.19}$$

onde $\boldsymbol{\varepsilon} = col_{1 \le i \le m}(\varepsilon_i)$ é um vector $m \times 1$ dos erros da sondagem com $\boldsymbol{\varepsilon} \overset{ind}{\sim} N(\mathbf{0}; \mathbf{R})$, sendo $\mathbf{R} = diag_{1 \le i \le m}(\sigma_i^2)$ uma matriz $m \times m$ com variâncias amostrais conhecidas. A introdução da dependência espacial entre pequenos domínios é efectuada através da especificação do seguinte modelo de ligação com efeitos aleatórios correlacionados espacialmente:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \tag{4.5.20}$$

onde $\mathbf{X} = col_{1 \le i \le m}(\mathbf{x}_i')$ é uma matriz $m \times p$ conhecida onde $\mathbf{x}_i = \left(x_{i1}, ..., x_{ip}\right)'$ é um vector p-dimensional de variáveis explicativas associadas ao i-ésimo pequeno domínio (i=1, ...,

m), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ é um vector $p\times1$ de coeficientes de regressão, $\mathbf{Z}$ é uma matriz $m\times m$ de constantes positivas conhecidas e $\mathbf{v} = col_{1\le i\le m}(v_i)$ é um vector $m\times1$ que representa a variação espacial de segunda ordem, a qual reflecte a estrutura de dependência espacial do processo. Neste caso, a inclusão de correlações espaciais entre os efeitos aleatórios específicos de domínio é efectuada através do seguinte processo auto-regressivo condicional (Besag, 1974):

$$v_i \mid \{v_{i'} \in A_i\} \sim \left(\phi \sum_{i'\in A_d} w_{ii'} v_{i'}, \sigma_v^2\right), \tag{4.5.21}$$

onde A_i representa o conjunto dos domínios vizinhos do i-ésimo pequeno domínio, ϕ é o coeficiente de associação espacial e $w_{ii'}$ é um elemento da matriz $\mathbf{W}$, a qual é uma matriz $m\times m$ de pesos espaciais conhecidos. O modelo combinado com erros correlacionados espacialmente, baseado em (4.5.19)-(4.5.21), é dado por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \boldsymbol{\varepsilon}, \tag{4.5.22}$$

no qual se assume que os termos de erro $\mathbf{v}$ e $\boldsymbol{\varepsilon}$ são mutuamente independentes, com $\mathbf{v} \sim \left(\mathbf{0}; \sigma_v^2(\mathbf{I}_m - \phi\mathbf{W})^{-1}\right)$. O modelo de Fay e Herriot (1979) também pode ser obtido a partir do modelo (4.5.22) quando se admite $\phi = 0$. O modelo (4.5.22) é um caso particular do modelo linear misto com a seguinte matriz de covariâncias de $\mathbf{y}$:

$$\mathbf{V} = \mathbf{R} + \sigma_v^2 \mathbf{Z}(\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{Z}'. \tag{4.5.23}$$

No caso do modelo CAR, a matriz $\mathbf{W}$ tem que ser simétrica e $(\mathbf{I}_m - \phi\mathbf{W})$ tem que ser estritamente definida positiva, de forma a assegurar a existência e a simetria de $\sigma_v^2(\mathbf{I}_m - \phi\mathbf{W})^{-1}$ no processo condicional (Upton e Fingleton, 1985). Isto é garantido se $\phi \in \left]1/\min(\lambda_i), 1/\max(\lambda_i)\right[$, onde os λ_i's são os valores próprios da matriz $\mathbf{W}$.

Sob o modelo apresentado, Salvati (2004) deduziu o BLUP de $\boldsymbol{\theta}$ e o respectivo EQMP a partir dos resultados gerais obtidos por Henderson (1975). Assumindo que o vector de parâmetros $\boldsymbol{\psi} = (\phi, \sigma_v^2)'$ é conhecido, então o BLUP espacial de $\boldsymbol{\theta}$ é dado por:

$$\tilde{\theta}_i = \tilde{\theta}_i^H(\boldsymbol{\psi}) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \mathbf{m}_i'\boldsymbol{\Lambda}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right), \tag{4.5.24}$$

onde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ é o estimador dos mínimos quadrados generalizados de $\boldsymbol{\beta}$, $\mathbf{m}_i' = (0,...,0,1,0,...,0)$ é um vector $1\times m$ com um na i-ésima posição e zeros nas restantes e $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\psi}) = \sigma_v^2\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}$ com $\mathbf{D} = \mathbf{I}_m - \phi\mathbf{W}$. O EMQP do BLUP (4.5.24) é dado por:

$$EQMP\left(\tilde{\theta}_i\right) = g_{1i}(\boldsymbol{\psi}) + g_{2i}(\boldsymbol{\psi}), \tag{4.5.25}$$

onde

$$g_{1i}(\boldsymbol{\psi}) = \mathbf{m}_i'\left\{\sigma_v^2(\mathbf{I}_m - \phi\mathbf{W})^{-1} - \sigma_v^2(\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2(\mathbf{I}_m - \phi\mathbf{W})^{-1}\right\}\mathbf{m}_i, \tag{4.5.26}$$

$$g_{2i}(\boldsymbol{\psi}) = \left[\mathbf{x}_i - \sigma_v^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}(\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{m}_i\right]'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{x}_i - \sigma_v^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}(\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{m}_i\right]. \tag{4.5.27}$$

Mais uma vez, o BLUP espacial de $\boldsymbol{\theta}$ depende de parâmetros, $\boldsymbol{\psi}$, que são geralmente desconhecidos na prática. Quando se substitui essas componentes de variância pelos seus estimadores assimptoticamente consistentes, $\hat{\boldsymbol{\psi}}$, então obtém-se o seguinte EBLUP espacial:

$$\hat{\theta}_i = \hat{\theta}_i^H(\hat{\boldsymbol{\psi}}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \mathbf{m}_i'\hat{\boldsymbol{\Lambda}}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{4.5.28}$$

onde $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\psi}})$ e $\hat{\boldsymbol{\Lambda}} = \hat{\sigma}_v^2\hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\psi}})$. Os algoritmos de estimação de ϕ e de σ_v^2 pelos métodos da MV e da MVR podem ser encontrados em Salvati (2004, 2005). Sob condições de normalidade de $\mathbf{v}$ e $\boldsymbol{\varepsilon}$, recorde-se que o EQMP do EBLUP é dada por (Kackar e Harville, 1984):

$$EQMP\left(\hat{\theta}_i\right) = g_{1i}(\boldsymbol{\psi}) + g_{2i}(\boldsymbol{\psi}) + E\left[\hat{\theta}_i(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_i(\boldsymbol{\psi})\right]^2. \tag{4.5.29}$$

As parcelas $g_{1i}(\boldsymbol{\psi})$ e $g_{2i}(\boldsymbol{\psi})$ são dadas por (4.5.26) e (4.5.27), respectivamente. Pelo facto do último termo da expressão (4.5.29) ser intratável, Salvati (2004) propôs uma aproximação em séries de Taylor de ordem $o\left(m^{-1}\right)$ baseada nos resultados de Prasad e Rao (1990). Uma aproximação para esse termo é dada por (Salvati, 2004):

$$E\left[\hat{\theta}_i(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_i(\boldsymbol{\psi})\right]^2 \approx tr\left[\mathbf{L}_i(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}_i'(\boldsymbol{\psi})\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})\right] = g_{3i}(\boldsymbol{\psi}), \tag{4.5.30}$$

onde $\mathbf{L}_i(\boldsymbol{\psi}) = \begin{bmatrix} \mathbf{m}_i'\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1} + \sigma_v^2\mathbf{m}_i'\mathbf{D}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) \\ \sigma_v^2\mathbf{m}_i'\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1} + \sigma_v^2\mathbf{m}_i'\mathbf{D}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right) \end{bmatrix}$ é uma matriz por blocos de dimensão $2\times m$ e $\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}}_{MVR}) \approx \overline{\mathbf{V}}(\hat{\boldsymbol{\psi}}_{MV}) \approx m^{-1}\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})$. A matriz de informação, de ordem 2×2, é dada por $\mathbf{I}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \{I_{ef}(\boldsymbol{\psi})\}$, cujos elementos são $I_{ef}(\boldsymbol{\psi}) = \frac{1}{2}tr\left[\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \psi_e}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \psi_f}\right]$, e,f=1, 2, com $\frac{\partial \mathbf{V}}{\partial \psi_e} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'$ e $\frac{\partial \mathbf{V}}{\partial \psi_f} = \sigma_v^2\mathbf{Z}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'$.

No contexto do modelo (4.5.22), Salvati (2004) propôs um estimador analítico do EQMP do EBLUP espacial, para o caso em que as componentes de variância são estimadas pelos métodos de verosimilhança. Esse estimador é dado por:

$$eqmp\left[\hat{\theta}_i(\hat{\boldsymbol{\psi}})\right] = g_{1i}(\hat{\boldsymbol{\psi}}) + g_{2i}(\hat{\boldsymbol{\psi}}) + 2g_{3i}(\hat{\boldsymbol{\psi}}) - g_{4i}(\hat{\boldsymbol{\psi}}), \tag{4.5.31}$$

onde $g_{3i}(\hat{\boldsymbol{\psi}}) - g_{4i}(\hat{\boldsymbol{\psi}})$ é o contributo para a variabilidade devido à estimação $\boldsymbol{\psi}$. A componente $g_{4i}(\boldsymbol{\psi})$, também de ordem inferior a $o(m^{-1})$, é dada por:

$$g_{4i}(\boldsymbol{\psi}) = \mathbf{c}_{\hat{\boldsymbol{\psi}}}'(\boldsymbol{\psi})\nabla g_{1i}(\hat{\boldsymbol{\psi}}), \tag{4.5.32}$$

onde $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MV}}(\boldsymbol{\psi}) \approx \frac{1}{2m}\left\{\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})\begin{bmatrix} tr\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{X}\right] \\ tr\left[\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}'\left(-\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{X}\right] \end{bmatrix}\right\}$ e $\mathbf{c}_{\hat{\boldsymbol{\psi}}_{MVR}}(\boldsymbol{\psi}) \approx \mathbf{0}$ são vectores de dimensão 2×1, e $\nabla g_{1i}(\boldsymbol{\psi}) = \left(\frac{\partial g_{1i}}{\partial \sigma_v^2}, \frac{\partial g_{1i}}{\partial \phi}\right)'$ é o gradiente de $g_{1i}(\boldsymbol{\psi})$ em σ_v^2 e ϕ, o qual é um vector de dimensão 2×1 com elementos:

$$\frac{\partial g_{1i}}{\partial \sigma_v^2} = \mathbf{m}_i'\left\{\mathbf{D}^{-1} - \left[\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1} + \sigma_v^2\mathbf{D}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{Z}\sigma_v^2\mathbf{D}^{-1} + \sigma_v^2\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}^{-1}\right]\right\}\mathbf{m}_i \tag{4.5.33}$$

e

$$\frac{\partial g_{1i}}{\partial \phi} = \mathbf{m}_i'\left\{\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1} - \left[\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1} + \sigma_v^2\mathbf{D}^{-1}\mathbf{Z}'\left(-\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\right)\mathbf{Z}\sigma_v^2\mathbf{D}^{-1} + \sigma_v^2\mathbf{D}^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\sigma_v^2\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\right]\right\}\mathbf{m}_i \tag{4.5.34}$$

Por último, é de notar que Salvati (2004) não estudou se o estimador (4.5.31) é centrado até à ordem $o\left(m^{-1}\right)$, tal como se verifica para a esmagadora maioria dos estimadores do EQMP dos EBLUP apresentados ao longo deste texto.

4.6 MODELO DE NÍVEL ÁREA COM DADOS ESPACIAIS E CRONOLÓGICOS

4.6.1 Introdução

A associação espacial existente entre pequenos domínios vizinhos, bem como a autocorrelação temporal existente entre as observações de um domínio particular ao longo do tempo, têm sido utilizadas, em separado, com sucesso em abordagens de estimação *model-based*, com o objectivo de melhorar a qualidade dos estimadores directos para pequenos domínios. Contudo, em raras situações se tem explorado simultaneamente a correlação espacial e a autocorrelação cronológica com esse objectivo. Singh *et al*. (2005) foram os primeiros investigadores a utilizar modelos espaciotemporais de nível área para estimação em pequenos domínios. Estes autores propuseram uma extensão de um modelo espacial para dados temporais, utilizando para tal a metodologia dos modelos *state space*. O modelo espaciotemporal proposto por Singh *et al*. (2005), apresentado na secção seguinte, enquadra-se na classe de modelos *state space* lineares exposta no subcapítulo 3.3.

4.6.2 Modelo *state space* de Singh-Shukla-Kundu

Sejam $\boldsymbol{\theta}_t = col_{1\le i\le m}\left(\theta_{it}\right)$ e $\mathbf{y}_t = col_{1\le i\le m}\left(y_{it}\right)$ vectores $m\times 1$, respectivamente, dos parâmetros populacionais da variável de interesse e dos seus estimadores directos não enviesados no desenho associados ao período de tempo t (i=1, ..., m; t=1, …, T), $\mathbf{X}_t = col_{1\le i\le m}(\mathbf{x}'_{it})$ uma matriz $m\times p$ conhecida onde $\mathbf{x}_{it} = \left(x_{it1}, ..., x_{itp}\right)'$ é um vector p-dimensional de variáveis explicativas associadas ao i-ésimo pequeno domínio no

período t (i=1,..., m; t=1,..., T) e $\mathbf{W}$ uma matriz $m \times m$ de pesos espaciais conhecidos. O modelo proposto por Singh *et al*. (2005) tem a seguinte forma:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}_t + \boldsymbol{\varepsilon}_t \\ \mathbf{Z} &= \left(\mathbf{I}_m - \phi\mathbf{W}\right)^{-1}, \end{aligned} \tag{4.6.1}$$

onde $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)'$ é um vector $p \times 1$ de coeficientes de regressão, $\mathbf{Z}$ é uma matriz $m \times m$ que contém os coeficientes dos efeitos aleatórios de domínio, ϕ é um coeficiente de associação espacial, $\mathbf{v}_t = col_{1\le i\le m}\left(v_{it}\right)$ é um vector $m \times 1$ dos efeitos aleatórios específicos de domínio e $\boldsymbol{\varepsilon} = col_{1\le i\le m}\left(\varepsilon_i\right)$ é um vector $m \times 1$ dos erros da sondagem com $\boldsymbol{\varepsilon}_t \overset{ind}{\sim} N\left(\mathbf{0}; \mathbf{R}_t\right)$, sendo $\mathbf{R}_t = diag_{1\le i\le m}\left(\sigma_{it}^2\right)$ uma matriz de dimensão $m \times m$ com variâncias amostrais conhecidas. A evolução dos efeitos aleatórios ao longo do tempo, t=1, ..., T, é determinada da seguinte forma:

$$\mathbf{v}_t = \rho\mathbf{v}_{t-1} + \boldsymbol{\eta}_t, \tag{4.6.2}$$

onde ρ é um coeficiente de autocorrelação temporal (admitindo-se que $|\rho| < 1$ de forma a garantir a estacionaridade) e $\boldsymbol{\eta}_t = col_{1\le i\le m}\left(\eta_{it}\right)$ é um vector $m \times 1$ de erros do modelo com $\boldsymbol{\eta}_t \overset{ind}{\sim} N\left(\mathbf{0}; \sigma_v^2\mathbf{I}_m\right)$. Como habitualmente, estes autores assumiram que $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são mutuamente independentes. Para além dos parâmetros dos efeitos fixos e dos efeitos aleatórios, este modelo contém mais três parâmetros de variância desconhecidos independentes do período de tempo, t, os quais são apresentados no seguinte vector de parâmetros: $\boldsymbol{\psi} = \left(\phi, \sigma_v^2, \rho\right)'$.

Quando reescrito de forma a obedecer à formulação clássica do modelo *state space*, o modelo definido por (4.6.1) e (4.6.2) é especificado pelas seguintes duas equações:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{U}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\alpha}_t &= \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\varsigma}_t, \end{aligned}, \tag{4.6.3}$$

onde $\mathbf{U}_t = \left(\mathbf{X}_t, \mathbf{Z}\right)$, $\boldsymbol{\alpha}_t = \left(\boldsymbol{\beta}_t', \mathbf{v}_t'\right)'$, $\mathbf{T} = diag\left(\mathbf{I}_p, \rho\mathbf{I}_m\right)$, $\boldsymbol{\varsigma}_t = \left(\boldsymbol{\xi}_t', \boldsymbol{\eta}_t'\right)'$ tal que $\boldsymbol{\varsigma}_t \sim N\left(\mathbf{0}; \mathbf{Q}\right)$ com $\mathbf{Q} = diag\left(\mathbf{0}_p, \sigma_v^2\mathbf{I}_m\right)$. Singh *et al.* (2005) deduziram o BLUP espaciotemporal de $\boldsymbol{\theta}_t$

e o seu EQMP tendo, por base as seguintes equações recorrentes do filtro de kalman: $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \tilde{\boldsymbol{\alpha}}_{t|t-1}(\boldsymbol{\psi}) = \mathbf{T}\tilde{\boldsymbol{\alpha}}_{t-1}$, $\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{t|t-1}(\boldsymbol{\psi}) = \mathbf{T}\boldsymbol{\Sigma}_{t-1}\mathbf{T}' + \mathbf{Q}$, $\mathbf{H}_t = \mathbf{H}_t(\boldsymbol{\psi}) = \mathbf{R}_t + \mathbf{U}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{U}_t'$, $\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_t(\boldsymbol{\psi}) = \tilde{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbf{U}_t'\mathbf{H}_t^{-1}(\mathbf{y}_t - \mathbf{U}_t\tilde{\boldsymbol{\alpha}}_{t|t-1})$, e $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t(\boldsymbol{\psi}) = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{U}_t'\mathbf{H}_t^{-1}\mathbf{U}_t\boldsymbol{\Sigma}_{t|t-1}$.

Assumindo que o vector de parâmetros, $\boldsymbol{\psi}$, é conhecido, então o BLUP de $\boldsymbol{\theta}_t$ e o seu EQMP são dados, respectivamente, por:

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\boldsymbol{\theta}}_t(\boldsymbol{\psi}) = \mathbf{U}_t\tilde{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{\Lambda}_t\mathbf{e}_t, \tag{4.6.4}$$

$$EQMP(\tilde{\boldsymbol{\theta}}_t) = g_{12t}(\boldsymbol{\psi}) = \mathbf{U}_t\boldsymbol{\Sigma}_t\mathbf{U}_t', \tag{4.6.5}$$

onde $\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_t(\boldsymbol{\psi}) = \mathbf{I}_m - \mathbf{R}_t\mathbf{H}_t^{-1}$ e $\mathbf{e}_t = \mathbf{e}_t(\boldsymbol{\psi}) = \mathbf{y}_t - \mathbf{U}_t\tilde{\boldsymbol{\alpha}}_{t|t-1}$. Nas aplicações práticas, o vector de parâmetros de variância, $\boldsymbol{\psi}$, é normalmente desconhecido. Singh *et al.* (2005) propuseram que esses parâmetros sejam estimados pelo método da MVR. Quando esses parâmetros desconhecidos são substituídos pelos seus estimadores na expressão do BLUP, obtém-se o seguinte EBLUP espaciotemporal:

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t(\hat{\boldsymbol{\psi}}) = \mathbf{U}_t(\hat{\boldsymbol{\psi}})\hat{\boldsymbol{\alpha}}_{t|t-1}(\hat{\boldsymbol{\psi}}) + \boldsymbol{\Lambda}_t(\hat{\boldsymbol{\psi}})\mathbf{e}_t(\hat{\boldsymbol{\psi}}). \tag{4.6.6}$$

Com base nos trabalhos de Kackar e Harville (1984), Singh *et al*. (2005) deduziram uma aproximação de segunda ordem para o EQMP do EBLUP para os casos em que os parâmetros, $\boldsymbol{\psi}$, são estimados pelo método da MVR. Assumindo que $m \to \infty$, que são ignorados todos os termos de ordem $o(m^{-1})$ e que se verificam as seguintes quatro condições de regularidade: (*i*) os elementos de $\mathbf{X}_t$, t=1, …, T são uniformemente limitados tal que $\mathbf{X}_t'\mathbf{V}_t^{-1}(\boldsymbol{\psi})\mathbf{X}_t = [o(m)]_{p\times p}$, onde $\mathbf{V}_t(\boldsymbol{\psi}) = \sigma_v^2\mathbf{A}^{-1}(\boldsymbol{\psi}) + \mathbf{R}_t$; (*ii*) m e T são finitos; (*iii*) $\boldsymbol{\Lambda}_t\mathbf{U}_t = [o(1)]_{m\times p}$, $\frac{\partial[\boldsymbol{\Lambda}_t\mathbf{U}_t]}{\partial\psi_e} = [o(1)]_{m\times p}$ e $\frac{\partial^2[\boldsymbol{\Lambda}_t]}{\partial\psi_e\partial\psi_f} = [o(1)]_{m\times m}$, para t=1, …, T e e,f=1, 2, 3; (*iv*) $\hat{\boldsymbol{\psi}}$ é o estimador de $\boldsymbol{\psi}$ que satisfaz as seguintes condições: $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} = o_p(m^{-1/2})$, $\hat{\boldsymbol{\psi}}(-\mathbf{y}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$, $\hat{\boldsymbol{\psi}}(\mathbf{y} - \mathbf{xh}) = \hat{\boldsymbol{\psi}}(\mathbf{y})$, $\forall\mathbf{h} \in \Re^p$ e $\forall\mathbf{y}$; então uma aproximação de segunda ordem do EQMP do EBLUP é dada por (Singh *et al.*, 2005):

$$EQMP(\hat{\boldsymbol{\theta}}_t) \approx g_{12t}(\boldsymbol{\psi}) + g_{3t}(\boldsymbol{\psi}), \tag{4.6.7}$$

onde $g_{12t}(\boldsymbol{\psi})$ é dada por (4.6.5). Por sua vez, a parcela $g_{3t}(\boldsymbol{\psi})$, que representa o enviesamento devido à estimação do vector de parâmetros $\boldsymbol{\psi}$ a partir dos dados amostrais, é de ordem $o(m^{-1})$ e é dada por:

$$g_{3t}(\boldsymbol{\psi}) = \mathbf{L}_t'(\boldsymbol{\psi})\left[\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi})\mathbf{K}(\boldsymbol{\psi})\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) \otimes \mathbf{H}_t\right] \mathbf{L}_t(\boldsymbol{\psi}), \tag{4.6.8}$$

onde $\mathbf{L}_t(\boldsymbol{\psi}) = col_{1\le i\le m}\left[\mathbf{L}_{ti}(\boldsymbol{\psi})\right]$ é uma matriz por blocos de dimensão $3m\times m$, na qual cada bloco de dimensão $m\times m$ é dado pela matriz $\mathbf{L}_{ti}(\boldsymbol{\psi}) = \dfrac{\partial \boldsymbol{\Lambda}_t(\boldsymbol{\psi})}{\partial \psi_e}$, e=1, 2, 3; $\mathbf{I}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \left\{I_{ef}(\boldsymbol{\psi})\right\}$ é uma matriz de informação de dimensão 3×3 com elemento genérico

$$\begin{aligned} I_{ef}(\boldsymbol{\psi}) = {} & \frac{1}{2}\sum_{t=1}^{T} tr\left(\mathbf{H}_t^{-1}\frac{\partial \mathbf{H}_t^{-1}}{\partial \psi_e}\mathbf{H}_t^{-1}\frac{\partial \mathbf{H}_t}{\partial \psi_f}\right) + \sum_{t=1}^{T}\left(\frac{\partial \mathbf{e}_t'}{\partial \psi_e}\mathbf{H}_t^{-1}\frac{\partial \mathbf{e}_t}{\partial \psi_f}\right) - \\ & -\frac{1}{2}tr\left[\left(\mathbf{X}_1'\mathbf{H}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{H}_1^{-1}\left(\frac{\partial^2 \mathbf{H}_1}{\partial \psi_e \partial \psi_f} - 2\frac{\partial \mathbf{H}_1}{\partial \psi_e}\mathbf{H}_1^{-1}\frac{\partial \mathbf{H}_1}{\partial \psi_f}\right)\mathbf{H}_1^{-1}\mathbf{X}_1\right] - \qquad , e,f=1, 2, 3 \\ & -\frac{1}{2}tr\left[\left(\mathbf{X}_1'\mathbf{H}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{H}_1^{-1}\frac{\partial \mathbf{H}_1}{\partial \psi_e}\mathbf{H}_1^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{H}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{H}_1^{-1}\frac{\partial \mathbf{H}_1}{\partial \psi_f}\mathbf{H}_1^{-1}\mathbf{X}_1\right] \end{aligned}$$

e $\mathbf{K}(\boldsymbol{\psi}) = \left\{k_{ef}(\boldsymbol{\psi})\right\}$ é uma matriz 3×3 com elemento genérico $k_{ef}(\boldsymbol{\psi}) = \dfrac{1}{2}\sum_{t=1}^{T} tr\left(\mathbf{H}_t^{-1}\dfrac{\partial \mathbf{H}_t}{\partial \psi_e}\mathbf{H}_t^{-1}\dfrac{\partial \mathbf{H}_t}{\partial \psi_f}\right)$, e, f=1, 2, 3. Tem-se ainda para t=1 que $\mathbf{X}_1$ é a matriz de desenho e $\mathbf{H}_1 = \mathbf{R}_1\sigma_v^2\mathbf{B}^{-1}$ com $\mathbf{B} = (\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})$.

Singh *et al*. (2005) também deduziram um estimador do EQMP do EBLUP, admitindo que $m \to \infty$ e que são ignorados todos os termos de ordem $o(m^{-1})$:

$$eqmp\left[\hat{\boldsymbol{\theta}}_t(\hat{\boldsymbol{\psi}})\right] = g_{12t}(\hat{\boldsymbol{\psi}}) + g_{3t}(\hat{\boldsymbol{\psi}}) + g_{31t}(\hat{\boldsymbol{\psi}}) - g_{4t}(\hat{\boldsymbol{\psi}}) - g_{5t}(\hat{\boldsymbol{\psi}}). \tag{4.6.9}$$

As parcelas $g_{12t}(\boldsymbol{\psi})$ e $g_{3t}(\boldsymbol{\psi})$ são dadas, respectivamente, por (4.6.5) e (4.6.8), enquanto as restantes são dadas por:

$$g_{31t}(\boldsymbol{\psi}) = \mathbf{L}_t'(\boldsymbol{\psi})\left[\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) \otimes \mathbf{H}_t\right]\mathbf{L}_t(\boldsymbol{\psi}), \tag{4.6.10}$$

$$g_{4t}(\boldsymbol{\psi}) = \left[\mathbf{c}'_{\hat{\psi}}(\boldsymbol{\psi}) \otimes \mathbf{I}_m\right] \frac{\partial g_{12t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}, \tag{4.6.11}$$

$$g_{5t}(\boldsymbol{\psi}) = \frac{1}{2} \underset{m}{tr}\left\{\left[\mathbf{I}_3 \otimes \left(\mathbf{R}_t \mathbf{H}_t^{-1}\right)\right] \frac{\partial^2 \mathbf{H}_t}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \left[\mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) \otimes \left(\mathbf{H}_t^{-1} \mathbf{R}_t\right)\right]\right\}, \tag{4.6.12}$$

onde $\mathbf{c}'_{\hat{\psi}}(\boldsymbol{\psi}) = \frac{1}{2} \mathbf{I}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) col_{1 \le e \le 3}\left\{tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}) \frac{\partial \mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})}{\partial \psi_e}\right]\right\}$ é uma matriz 3×1 e $\frac{\partial g_{12t}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = col_{1 \le e \le 3}\left[\frac{\partial g_{12t}(\boldsymbol{\psi})}{\partial \psi_e}\right]$ é uma matriz por blocos de dimensão $3m \times m$, sendo cada bloco de dimensão $m \times m$ dado por $\frac{\partial g_{12t}(\boldsymbol{\psi})}{\partial \psi_e} = \mathbf{R}_t \mathbf{H}_t^{-1} \frac{\partial \mathbf{R}_t}{\partial \psi_e} \mathbf{H}_t^{-1} \mathbf{H}_t^{-1}$. A expressão (4.6.9) representa uma matriz de ordem $m \times m$, cujos elementos principais são os estimadores do EQMP do EBLUP relativos aos *m* pequenos domínios particulares.

5. ESTIMAÇÃO EM PEQUENOS DOMÍNIOS UTILIZANDO UM MODELO LINEAR MISTO COM DADOS ESPACIAIS E CRONOLÓGICOS

5.1 INTRODUÇÃO

Tal como foi referido no subcapítulo 4.6, a associação espacial existente entre pequenos domínios, bem como a autocorrelação cronológica associada a um particular pequeno domínio, têm sido utilizadas, em separado, com sucesso em abordagens de estimação *model-based*, com o objectivo de alcançar melhores níveis de qualidade do que aqueles apresentados pelos estimadores directos para pequenos domínios. Contudo, até ao momento ainda não foi explorada simultaneamente a associação espacial e a autocorrelação cronológica em estudos publicados de estimação em pequenos domínios baseados em modelos lineares mistos de nível área. Neste contexto, pretende apresentar-se neste capítulo uma primeira metodologia de estimação em pequenos domínios com base num modelo espaciotemporal de nível área, incluído na classe dos modelos lineares mistos.

A classe dos modelos lineares mistos apresenta-se bastante adequada para representar realidades que possam ser descritas através de dados de natureza seccional/espacial e cronológica. Esta característica dos modelos lineares mistos é sublinhada, por exemplo, por Baltagi *et al*. (2007). No contexto da estimação em pequenos domínios com dados de natureza seccional/espacial e/ou cronológica, encontram-se o modelo longitudinal de Rao-Yu e os modelos *state space* de Pfeffermann-Burck e de Singh-Shukla-Kundu, apresentados no subcapítulos 4.3, 4.4 e 4.6, respectivamente. No primeiro modelo, admite-se que os efeitos aleatórios específicos de domínio são independentes, o que

raramente se verifica nas aplicações práticas, enquanto no segundo tipo de modelos é necessário recorrer a uma equação de transição rígida para especificar completamente o modelo. Naturalmente que estes modelos podem não ser suficientemente flexíveis ao ponto de conseguirem representar todo o tipo de realidades que podem estar presentes em dados de natureza seccional/espacial e cronológica. Deste modo, parece potencialmente interessante explorar toda a flexibilidade oferecida pelo modelo linear misto para representar essas realidades, através da utilização de um modelo com uma estrutura de erro espacialmente associada e temporalmente autocorrelacionada.

A abordagem proposta em seguida para estimar parâmetros em pequenos domínios, quando está disponível informação de natureza seccional/espacial e cronológica, tenta explorar toda essa flexibilidade. Nessa abordagem considera-se que:

1. as observações disponíveis resultam de um inquérito longitudinal (com ou sem rotação) realizado ao longo de T períodos de tempo;

2. é utilizado um modelo que integra dados amostrais agrupados ao nível de domínio, pelo facto de não ser possível fazer a ligação entre as observações da variável de interesse e das variáveis auxiliares ao nível individual;

3. é utilizado um modelo linear misto para incorporar informação de natureza espacial e cronológica na estrutura de erro;

4. a variabilidade existente entre pequenos domínios não explicada pelos efeitos fixos do modelo, pode ser acomodada no modelo através de efeitos aleatórios associados a diferentes níveis de agregação: domínio e domínio-tempo;

5. podem ser especificadas diferentes estruturas de covariância sobre os efeitos aleatórios do modelo, dependendo da natureza dos dados;

6. o objectivo principal consiste em estimar características de uma população finita (parâmetro da variável de interesse) e não de uma superpopulação.

Esta abordagem de modelação conjunta de dados de diversos períodos permite a estimação de parâmetros em períodos passados, utilizando informação referente a períodos mais recentes. Desta forma, torna-se igualmente possível actualizar estimativas para momentos passados, à medida que mais informação se vai tornando disponível.

Neste capítulo é proposto um estimador para um parâmetro de interesse em pequenos domínios, assistido por um modelo espaciotemporal que se enquadra na classe dos modelos lineares mistos. O estimador proposto, considerado como um EBLUP dos parâmetros de interesse, pretende constituir uma alternativa aos estimadores directos e indirectos "tradicionais" utilizados na estimação em pequenos domínios. Neste capítulo também é considerado o difícil problema da medição da incerteza associada ao EBLUP, tomando em consideração a variabilidade introduzida pela estimação das componentes de variância. Assim, no subcapítulo 5.2 é efectuada a especificação do modelo espaciotemporal de nível área. Em seguida, nos subcapítulos 5.3 e 5.4 são deduzidos o BLUP e o EBLUP espaciotemporal dos parâmetros de interesse, respectivamente. Os estimadores das componentes de variância são propostos no subcapítulo 5.5. Nos subcapítulos 5.6, 5.7 e 5.8 são propostos estimadores do EQMP do EBLUP espaciotemporal. Por último, no subcapítulo 5.9 é apresentado o modelo espacial de Salvati, como caso particular do modelo espaciotemporal aqui proposto, para o qual é proposto um estimador da componente de variância pelo método dos momentos.

5.2 ESPECIFICAÇÃO DO MODELO ESPACIOTEMPORAL DE NÍVEL ÁREA

Sejam $\boldsymbol{\theta} = col_{1\le i\le m}(\boldsymbol{\theta}_i)$ e $\mathbf{y} = col_{1\le i\le m}(\mathbf{y}_i)$, vectores $mT\times 1$, respectivamente, dos parâmetros populacionais da variável de interesse e dos seus estimadores directos não enviesados no desenho associados ao domínio i, onde $\boldsymbol{\theta}_i = col_{1\le t\le T}(\theta_{it})$ e $\mathbf{y}_i = col_{1\le t\le T}(y_{it})$ (i=1,..., m; t=1, …, T). Assume-se que os estimadores directos existem sempre que a dimensão amostral no domínio seja não nula, $n_{it} \ge 1$. Seja ainda $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$ uma matriz $mT\times p$ conhecida, onde $\mathbf{X}_i = col_{1\le t\le T}(\mathbf{x}'_{it})$ e na qual $\mathbf{x}_{it} = (x_{it1}, ..., x_{itp})'$ é um vector p-dimensional de variáveis explicativas associadas ao i-ésimo pequeno domínio no período t (i=1, ..., m; t=1, ..., T). O modelo de erro da sondagem é dado por:

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{5.2.1}$$

onde $\boldsymbol{\varepsilon} = col_{1\le i\le m}(\boldsymbol{\varepsilon}_i)$ é um vector $mT\times 1$ de erros da sondagem com $\boldsymbol{\varepsilon}_i = col_{1\le t\le T}(\varepsilon_{it})$, e satisfazendo $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0};\mathbf{R})$, sendo $\mathbf{R} = diag_{1\le i\le m;1\le t\le T}(\sigma_{it}^2)$ uma matriz $mT\times mT$ com variâncias amostrais conhecidas, dados os θ_{it}. Propõe-se o seguinte modelo de ligação, no qual o vector de parâmetros de interesse está relacionado com a matriz de variáveis auxiliares através de um modelo linear com efeitos aleatórios:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{v} + \mathbf{u}_2, \tag{5.2.2}$$

onde $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ é um vector $p\times 1$ de parâmetros de regressão; $\mathbf{v} = col_{1\le i\le m}(v_i)$ é um vector $m\times 1$ que representa a variação espacial de segunda ordem ao nível de domínio, a qual reflecte a estrutura de dependência espacial do processo; $\mathbf{u}_2 = col_{1\le i\le m}(\mathbf{u}_{2i})$ é um vector $mT\times 1$ de efeitos aleatórios específicos de domínio-tempo com $\mathbf{u}_{2i} = col_{1\le t\le T}(u_{2it})$, que incorporam a estrutura de dependência temporal do processo; e $\mathbf{Z}_1 = \mathbf{I}_m \otimes \mathbf{1}_T$. A inclusão de associações espaciais entre os efeitos aleatórios específicos de domínio é efectuada através do seguinte processo SAR:

$$\mathbf{v} = \phi\mathbf{W}\mathbf{v} + \mathbf{u}_1 \Rightarrow \mathbf{v} = (\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{u}_1, \tag{5.2.3}$$

onde ϕ é um coeficiente de associação espacial, $\mathbf{W}$ é uma matriz $m\times m$ de pesos espaciais conhecidos e com os elementos principais nulos, e $\mathbf{u}_1 = col_{1\le i\le m}(u_{1i})$ é um vector $m\times 1$ de erros do processo com $\mathbf{u}_1 \overset{iid}{\sim} N(\mathbf{0};\sigma_u^2\mathbf{I}_m)$. Desta forma, assume-se que os efeitos aleatórios de domínios vizinhos não são independentes entre si, podendo as unidades populacionais de um lado da fronteira estar ligadas, através dos pesos espaciais, às unidades populacionais do outro lado da fronteira. Note-se que a matriz $\mathbf{W}$ descreve como os efeitos aleatórios de domínios vizinhos estão relacionados, enquanto ϕ define a força dessa relação espacial. Por sua vez, a inclusão de correlações temporais associadas aos efeitos aleatórios específicos de domínio-tempo é efectuada através do seguinte processo AR(1):

$$u_{2,it} = \rho u_{2,i,t-1} + \xi_{it}, \quad |\rho| < 1 \tag{5.2.4}$$

onde ξ_{it}'s são os erros do processo a satisfazer $\xi_{it} \overset{iid}{\sim} N(0;\sigma^2)$ e ρ é um coeficiente de autocorrelação temporal. Assume-se, portanto, que existe uma relação explícita entre os efeitos aleatórios de cada domínio num dado momento do tempo e no momento de tempo imediatamente anterior[51]. O modelo combinado com erros correlacionados espacialmente e temporalmente, baseado em (5.2.1)-(5.2.4), é dado por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \tag{5.2.5}$$

onde $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{I}_{mT}]$, $\mathbf{Z}_1 = \mathbf{I}_m \otimes \mathbf{1}_T$ e $\boldsymbol{\upsilon} = [\mathbf{v}' \quad \mathbf{u}_2']'$ e no qual se assume que os termos de erro $\mathbf{v} = (\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{u}_1$, $\mathbf{u}_2$ e $\boldsymbol{\varepsilon}$ são mutuamente independentes. Tem-se neste modelo que $\mathbf{u}_1 \overset{iid}{\sim} N(\mathbf{0};\sigma_u^2\mathbf{I}_m)$, $\mathbf{u}_2 \sim N(\mathbf{0};\sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma})$ e $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0};\mathbf{R})$, onde $\boldsymbol{\Gamma} = \{\gamma_{rs}\}$ é uma matriz $T \times T$ com elementos $\gamma_{rs} = \rho^{|r-s|}/(1-\rho^2)$, r, s=1, ..., T e $\mathbf{R} = diag_{1\le i\le m}(\mathbf{R}_i)$ com $\mathbf{R}_i = diag_{1\le t\le T}(\sigma_{it}^2)$.

É fácil verificar que o modelo (5.2.5) é um caso particular do modelo linear misto (3.2.1), com matriz de covariâncias de $\boldsymbol{\upsilon}$ diagonal por blocos, dada por $\mathbf{G} = diag_{1\le k\le 2}(\mathbf{G}_k)$, onde $\mathbf{G}_1$ e $\mathbf{G}_2$ representam, respectivamente, as matrizes de covariâncias de $\mathbf{v}$ e de $\mathbf{u}_2$. Tal como demonstrado por Salvati (2004) e por Rao e Yu (1994), estas estruturas de covariância são dadas, respectivamente, por $\mathbf{G}_1 = E(\mathbf{v}\mathbf{v}') = \sigma_u^2\left[(\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})\right]^{-1}$ e por $\mathbf{G}_2 = E(\mathbf{u}_2\mathbf{u}_2') = \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma}$. Neste caso, a matriz de covariâncias de $\mathbf{y}$ é então dada por:

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}' = diag_{1\le i\le m;1\le t\le T}(\sigma_{it}^2) + \mathbf{Z}_1\sigma_u^2\mathbf{B}^{-1}\mathbf{Z}_1' + \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma}, \tag{5.2.6}$$

onde $\mathbf{B} = (\mathbf{I}_m - \phi\mathbf{W})'(\mathbf{I}_m - \phi\mathbf{W})$. Note-se que a matriz (5.2.6) não é uma matriz diagonal por blocos, ao contrário do que se verifica nos conhecidos modelos de Rao-Yu e de Fay-Herriot.

[51] Naturalmente que esta relação entre os efeitos aleatórios associados a um determinado domínio nem sempre se verifica, podendo eventualmente ser do tipo AR(p) ou MA(q), com $p\ge2$ ou $q\ge1$. Como será apresentado no estudo empírico, no âmbito do problema de estimação em pequenos domínios em estudo, basta assumir que os efeitos aleatórios seguem um processo AR(1).

No contexto do modelo (5.2.5), supõe-se que as variáveis explicativas (ou auxiliares), $\mathbf{x}_{it}$, representam informação conhecida sobre toda a população ao nível de cada domínio em cada período de tempo, sendo na maior parte dos casos oriunda de fontes administrativas ou de recenseamentos. Não existe nenhum obstáculo relativamente à disponibilidade dessas variáveis a um nível mais desagregado (subdomínios) ou até mesmo ao nível das unidades individuais. Nestes casos, pode calcular-se uma medida que agregue toda essa informação para cada variável ao nível de domínio (por exemplo, uma média aritmética ou um total). Por sua vez, admite-se que os efeitos aleatórios representam as características específicas associadas aos domínios de interesse que afectam o parâmetro da variável de interesse, mas que não são consideradas nas variáveis explicativas do modelo. É, ainda, de salientar que esses efeitos só representam características que não ultrapassam as fronteiras do domínio.

O facto do modelo proposto envolver efeitos aleatórios de domínios espacialmente dependentes através de um processo SAR e efeitos aleatórios de domínio-tempo temporalmente autocorrelacionados através de um processo AR(1), confere-lhe um vasto leque de aplicação a problemas práticos nas áreas da economia, gestão, ambiente, saúde, entre outras. Na maior parte dos problemas é razoável assumir que os efeitos aleatórios associados a domínios vizinhos estão associados, assim como efeitos aleatórios associados a um determinado domínio e referentes a períodos de tempo próximos estão correlacionados e que essa autocorrelação decai para zero com o aumento da distância temporal. É ainda de salientar que a especificação de uma estrutura que estabelece uma correlação temporal entre os efeitos aleatórios associados a um determinado domínio, confere uma grande mais-valia ao modelo. Na verdade, a consideração de correlação temporal entre os efeitos aleatórios leva ao aumento da informação disponível, relativa a outros períodos de tempo passados, para a estimação/predição do modelo, permitindo melhorar as propriedades dos estimadores dos parâmetros de interesse em cada um dos domínios. Baltagi (2005) refere também que a não consideração de uma estrutura de correlação temporal na estrutura de erro do modelo linear misto quando os dados são temporais, pode conduzir a estimativas dos coeficientes de regressão ineficientes, embora consistentes.

Por último, note-se que a partir do modelo (5.2.5) pode ser obtido o modelo de Rao-Yu (4.3.5) fazendo $\phi = 0$, bem como o modelo de Salvati (4.5.4) fazendo T=1, ρ=0 e

$\sigma^2 = 0$. A partir do modelo (5.2.5) pode ainda ser obtido o modelo de Fay-Herriot (4.2.5) fazendo $\phi = 0$, T=1, ρ=0 e $\sigma^2 = 0$.

5.3 O *BEST LINEAR UNBIASED PREDICTOR* (BLUP)

O valor do parâmetro da variável de interesse num domínio i no período t, $\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + v_i + u_{2it}$, é um caso especial da combinação linear $\tau = \mathbf{k}'_{it}\boldsymbol{\beta} + \mathbf{m}'_{it}\boldsymbol{\upsilon}$, onde $\mathbf{k}'_{it} = \mathbf{x}'_{it}$ e $\mathbf{m}'_{it} = [\mathbf{m}'_{1i} \quad \mathbf{m}'_{2it}]$, no qual $\mathbf{m}'_{1i} = (0, ..., 0, 1, 0, ..., 0)$ é um vector linha m-dimensional com o número um na i-ésima posição e zeros nas outras posições, e $\mathbf{m}'_{2it} = (0, ..., 0, 1, 0, ..., 0)$ é um vector linha mT-dimensional com o número um na (it)-ésima posição e zeros nas outras posições. Notando-se que o modelo (5.2.5) escrito na forma compacta é um caso especial do modelo linear misto, então o BLUP de $\tau = \theta_{it}$ também pode ser obtido a partir dos resultados gerais deduzidos por Henderson (1975).

Assumindo que as componentes de variância $\boldsymbol{\psi} = (\sigma^2, \sigma_u^2, \phi, \rho)'$ são conhecidas, ou seja, que a matriz $\mathbf{G}$ é conhecida, então o BLUP de θ_{it} é dado por:

$$\tilde{\theta}_{it} = \tilde{\theta}_{it}^{H}(\boldsymbol{\psi}) = \mathbf{k}'_{it}\tilde{\boldsymbol{\beta}} + \mathbf{m}'_{it}\tilde{\boldsymbol{\upsilon}}, \tag{5.3.1}$$

onde $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ é o estimador dos mínimos quadrados generalizados de $\boldsymbol{\beta}$ e $\tilde{\boldsymbol{\upsilon}} = \tilde{\boldsymbol{\upsilon}}(\boldsymbol{\psi}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ é o BLUP de $\boldsymbol{\upsilon}$. Utilizando as estruturas $\mathbf{k}'_{it}$, $\mathbf{m}'_{it}$, $\mathbf{G}$ e $\mathbf{V}$, e assumindo que são conhecidas as quatro componentes de variância do modelo, $\boldsymbol{\psi}$, obtém-se o seguinte BLUP de $\boldsymbol{\upsilon}$:

$$
\begin{aligned}
\tilde{\mathbf{\upsilon}} &= diag_{1\le k\le 2}(\mathbf{G}_k)\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)= \\
&= \begin{bmatrix} \sigma_u^2\mathbf{B}^{-1} & \mathbf{0}_{m\times mT} \\ \mathbf{0}_{mT\times m} & \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma} \end{bmatrix}\begin{bmatrix} \mathbf{I}_m\otimes\mathbf{1}_T' \\ \mathbf{I}_{mT} \end{bmatrix}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \begin{bmatrix} \sigma_u^2\mathbf{B}^{-1}\left(\mathbf{I}_m\otimes\mathbf{1}_T'\right) \\ \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma} \end{bmatrix}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \begin{bmatrix} \sigma_u^2\mathbf{B}^{-1}\otimes\mathbf{1}_T' \\ \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma} \end{bmatrix}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{\Lambda}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)
\end{aligned}
\qquad (5.3.2)
$$

assim como o seguinte BLUP espaciotemporal de θ_{it}:

$$
\begin{aligned}
\tilde{\theta}_{it} = \tilde{\theta}_{it}^{H}(\boldsymbol{\psi}) &= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}+\mathbf{m}_{it}'\boldsymbol{\Lambda}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}+\begin{bmatrix}\mathbf{m}_{1i}' & \mathbf{m}_{2it}'\end{bmatrix}\begin{bmatrix} \sigma_u^2\mathbf{B}^{-1}\otimes\mathbf{1}_T' \\ \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma} \end{bmatrix}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}+\left(\mathbf{m}_{1i}'\sigma_u^2\mathbf{B}^{-1}\otimes\mathbf{1}_T'+\mathbf{m}_{2it}'\sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma}\right)\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}+\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\
&= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}+\mathbf{h}_i'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right),
\end{aligned}
\qquad (5.3.3)
$$

onde $\mathbf{h}_i'=\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'$, $\boldsymbol{\varsigma}_i'=\{\varsigma_{ii'}\}$ é a *i*-ésima linha da matriz $\mathbf{B}^{-1}$ e $\boldsymbol{\zeta}_{it}'$ é um vector linha de dimensão $1\times mT$ com *m* blocos, no qual cada bloco é formado por um vector linha *T*-dimensional, sendo o *i*-ésimo bloco formado pela *t*-ésima linha da matriz $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}_t$, e os restantes blocos por vectores nulos $\mathbf{0}_{1\times T}$, *i*,*i*'=1, …, *m*. Note-se que não é possível apresentar a segunda parcela do BLUP ao nível de domínio-tempo, porque a matriz $\mathbf{V}$ não é uma matriz diagonal por blocos.

O estimador obtido pode ser classificado como um estimador combinado, uma vez que pode ser decomposto em duas componentes: um estimador sintético, $\mathbf{x}_{it}'\tilde{\boldsymbol{\beta}}$, e um factor de correcção, $\mathbf{h}_i'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)$, que é uma função das diferenças entre as estimativas directas e as estimativas sintéticas do parâmetro de interesse. Pode afirmar-se que os pesos reflectidos em $\mathbf{h}_i'\mathbf{V}^{-1}$ permitem que o estimador sintético seja corrigido pelos erros de predição do domínio que é alvo de inferência no período *t* e nos períodos anteriores correlacionados cronologicamente com o período *t*, mas também pelos erros

de predição referentes ao período *t* de outros domínios, associados espacialmente com o domínio alvo de inferência.

Notando que $\mathbf{h}_i' = \{\mathbf{h}_{ii'}'\}$ é um vector linha de dimensão $1\times mT$ com *m* blocos, no qual o *i*-ésimo bloco, de dimensão $1\times T$, é dado por $\mathbf{h}_{ii}' = \sigma_u^2 \varsigma_{ii} \mathbf{1}_T' + \sigma^2 \boldsymbol{\gamma}_t'$ e os restantes blocos são dados por $\mathbf{h}_{ii'}' = \sigma_u^2 \varsigma_{ii'} \mathbf{1}_T'$, $i \neq i'$, $i,i'=1, \ldots, m$, então o BLUP (5.3.3) pode ainda ser apresentado como:

$$\begin{aligned} \tilde{\theta}_{it} &= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}} + \mathbf{h}_i'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) \\ &= \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}} + \left[col_{1\le i'\le m}\left(\mathbf{h}_{ii'}\right)\right]'\mathbf{V}^{-1}\left[col_{1\le i'\le m}\left(\mathbf{y}_{i'} - \mathbf{X}_{i'}\tilde{\boldsymbol{\beta}}\right)\right]. \end{aligned} \tag{5.3.4}$$

A partir da expressão anterior é possível observar que quando um determinado domínio *i* não está representado na amostra da *t*-ésima vaga, continua a ser possível fazer predições para o factor de correcção associado a esse domínio, tirando partido da sua potencial associação espacial e autocorrelação cronológica. Para tal, basta que esse domínio *i* apresente pelo menos uma relação de dependência espacial com outro domínio *i*' de dimensão amostral não nula, e/ou que existam observações amostrais referentes ao domínio *i* em pelo menos uma das vagas anteriores. Esta é, sem dúvida, uma característica muito apelativa do estimador proposto e uma grande vantagem em relação aos estimadores apresentados no capítulo quarto, e em particular ao estimador EBLUP temporal, pois é possível evitar que o estimador proposto se reduza a um estimador sintético "puro", mesmo quando a dimensão amostral observada no domínio que é alvo de inferência é nula para todos os períodos de tempo.

5.4 O EBLUP ESPACIOTEMPORAL

O BLUP espaciotemporal de θ_{it} (5.3.3) depende das componentes de variância, $\boldsymbol{\psi}$, as quais são geralmente desconhecidas na prática. Assumindo-se que as componentes de associação espacial e de autocorrelação temporal são conhecidas[52], à semelhança do que foi admitido no modelo de Rao e Yu (1994), então o preditor em dois passos de θ_{it} é

[52] Segundo Rao (2003), é frequentemente admitido que este tipo de parâmetros é conhecido devido à dificuldade em estimá-los de forma consistente e admissível.

obtido quando se substitui em (5.3.3) σ^2 e σ_u^2 por estimadores assimptoticamente consistentes, $\hat{\sigma}^2$ e $\hat{\sigma}_u^2$, respectivamente. O EBLUP espaciotemporal resultante é então dado por:

$$\hat{\theta}_{it} = \hat{\theta}_{it}^{H}\left(\phi, \rho, \hat{\sigma}^2, \hat{\sigma}_u^2\right) = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}(\phi,\rho) + \left(\hat{\sigma}_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \hat{\sigma}^2 \boldsymbol{\zeta}_{it}'\right)\hat{\mathbf{V}}^{-1}(\phi,\rho)\left[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\phi,\rho)\right], \quad (5.4.1)$$

onde $\hat{\boldsymbol{\beta}}(\phi,\rho)$ e $\hat{\mathbf{V}}^{-1}(\phi,\rho)$ são os estimadores de $\boldsymbol{\beta}(\phi,\rho)$ e de $\mathbf{V}^{-1}(\phi,\rho)$, respectivamente, quando $\sigma^2(\phi,\rho)$ e $\sigma_u^2(\phi,\rho)$ são substituídos por $\hat{\sigma}^2(\phi,\rho)$ e $\hat{\sigma}_u^2(\phi,\rho)$, respectivamente.

No contexto deste modelo, define-se a partir deste ponto o vector de componentes de variância como $\boldsymbol{\psi} = \left[\sigma^2(\phi,\rho), \sigma_u^2(\phi,\rho)\right]'$, pelo facto de se assumir que ϕ e ρ são conhecidos.

5.5 ESTIMAÇÃO DAS COMPONENTES DE VARIÂNCIA

Tal como já foi referido anteriormente, as componentes de variância são geralmente desconhecidas nas aplicações práticas, sendo necessária a sua estimação. Nesta secção propõe-se uma extensão do método III de Henderson (1953)[53] para a estimação das componentes de variância, σ^2 e σ_u^2, no contexto do modelo (5.2.5), com erros associados espacialmente através de um processo SAR, v_i, erros autocorrelacionados temporalmente através de um processo AR(1), $u_{2,it}$, e erros da sondagem independentes, ε_{it}. Admite-se, contudo, que ϕ e ρ são conhecidos. Os estimadores propostos para as componentes de variância são baseados nos resíduos de regressões lineares efectuadas pelo método dos mínimos quadrados ordinários sobre o modelo (5.2.5) transformado.

[53] Rao e Yu (1994) também propuseram uma extensão do método III de Henderson (1953) para a estimação das componentes de variância, σ^2 e σ_v^2, no contexto do seu modelo, o qual envolve efeitos aleatórios de domínio-tempo autocorrelacionados temporalmente, mas efeitos aleatórios de domínio independentes.

Começa por propor-se um estimador não enviesado para σ^2. Para tal é necessário transformar o modelo (5.2.5) de forma a eliminar o vector de efeitos aleatórios $\mathbf{v}$. Em primeiro lugar transforma-se $\mathbf{y}_i$ em $\mathbf{z}_i = \mathbf{P}\mathbf{y}_i$, de forma a que $V(\mathbf{P}\mathbf{u}_{2i}) = \sigma^2\mathbf{I}_T$, ou seja, transformam-se os efeitos aleatórios de domínio-tempo em efeitos aleatórios não correlacionados. Nessa transformação, conhecida como transformação de Prais-Winsten, utiliza-se a decomposição $\boldsymbol{\Gamma} = \mathbf{P}^{-1}(\mathbf{P}^{-1})'$, onde $\mathbf{P}$ é uma matriz $T\times T$ com a seguinte forma: $p_{11} = (1-\rho^2)^{1/2}$, $p_{t,t'} = 1, \forall t = t'$ para $t, t' = 2,...,T$, $p_{t+1,t} = -\rho$ para $t = 1,...,T-1$ e os restantes elementos são $p_{t,t'} = 0$ (Judge *et al.*, 1985). Pré-multiplicando o modelo (5.2.5) por $diag_{1\le i\le m}(\mathbf{P})$, obtém-se o seguinte modelo transformado:

$$diag_{1\le i\le m}(\mathbf{P})\mathbf{y} = diag_{1\le i\le m}(\mathbf{P})\mathbf{X}\boldsymbol{\beta} + diag_{1\le i\le m}(\mathbf{P})\mathbf{Z}_1\mathbf{v} + diag_{1\le i\le m}(\mathbf{P})\mathbf{u}_2 + diag_{1\le i\le m}(\mathbf{P})\boldsymbol{\varepsilon}$$

$$col_{1\le i\le m}(\mathbf{P}\mathbf{y}_i) = col_{1\le i\le m}(\mathbf{P}\mathbf{X}_i)\boldsymbol{\beta} + diag_{1\le i\le m}(\mathbf{P})diag_{1\le i\le m}(\mathbf{1}_T)\mathbf{v} + col_{1\le i\le m}(\mathbf{P}\mathbf{u}_{2i}) + col_{1\le i\le m}(\mathbf{P}\boldsymbol{\varepsilon}_i)$$

$$col_{1\le i\le m}(\mathbf{z}_i) = col_{1\le i\le m}(\mathbf{H}_i)\boldsymbol{\beta} + diag_{1\le i\le m}(\mathbf{f})\mathbf{v} + col_{1\le i\le m}(\mathbf{P}\mathbf{u}_{2i}) + col_{1\le i\le m}(\mathbf{P}\boldsymbol{\varepsilon}_i), \qquad (5.5.1)$$

onde $\mathbf{z}_i = \mathbf{P}\mathbf{y}_i$, $\mathbf{H}_i = \mathbf{P}\mathbf{X}_i$ e $\mathbf{f} = \mathbf{P}\mathbf{1}_T = col_{1\le t\le T}(f_t)$, com $f_1 = (1-\rho^2)^{1/2}$ e $f_t = 1-\rho$ para $2 \le t \le T$.

Em segundo lugar transforma-se $\mathbf{z}_i$ em $\mathbf{z}_i^{(1)} = (\mathbf{I}_T - \mathbf{D})\mathbf{z}_i$, de forma a eliminar o vector de efeitos aleatórios $\mathbf{v}$, onde $\mathbf{D} = (\mathbf{f}\mathbf{f}')/c$ é uma matriz $T\times T$ com $c = \mathbf{f}'\mathbf{f} = (1-\rho)[T-(T-2)\rho]$. Pré-multiplicando agora o modelo (5.5.1) por $diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})$, obtém-se o seguinte modelo transformado:

$$\begin{aligned} diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})col_{1\le i\le m}(\mathbf{z}_i) = {} & diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})col_{1\le i\le m}(\mathbf{H}_i)\boldsymbol{\beta} + diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})diag_{1\le i\le m}(\mathbf{f})\mathbf{v} + \\ & + diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})col_{1\le i\le m}(\mathbf{P}\mathbf{u}_{2i}) + diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})col_{1\le i\le m}(\mathbf{P}\boldsymbol{\varepsilon}_i) \end{aligned}$$

$$\begin{aligned} col_{1\le i\le m}[(\mathbf{I}_T - \mathbf{D})\mathbf{z}_i] = {} & col_{1\le i\le m}[(\mathbf{I}_T - \mathbf{D})\mathbf{H}_i]\boldsymbol{\beta} + diag_{1\le i\le m}[(\mathbf{I}_T - \mathbf{D})\mathbf{f}]\mathbf{v} + \\ & + col_{1\le i\le m}[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{u}_{2i}] + col_{1\le i\le m}[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\boldsymbol{\varepsilon}_i], \end{aligned}$$

e notando que $(\mathbf{I}_T - \mathbf{D})\mathbf{f} = \mathbf{I}_T\mathbf{f} - \mathbf{D}\mathbf{f} = \mathbf{f} - (\mathbf{f}\mathbf{f}'\mathbf{f})/c = (\mathbf{f}\mathbf{f}'\mathbf{f} - \mathbf{f}\mathbf{f}'\mathbf{f})/\mathbf{f}'\mathbf{f} = \mathbf{0}_{T\times 1}$, então o modelo transformado reduz-se a:

$$\mathbf{z}^{(1)} = \mathbf{H}^{(1)}\boldsymbol{\beta} + \mathbf{e}^{(1)}, \tag{5.5.2}$$

onde $\mathbf{z}^{(1)} = col_{1\le i\le m}\left(\mathbf{z}_i^{(1)}\right)$, $\mathbf{H}^{(1)} = col_{1\le i\le m}\left(\mathbf{H}_i^{(1)}\right)$, $\mathbf{H}_i^{(1)} = (\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{X}_i$ e $\mathbf{e}^{(1)} = col_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)\right]$ e satisfazendo $E\left(\mathbf{e}^{(1)}\right) = \mathbf{0}_{mT\times 1}$. Uma vez que no modelo transformado (5.5.2) se verifica que a matriz de covariâncias do termo de erro,

$$\begin{aligned}
V\left(\mathbf{e}^{(1)}\right) &= E\left(\left\{col_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)\right]\right\}\left\{col_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)\right]\right\}'\right) \\
&= \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right].E\left\{diag_{1\le i\le m}\left[\mathbf{P}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)'\mathbf{P}'\right]\right\}\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right]' \\
&= \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right].\left\{diag_{1\le i\le m}\left[\mathbf{P}E(\mathbf{u}_{2i}\mathbf{u}_{2i}')\mathbf{P}' + \mathbf{P}E(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i')\mathbf{P}' + \mathbf{P}E(\mathbf{u}_{2i}\boldsymbol{\varepsilon}_i')\mathbf{P}' + \mathbf{P}E(\boldsymbol{\varepsilon}_i\mathbf{u}_{2i}')\mathbf{P}'\right]\right\}\times \\
&\quad \times\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right] \\
&= \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right].\left\{diag_{1\le i\le m}\left[\mathbf{P}\sigma^2\boldsymbol{\Gamma}\mathbf{P}' + \mathbf{P}\mathbf{R}_i\mathbf{P}' + \mathbf{P}\mathbf{0}_{m\times m}\mathbf{P}' + \mathbf{P}\mathbf{0}_{m\times m}\mathbf{P}'\right]\right\}\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right] \\
&= \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right].\left\{diag_{1\le i\le m}\left[\sigma^2\mathbf{P}\mathbf{P}^{-1}\left(\mathbf{P}^{-1}\right)'\mathbf{P}' + \mathbf{P}\mathbf{R}_i\mathbf{P}'\right]\right\}\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right] \\
&= \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right].\left[diag_{1\le i\le m}\left(\sigma^2\mathbf{I}_T + \mathbf{P}\mathbf{R}_i\mathbf{P}'\right)\right].\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D})\right]' \\
&= diag_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\left(\sigma^2\mathbf{I}_T + \mathbf{P}\mathbf{R}_i\mathbf{P}'\right)(\mathbf{I}_T - \mathbf{D})'\right],
\end{aligned} \tag{5.5.3}$$

não depende de σ_u^2, então σ^2 pode ser estimado através da soma dos quadrados dos resíduos obtidos pela estimação do modelo reduzido (5.5.2). Seja então $\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}$ a soma dos quadrados dos resíduos obtidos na regressão de $\mathbf{z}^{(1)}$ sobre $\mathbf{H}^{(1)}$ pelo método dos mínimos quadrados ordinários. A partir destes resíduos, obtém-se o seguinte estimador não enviesado e consistente de σ^2, dado por:

$$\begin{aligned}
\tilde{\sigma}^2 = &\left(\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} - tr\left\{\left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D}) - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}\right]\left[diag_{1\le i\le m}(\mathbf{P}\mathbf{R}_i\mathbf{P})\right]\right\}\right)\times \\
&\times\left[m(T-1) - r\left(\mathbf{H}^{(1)}\right)\right]^{-1}.
\end{aligned} \tag{5.5.4}$$

Este estimador é exactamente igual ao estimador proposto por Rao e Yu (1994), no âmbito do modelo temporal de Rao-Yu, para a componente de variância associada aos efeitos aleatórios específicos de domínio-tempo, apesar de agora se estar a trabalhar com um modelo espaciotemporal. Isto deve-se ao facto de se ter transformado o modelo (5.2.5) de forma a eliminar o vector de efeitos aleatórios $\mathbf{v}$, os quais envolvem a

associação espacial. A demonstração do não enviesamento e da consistência assimptótica do estimador (5.5.4) foi efectuada por Yu na sua tese de doutoramento (Yu, 1993, p. 117).

Passa-se agora para a estimação de σ_u^2. Em primeiro lugar transforma-se $\mathbf{z}_i$ em $z_i^{(2)} = c^{-1/2}\mathbf{f}'\mathbf{z}_i$ de forma a que $u_{2i}^{(2)} = c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{u}_{2i}$ tenha média nula e variância σ^2. Pré-multiplicando o modelo (5.5.1) por $diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)$, obtém-se o seguinte modelo transformado:

$$\begin{aligned} diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{z}_i\right) &= diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{H}_i\right)\boldsymbol{\beta} + diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)diag_{1\le i\le m}\left(\mathbf{f}\right)\mathbf{v} \\ &\quad + diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\mathbf{u}_{2i}\right) + diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\boldsymbol{\varepsilon}_i\right) \end{aligned}$$

$$\begin{aligned} col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{z}_i\right) &= col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{H}_i\right)\boldsymbol{\beta} + diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{f}\right)\mathbf{v} + \\ &\quad + col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{u}_{2i}\right) + col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\boldsymbol{\varepsilon}_i\right), \end{aligned}$$

e notando que $c^{-1/2}\mathbf{f}'\mathbf{f} = c^{-1/2}c = c^{1/2}$, então o modelo transformado reduz-se a:

$$\mathbf{z}^{(2)} = \mathbf{H}^{(2)}\boldsymbol{\beta} + \mathbf{e}^{(2)}, \tag{5.5.5}$$

onde $\mathbf{z}^{(2)} = col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{z}_i\right)$, $\mathbf{H}^{(2)} = col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{H}_i\right)$ e $\mathbf{e}^{(2)} = c^{1/2}\mathbf{v} + \mathbf{u}_2^{(2)} + \boldsymbol{\varepsilon}^{(2)}$, no qual $\mathbf{u}_2^{(2)} = col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\mathbf{u}_{2i}\right)$ e $\boldsymbol{\varepsilon}^{(2)} = col_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\mathbf{P}\boldsymbol{\varepsilon}_i\right)$ e satisfazendo $E\left(\mathbf{e}^{(2)}\right) = \mathbf{0}_{m\times 1}$. Note-se que:

$$\begin{aligned} V\left(\mathbf{u}_2^{(2)}\right) &= E\left\{\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\mathbf{u}_{2i}\right)\right]\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\mathbf{u}_{2i}\right)\right]'\right\} \\ &= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].E\left[diag_{1\le i\le m}\left(\mathbf{P}\mathbf{u}_{2i}\mathbf{u}_{2i}'\mathbf{P}'\right)\right].\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\ &= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].\left\{diag_{1\le i\le m}\left[\mathbf{P}E\left(\mathbf{u}_{2i}\mathbf{u}_{2i}'\right)\mathbf{P}'\right]\right\}.\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\ &= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].\left[diag_{1\le i\le m}\left(\sigma^2\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}'\right)\right].\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\ &= \sigma^2\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]\left\{diag_{1\le i\le m}\left[\mathbf{P}\mathbf{P}^{-1}\left(\mathbf{P}^{-1}\right)'\mathbf{P}'\right]\right\}\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\ &= \sigma^2\left[diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{I}_T\mathbf{f}\right)\right] \\ &= \sigma^2\left[diag_{1\le i\le m}\left(c^{-1}c\right)\right] \\ &= \sigma^2\mathbf{I}_m \end{aligned}$$

e

$$
\begin{aligned}
V\left(\boldsymbol{\varepsilon}^{(2)}\right) &= E\left\{\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\boldsymbol{\varepsilon}_i\right)\right]\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)col_{1\le i\le m}\left(\mathbf{P}\boldsymbol{\varepsilon}_i\right)\right]'\right\} \\
&= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].E\left[diag_{1\le i\le m}\left(\mathbf{P}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\mathbf{P}'\right)\right].\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\
&= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].\left\{diag_{1\le i\le m}\left[\mathbf{P}E\left(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right)\mathbf{P}'\right]\right\}.\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\
&= \left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right].\left[diag_{1\le i\le m}\left(\mathbf{P}\mathbf{R}_i\mathbf{P}'\right)\right].\left[diag_{1\le i\le m}\left(c^{-1/2}\mathbf{f}'\right)\right]' \\
&= diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right).
\end{aligned}
$$

O modelo transformado (5.5.5) tem matriz de covariâncias do termo de erro dada por:

$$
\begin{aligned}
V\left(\mathbf{e}^{(2)}\right) &= V\left(c^{1/2}\mathbf{v}\right)+V\left(\mathbf{u}_2^{(2)}\right)+V\left(\boldsymbol{\varepsilon}^{(2)}\right) \\
&= c\sigma_u^2\mathbf{B}^{-1}+\sigma^2\mathbf{I}_m+diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right).
\end{aligned}
\tag{5.5.6}
$$

Seja então $\hat{\mathbf{e}}^{(2)\prime}\hat{\mathbf{e}}^{(2)}$ a soma dos quadrados dos resíduos obtidos na regressão de $\mathbf{z}^{(2)}$ sobre $\mathbf{H}^{(2)}$ pelo método dos mínimos quadrados ordinários. A partir destes resíduos, obtém-se o seguinte estimador não enviesado de σ_u^2, dado por:

$$
\tilde{\sigma}_u^2 = \frac{1}{c\times tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)}\left\{\hat{\mathbf{e}}^{(2)\prime}\hat{\mathbf{e}}^{(2)} - tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]-\tilde{\sigma}^2\left[m-r\left(\mathbf{H}^{(2)}\right)\right]\right\}, \tag{5.5.7}
$$

onde $\mathbf{P}_{H^{(2)}} = \mathbf{I}_m - \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)\prime}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)\prime}$[54] e $\tilde{\sigma}^2$ é dado por (5.5.4).

Lema 5.1: *O estimador $\tilde{\sigma}_u^2$ é um estimador não enviesado de σ_u^2.*

Demonstração: A demonstração do não enviesamento de $\tilde{\sigma}_u^2$ é baseada nas propriedades dos resíduos dos mínimos quadrados. Antes de se avançar para essa demonstração, note-se que $\mathbf{P}_{H^{(2)}}$ é uma matriz simétrica[55], idempotente[56] e ortogonal aos regressores,

[54] Decidiu-se usar a particular matriz inversa generalizada de Moore-Penrose em $\mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}$ e em $\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)\prime}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)\prime}$.

[55] Basta notar que no caso da matriz inversa generalizada de Moore-Penrose, se **B** for a matriz inversa generalizada de **A**, então $(\mathbf{B}^{-})' = (\mathbf{B}')^{-}$ (Peterson e Pederson, 2008).

[56] Basta notar que no caso da matriz inversa generalizada de Moore-Penrose, se **B** for a matriz inversa generalizada de **A**, então **BAB=B** (Peterson e Pederson, 2008).

$\mathbf{P}_{H^{(2)}}\mathbf{H}^{(2)} = \mathbf{0}_{m\times p}$ se $\mathbf{H}^{(2)'}\mathbf{H}^{(2)}$ for regular. Para além disso verifica-se que $tr\left(\mathbf{P}_{H^{(2)}}\right) = m - r\left(\mathbf{H}^{(2)}\right)$:

$$\begin{aligned} tr\left(\mathbf{P}_{H^{(2)}}\right) &= tr\left[\mathbf{I}_m - \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)'}\right] \\ &= tr\left(\mathbf{I}_m\right) - tr\left[\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)'}\right] \\ &= m - tr\left[\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\right]. \end{aligned}$$

Notando agora que: (*i*) $\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}$ é uma matriz idempotente, pelo que o seu traço é igual à sua característica; e (*ii*) $\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}$ é a matriz inversa generalizada de Moore-Penrose de $\mathbf{H}^{(2)'}\mathbf{H}^{(2)}$, pelo que $r\left[\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\right] = r\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)$, então demonstra-se que $tr\left(\mathbf{P}_{H^{(2)}}\right) = m - r\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right) = m - r\left(\mathbf{H}^{(2)}\right)$ (Peterson e Pederson, 2008). A partir da regressão pelo método dos mínimos quadrados ordinários de $\mathbf{z}^{(2)}$ sobre $\mathbf{H}^{(2)}$, sob o modelo (5.5.5), obtém-se o BLUE de $\boldsymbol{\beta}$, dado por $\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)'}\mathbf{z}^{(2)}$, e os resíduos dos mínimos quadrados dados por:

$$\begin{aligned} \hat{\mathbf{e}}^{(2)} &= \mathbf{z}^{(2)} - \mathbf{H}^{(2)}\hat{\boldsymbol{\beta}} \\ &= \mathbf{z}^{(2)} - \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)'}\mathbf{z}^{(2)} \\ &= \left[\mathbf{I}_m - \mathbf{H}^{(2)}\left(\mathbf{H}^{(2)'}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)'}\right]\mathbf{z}^{(2)} \\ &= \mathbf{P}_{H^{(2)}}\left(\mathbf{H}^{(2)}\boldsymbol{\beta} + \mathbf{e}^{(2)}\right) \\ &= \mathbf{P}_{H^{(2)}}\mathbf{H}^{(2)}\boldsymbol{\beta} + \mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)} \\ &= \mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)}. \end{aligned}$$

O valor esperado da soma dos quadrados dos resíduos é dado por:

$$
\begin{aligned}
E\left[\left(\hat{\mathbf{e}}^{(2)}\right)'\hat{\mathbf{e}}^{(2)}\right] &= E\left[\left(\mathbf{e}^{(2)}\right)'\mathbf{P}_{H^{(2)}}\mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)}\right] \\
&= E\left[\left(\mathbf{e}^{(2)}\right)'\mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)}\right] \\
&= E\left\{tr\left[\left(\mathbf{e}^{(2)}\right)'\mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)}\right]\right\} \\
&= tr\left\{E\left[\mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)}\left(\mathbf{e}^{(2)}\right)'\right]\right\} \\
&= tr\left\{\mathbf{P}_{H^{(2)}}E\left[\mathbf{e}^{(2)}\left(\mathbf{e}^{(2)}\right)'\right]\right\} \\
&= tr\left\{\mathbf{P}_{H^{(2)}}\left[c\sigma_u^2\mathbf{B}^{-1}+\sigma^2\mathbf{I}_m+diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]\right\} \\
&= c\sigma_u^2 tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)+\sigma^2 tr\left(\mathbf{P}_{H^{(2)}}\right)+tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right] \\
&= c\sigma_u^2 tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)+\sigma^2\left[m-r\left(\mathbf{H}^{(2)}\right)\right]+tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right].
\end{aligned}
$$

Com base no resultado anterior e relembrando que $E\left(\tilde{\sigma}^2\right)=\sigma^2$, verifica-se facilmente que $\tilde{\sigma}_u^2$ é um estimador não enviesado de σ_u^2:

$$
\begin{aligned}
E\left(\tilde{\sigma}_u^2\right) &= \frac{1}{c\times tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)}\left\{E\left(\hat{\mathbf{e}}^{(2)\,\prime}\hat{\mathbf{e}}^{(2)}\right)-tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]-E\left(\tilde{\sigma}^2\right)\left[m-r\left(\mathbf{H}^{(2)}\right)\right]\right\} \\
&= \frac{1}{c\times tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)}\left\{c\sigma_u^2 tr\left(\mathbf{P}_{H^{(2)}}\mathbf{B}^{-1}\right)+\sigma^2\left[m-r\left(\mathbf{H}^{(2)}\right)\right]+tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]-\right. \\
&\quad \left.-tr\left[\mathbf{P}_{H^{(2)}}diag_{1\le i\le m}\left(c^{-1}\mathbf{f}'\mathbf{P}\mathbf{R}_i\mathbf{P}'\mathbf{f}\right)\right]-\sigma^2\left[m-r\left(\mathbf{H}^{(2)}\right)\right]\right\} \\
&= \sigma_u^2. \qquad \blacksquare
\end{aligned}
$$

À semelhança do que se verifica no estimador (5.5.4), que é um estimador assimptoticamente consistente, também se admite que o estimador (5.5.7) seja consistente quando $m\to\infty$, tendo em conta os resultados apresentados na tese de doutoramento de Yu (1993, p. 117), e assumindo adequadas condições de regularidade para o modelo proposto (5.2.5).

Uma vez que os estimadores $\tilde{\sigma}^2$ e $\tilde{\sigma}_u^2$ podem assumir valores negativos, então trunca-se estes estimadores a zero sempre que eles assumirem valores negativos, obtendo-se:

$$
\hat{\sigma}^2 = \max\left\{0;\tilde{\sigma}^2\right\}, \tag{5.5.8}
$$

$$
\hat{\sigma}_u^2 = \max\left\{0;\tilde{\sigma}_u^2\right\}. \tag{5.5.9}
$$

Os estimadores truncados $\hat{\sigma}^2$ e $\hat{\sigma}_u^2$ já não são centrados, mas continuarão a ser consistentes quando $m\to\infty$.

Por último, é de salientar que o EBLUP espaciotemporal (5.4.1) é um estimador não enviesado no modelo pelo facto de $\hat{\sigma}^2$ e $\hat{\sigma}_u^2$ serem funções ímpares em **y** e invariantes a translações, ou seja, os estimadores das componentes de variância não se alteram quando **y** passa a –**y** ou a **y**-**Xh**, $\forall \mathbf{h} \in \Re^p, \forall \mathbf{y}$ (de acordo com os resultados gerais de Kackar e Harville (1984)).

5.6 APROXIMAÇÃO ANALÍTICA DO EQMP DO EBLUP ESPACIOTEMPORAL

Uma medida de variabilidade associada ao EBLUP, $\hat{\theta}_{it}$, é dada pelo seu EQMP. Tal como foi apresentado na secção 3.2.5, sob condições de normalidade dos efeitos aleatórios e dos erros da sondagem, essa medida de variabilidade pode ser decomposta em três componentes (com base nos trabalhos de Kackar e Harville (1984) e utilizando os resultados gerais de Henderson (1975)):

$$EQMP\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\boldsymbol{\psi}) + g_{2it}(\boldsymbol{\psi}) + E\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}}) - \tilde{\theta}_{it}(\boldsymbol{\psi})\right]^2, \qquad (5.6.1)$$

onde *E* representa o valor esperado relativo ao modelo (5.2.5). Utilizando os referidos resultados gerais de Henderson (1975), as duas componentes do EQMP do BLUP podem ser analiticamente avaliadas a partir de expressões "fechadas" sem que se verifique a normalidade dos erros do modelo. Basta apenas que os erros do modelo estejam simetricamente distribuídos. A primeira componente, $g_{1it}(\boldsymbol{\psi})$, que representa a incerteza devida à estimação dos efeitos aleatórios e é de ordem $o(1)$, é dada por:

$$
\begin{aligned}
g_{1it}(\boldsymbol{\psi}) &= \mathbf{m}_{it}'\left(\mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\right)\mathbf{m}_{it} \\
&= \begin{bmatrix}\mathbf{m}_{1i}' & \mathbf{m}_{2it}'\end{bmatrix}\begin{bmatrix}\sigma_u^2\mathbf{B}^{-1} & \mathbf{0}_{m\times mT}\\ \mathbf{0}_{mT\times m} & \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma}\end{bmatrix}\begin{bmatrix}\mathbf{m}_{1i}\\ \mathbf{m}_{2it}\end{bmatrix}- \\
&\quad -\begin{bmatrix}\mathbf{m}_{1i}' & \mathbf{m}_{2it}'\end{bmatrix}\begin{bmatrix}\sigma_u^2\mathbf{B}^{-1} & \mathbf{0}_{m\times mT}\\ \mathbf{0}_{mT\times m} & \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m\otimes\mathbf{1}_T'\\ \mathbf{I}_{mT}\end{bmatrix}\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it} \\
&= \mathbf{m}_{1i}'\sigma_u^2\mathbf{B}^{-1}\mathbf{m}_{1i}+\mathbf{m}_{2it}'\sigma^2\left(\mathbf{I}_m\otimes\boldsymbol{\Gamma}\right)\mathbf{m}_{2it}- \\
&\quad -\left[\mathbf{m}_{1i}'\sigma_u^2\mathbf{B}^{-1}\left(\mathbf{I}_m\otimes\mathbf{1}_T'\right)+\mathbf{m}_{2it}'\sigma^2\left(\mathbf{I}_m\otimes\boldsymbol{\Gamma}\right)\right]\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it} \\
&= \sigma_u^2\boldsymbol{\varsigma}_i'\mathbf{m}_{1i}+\frac{\sigma^2}{1-\rho^2}-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it} \\
&= \sigma_u^2\varsigma_{ii}+\frac{\sigma^2}{1-\rho^2}-\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right)'\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right),
\end{aligned}
\tag{5.6.2}
$$

onde $\boldsymbol{\varsigma}_i'=\{\varsigma_{ii'}\}$ é a *i*-ésima linha da matriz $\mathbf{B}^{-1}$ e $\boldsymbol{\zeta}_{it}'$ é um vector linha de dimensão $1\times mT$ com *m* blocos, no qual cada bloco é formado por um vector linha *T*-dimensional, sendo o *i*-ésimo bloco formado pela *t*-ésima linha da matriz $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}_t$, e os restantes blocos por vectores nulos $\mathbf{0}_{1\times T}$, *i,i'*=1, …, *m*. Por sua vez, a segunda componente, $g_{2it}(\boldsymbol{\psi})$, que mede a incerteza devida à estimação dos efeitos fixos e é de ordem $o\left(m^{-1}\right)$, é dada por:

$$
\begin{aligned}
g_{2it}(\boldsymbol{\psi}) &= \left(\mathbf{k}_{it}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{k}_{it}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it}\right) \\
&= \left(\mathbf{k}_{it}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\times \\
&\quad \times\left\{\mathbf{x}_{it}-\mathbf{X}'\mathbf{V}^{-1}\begin{bmatrix}\mathbf{I}_m\otimes\mathbf{1}_T & \mathbf{I}_{mT}\end{bmatrix}\begin{bmatrix}\sigma_u^2\mathbf{B}^{-1} & \mathbf{0}_{m\times mT}\\ \mathbf{0}_{mT\times m} & \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma}\end{bmatrix}\begin{bmatrix}\mathbf{m}_{1i}\\ \mathbf{m}_{2it}\end{bmatrix}\right\} \\
&= \left(\mathbf{k}_{it}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left\{\mathbf{x}_{it}-\mathbf{X}'\mathbf{V}^{-1}\begin{bmatrix}\sigma_u^2\mathbf{B}^{-1}\left(\mathbf{I}_m\otimes\mathbf{1}_T\right) & \sigma^2\mathbf{I}_m\otimes\boldsymbol{\Gamma}\end{bmatrix}\begin{bmatrix}\mathbf{m}_{1i}\\ \mathbf{m}_{2it}\end{bmatrix}\right\} \\
&= \left(\mathbf{k}_{it}-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{m}_{it}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left\{\mathbf{x}_{it}-\mathbf{X}'\mathbf{V}^{-1}\left[\sigma_u^2\mathbf{B}^{-1}\left(\mathbf{I}_m\otimes\mathbf{1}_T\right)\mathbf{m}_{1i}+\sigma^2\left(\mathbf{I}_m\otimes\boldsymbol{\Gamma}\right)\mathbf{m}_{2it}\right]\right\} \\
&= \left[\mathbf{x}_{it}-\mathbf{X}'\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right)\right]'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{x}_{it}-\mathbf{X}'\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right)\right].
\end{aligned}
\tag{5.6.3}
$$

Como já foi referido no contexto do modelo linear misto geral, a última parcela do membro direito da expressão (5.6.1) não pode ser analiticamente avaliada de forma exacta, sendo por este motivo necessária uma aproximação. Uma vez que o modelo proposto é um caso particular do modelo linear misto longitudinal Gaussiano e envolve a estimação de componentes de variância pelo método ANOVA, então propõe utilizar-

se uma aproximação baseada em desenvolvimentos em série de Taylor proposta por Prasad e Rao (1990) para o caso geral:

$$E\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})-\tilde{\theta}_{it}(\boldsymbol{\psi})\right]^2 \approx tr\left[\mathbf{L}_{it}(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}_{it}'(\boldsymbol{\psi})\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})\right]=g_{3it}(\boldsymbol{\psi}), \tag{5.6.4}$$

onde $\mathbf{L}_{it}(\boldsymbol{\psi})=\dfrac{\partial \mathbf{b}_{it}'(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$, $\mathbf{b}_{it}'(\boldsymbol{\psi})=\mathbf{m}_{it}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$, $\mathbf{V}(\boldsymbol{\psi})$ é dada por (5.2.6) e $\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})$ é a matriz de covariâncias assimptóticas de $\hat{\boldsymbol{\psi}}$. Apesar da aproximação (5.6.4) parecer relativamente simples, na verdade a obtenção de uma expressão para $g_{3it}(\boldsymbol{\psi})$ no âmbito do modelo proposto envolve alguns desafios, nomeadamente na obtenção de $\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})$. Tem-se então que $\mathbf{b}_{it}'(\boldsymbol{\psi})=\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}$, pelo que $\mathbf{L}_{it}(\boldsymbol{\psi})=\left(\dfrac{\partial \mathbf{b}_{it}}{\partial\sigma^2},\dfrac{\partial \mathbf{b}_{it}}{\partial\sigma_u^2}\right)'$ é uma matriz por blocos de ordem $2\times mT$, onde:

$$\begin{aligned}\frac{\partial \mathbf{b}_{it}'}{\partial\sigma^2}&=\boldsymbol{\zeta}_{it}'\mathbf{V}^{-1}-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2}\mathbf{V}^{-1}\\&=\boldsymbol{\zeta}_{it}'\mathbf{V}^{-1}-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\mathbf{V}^{-1}\\&=\left[\boldsymbol{\zeta}_{it}'-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\right]\mathbf{V}^{-1}\end{aligned} \tag{5.6.5}$$

e

$$\begin{aligned}\frac{\partial \mathbf{b}_{it}'}{\partial\sigma_u^2}&=\left(\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'\right)\mathbf{V}^{-1}-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_u^2}\mathbf{V}^{-1}\\&=\left(\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'\right)\mathbf{V}^{-1}-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\mathbf{V}^{-1}\\&=\left[\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\right]\mathbf{V}^{-1}\end{aligned} \tag{5.6.6}$$

Se se definir $\mathbf{A}_{it}(\boldsymbol{\psi})=\mathbf{L}_{it}(\boldsymbol{\psi})\mathbf{V}(\boldsymbol{\psi})\mathbf{L}_{it}'(\boldsymbol{\psi})=\{a_{kl}\}$, então esta é uma matriz 2×2 simétrica com elementos dados por:

$$\begin{aligned}a_{11}&=\left[\boldsymbol{\zeta}_{it}'-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\right]\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\left[\boldsymbol{\zeta}_{it}'-\left(\sigma_u^2\boldsymbol{\varsigma}_i'\otimes\mathbf{1}_T'+\sigma^2\boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\right]'\\&=\left[\boldsymbol{\zeta}_{it}-(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right)\right]'\mathbf{V}^{-1}\left[\boldsymbol{\zeta}_{it}-(\mathbf{I}_m\otimes\boldsymbol{\Gamma})\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i\otimes\mathbf{1}_T+\sigma^2\boldsymbol{\zeta}_{it}\right)\right]\end{aligned} \tag{5.6.7}$$

$$
\begin{aligned}
a_{22} &= \left[\boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' - \left(\sigma_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \sigma^2 \boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\right]\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1} \\
&\quad \times \left[\boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' - \left(\sigma_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \sigma^2 \boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\right]' \\
&= \left[\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T - \left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\mathbf{V}^{-1}\left(\sigma_u^2 \boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2 \boldsymbol{\zeta}_{it}\right)\right]'\mathbf{V}^{-1} \\
&\quad \times \left[\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T - \left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\mathbf{V}^{-1}\left(\sigma_u^2 \boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2 \boldsymbol{\zeta}_{it}\right)\right]
\end{aligned}
\tag{5.6.8}
$$

e

$$
\begin{aligned}
a_{12} = a_{21} &= \left[\boldsymbol{\zeta}_{it}' - \left(\sigma_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \sigma^2 \boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{I}_m \otimes \boldsymbol{\Gamma}\right)\right]\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1} \\
&\quad \times \left[\boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' - \left(\sigma_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \sigma^2 \boldsymbol{\zeta}_{it}'\right)\mathbf{V}^{-1}\left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\right]' \\
&= \left[\boldsymbol{\zeta}_{it} - \left(\mathbf{I}_m \otimes \boldsymbol{\Gamma}\right)\mathbf{V}^{-1}\left(\sigma_u^2 \boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2 \boldsymbol{\zeta}_{it}\right)\right]'\mathbf{V}^{-1} \\
&\quad \times \left[\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T - \left(\mathbf{Z}_1\mathbf{B}^{-1}\mathbf{Z}_1'\right)\mathbf{V}^{-1}\left(\sigma_u^2 \boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2 \boldsymbol{\zeta}_{it}\right)\right].
\end{aligned}
\tag{5.6.9}
$$

A avaliação dos elementos da matriz de covariâncias assimptótica de $\hat{\boldsymbol{\psi}}$ de dimensão 2×2, formada pelos estimadores centrados das componentes de variância, $\tilde{\sigma}^2$ e $\tilde{\sigma}_u^2$, tem que ser efectuada com o apoio de um lema sobre a avaliação da covariância entre duas formas quadráticas de variáveis normalmente distribuídas (Jiang, 2007, p. 238):

Lema: Se $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, então $Cov(\mathbf{y}'\mathbf{N}_1\mathbf{y}; \mathbf{y}'\mathbf{N}_2\mathbf{y}) = 2tr(\mathbf{N}_1\boldsymbol{\Omega}\mathbf{N}_2\boldsymbol{\Omega}) + 4\boldsymbol{\mu}'\mathbf{N}_1\boldsymbol{\Omega}\mathbf{N}_2\boldsymbol{\mu}$, onde $\mathbf{N}_1$ e $\mathbf{N}_2$ são duas matrizes simétricas.

Para que este lema seja passível de ser aplicado aos estimadores $\tilde{\sigma}^2$ e $\tilde{\sigma}_u^2$, é necessário que eles sejam apresentados como formas quadráticas de variáveis normalmente distribuídas.

Lema 5.3: *Os estimadores $\tilde{\sigma}^2$ e $\tilde{\sigma}_u^2$ podem ser apresentados, respectivamente, através das seguintes formas quadráticas*

$$\tilde{\sigma}^2 = k_1\mathbf{a}'\mathbf{C}_1\mathbf{a} + k_2, \tag{5.6.10}$$

$$\tilde{\sigma}_u^2 = k_3\mathbf{a}'\mathbf{C}_2\mathbf{a} + k_4\mathbf{a}'\mathbf{C}_1\mathbf{a} + k_5, \tag{5.6.11}$$

onde $\mathbf{a} = \mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon} \sim N(\mathbf{0}; \mathbf{V})$, $\mathbf{C}_1$ e $\mathbf{C}_2$ são matrizes simétricas e k_1, k_2, k_3, k_4 e k_5 são constantes.

Demonstração: Comece por notar-se que os erros do modelo (5.2.5) representam uma variável normalmente distribuída, $\mathbf{a} = \mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon} \sim N(\mathbf{0}; \mathbf{V})$, onde $\mathbf{V}$ é dada por (5.2.6). Note-se, também, que as partes aleatórias dos estimadores $\tilde{\sigma}^2$ e $\tilde{\sigma}_u^2$ se encontram nas componentes dadas pela soma dos quadrados dos resíduos obtidos na regressão, sendo essas componentes apresentadas como formas quadráticas. As restantes componentes desses estimadores são constantes, pelo que podem ser menosprezadas nesta demonstração.

Em primeiro lugar vai reescrever-se o estimador $\tilde{\sigma}^2$ como uma forma quadrática de variável aleatória normalmente distribuída. Tem-se a partir de (5.5.4) que:

$$\begin{aligned}\tilde{\sigma}^2 &= \left(\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} - tr\left\{ \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D}) - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime} \right] \left[diag_{1\le i\le m}(\mathbf{P}\mathbf{R}_i\mathbf{P}) \right] \right\} \right) \\ &\quad \times \left[m(T-1) - r\left(\mathbf{H}^{(1)}\right) \right]^{-1} \\ &= k_1 \hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} + k_2 ,\end{aligned}$$

onde $k_1 = \left[m(T-1) - r\left(\mathbf{H}^{(1)}\right) \right]^{-1}$ e

$$k_2 = -tr\left\{ \left[diag_{1\le i\le m}(\mathbf{I}_T - \mathbf{D}) - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime} \right] \left[diag_{1\le i\le m}(\mathbf{P}\mathbf{R}_i\mathbf{P}) \right] \right\} \left[m(T-1) - r\left(\mathbf{H}^{(1)}\right) \right]^{-1}$$

são constantes. A soma dos quadrados dos resíduos, $\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}$, pode ser apresentada como:

$$\begin{aligned}\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} &= \hat{\mathbf{e}}^{(1)\prime}\left(\mathbf{z}^{(1)} - \mathbf{H}^{(1)}\hat{\boldsymbol{\beta}}\right) \\ &= \hat{\mathbf{e}}^{(1)\prime}\left[\mathbf{z}^{(1)} - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}\mathbf{z}^{(1)} \right] \\ &= \hat{\mathbf{e}}^{(1)\prime}\left\{ \left[\mathbf{I}_{mT} - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime} \right]\mathbf{z}^{(1)} \right\} \\ &= \hat{\mathbf{e}}^{(1)\prime}\left[\mathbf{P}_{H^{(1)}}\left(\mathbf{H}^{(1)}\boldsymbol{\beta} + \mathbf{e}^{(1)}\right) \right] \\ &= \hat{\mathbf{e}}^{(1)\prime}\left[\mathbf{P}_{H^{(1)}}\mathbf{H}^{(1)}\boldsymbol{\beta} + \mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)} \right] \\ &= \hat{\mathbf{e}}^{(1)\prime}\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\end{aligned}$$

uma vez que $\mathbf{P}_{H^{(1)}} = \mathbf{I}_{mT} - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}$ é uma matriz simétrica, idempotente e ortogonal aos regressores, $\mathbf{P}_{H^{(1)}}\mathbf{H}^{(1)} = \mathbf{0}_{mT\times p}$ se $\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}$ for regular. Notando-se agora que $\mathbf{e}^{(1)} = col_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i)\right]$ e $\mathbf{H}^{(1)} = col_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{X}_i\right]$, e definindo-se $\mathbf{C}^{(1)} = diag_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\right]$, tem-se que $\mathbf{e}^{(1)} = diag_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\right]col_{1\le i\le m}(\mathbf{u}_{2i} + \boldsymbol{\varepsilon}_i) = \mathbf{C}^{(1)}(\mathbf{u}_2 + \boldsymbol{\varepsilon})$ e $\mathbf{H}^{(1)} = diag_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\right]col_{1\le i\le m}(\mathbf{X}_i) = \mathbf{C}^{(1)}\mathbf{X}$, pelo que:

$$\begin{aligned}\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} &= \mathbf{e}^{(1)\prime}\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\\ &= (\mathbf{u}_2 + \boldsymbol{\varepsilon})'\mathbf{C}^{(1)\prime}\left[\mathbf{I}_{mT} - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}\right]\mathbf{C}^{(1)}(\mathbf{u}_2 + \boldsymbol{\varepsilon})\\ &= (\mathbf{u}_2 + \boldsymbol{\varepsilon})'\mathbf{C}^{(1)\prime}\left[\mathbf{I}_{mT} - \mathbf{C}^{(1)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(1)\prime}\mathbf{C}^{(1)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(1)\prime}\right]\mathbf{C}^{(1)}(\mathbf{u}_2 + \boldsymbol{\varepsilon}).\end{aligned}$$

Relembrando-se ainda que $(\mathbf{I}_T - \mathbf{D})\mathbf{f} = (\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{1}_T = \mathbf{0}_{T\times 1}$, tem-se que $\mathbf{C}^{(1)}(\mathbf{u}_2 + \boldsymbol{\varepsilon}) = \mathbf{C}^{(1)}(\mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon})$ porque $\mathbf{C}^{(1)}\mathbf{Z}_1\mathbf{v} = diag_{1\le i\le m}\left[(\mathbf{I}_T - \mathbf{D})\mathbf{P}\mathbf{1}_T\right]\mathbf{v} = \mathbf{0}_{mT\times 1}$, pelo que $\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}$ pode ser apresentada como a seguinte forma quadrática de variável aleatória normalmente distribuída:

$$\begin{aligned}\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)} &= (\mathbf{u}_2 + \boldsymbol{\varepsilon})'\mathbf{C}^{(1)\prime}\left[\mathbf{I}_{mT} - \mathbf{C}^{(1)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(1)\prime}\mathbf{C}^{(1)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(1)\prime}\right]\mathbf{C}^{(1)}(\mathbf{u}_2 + \boldsymbol{\varepsilon})\\ &= (\mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon})'\mathbf{C}^{(1)\prime}\left[\mathbf{I}_{mT} - \mathbf{C}^{(1)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(1)\prime}\mathbf{C}^{(1)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(1)\prime}\right]\mathbf{C}^{(1)}(\mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon})\\ &= \mathbf{a}'\mathbf{C}_1\mathbf{a},\end{aligned}$$

onde $\mathbf{a} = \mathbf{Z}_1\mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon} \sim N(\mathbf{0};\mathbf{V})$ e $\mathbf{C}_1 = \mathbf{C}^{(1)\prime}\left[\mathbf{I}_{mT} - \mathbf{C}^{(1)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(1)\prime}\mathbf{C}^{(1)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(1)\prime}\right]\mathbf{C}^{(1)}$ é uma matriz simétrica, porque $\mathbf{P}_{H^{(1)}} = \mathbf{I}_{mT} - \mathbf{C}^{(1)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(1)\prime}\mathbf{C}^{(1)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(1)\prime}$ é uma matriz simétrica. Mostrou-se, portanto, que o estimador $\tilde{\sigma}^2$ pode ser apresentado como uma forma quadrática de variável aleatória normalmente distribuída:

$$\tilde{\sigma}^2 = k_1 \mathbf{a}' \mathbf{C}_1 \mathbf{a} + k_2 .$$

Passa-se agora para a demonstração de como o estimador $\tilde{\sigma}_u^2$ pode ser reescrito como uma forma quadrática de variável aleatória normalmente distribuída. Tem-se a partir de (5.5.7) que:

$$\begin{aligned}
\tilde{\sigma}_u^2 &= \frac{1}{c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)} \left\{ \hat{\mathbf{e}}^{(2)\prime} \hat{\mathbf{e}}^{(2)} - tr\left[\mathbf{P}_{H^{(2)}} diag_{1 \le i \le m}\left(c^{-1} \mathbf{f}' \mathbf{P} \mathbf{R}_i \mathbf{P}' \mathbf{f}\right)\right] - \tilde{\sigma}^2 \left[m - r\left(\mathbf{H}^{(2)}\right)\right] \right\} \\
&= \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1} \hat{\mathbf{e}}^{(2)\prime} \hat{\mathbf{e}}^{(2)} - tr\left[\mathbf{P}_{H^{(2)}} diag_{1 \le i \le m}\left(c^{-1} \mathbf{f}' \mathbf{P} \mathbf{R}_i \mathbf{P}' \mathbf{f}\right)\right] \times \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1} - \\
&\quad - \left(k_1 \mathbf{a}' \mathbf{C}_1 \mathbf{a} + k_2\right) \left[m - r\left(\mathbf{H}^{(2)}\right)\right] \times \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1} \\
&= k_3 \hat{\mathbf{e}}^{(2)\prime} \hat{\mathbf{e}}^{(2)} + k_4 \mathbf{a}' \mathbf{C}_1 \mathbf{a} + k_5 ,
\end{aligned}$$

onde $k_3 = \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1}$, $k_4 = -k_1 \left[m - r\left(\mathbf{H}^{(2)}\right)\right] \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1}$ e $k_5 = \left\{ -k_2 \left[m - r\left(\mathbf{H}^{(2)}\right)\right] - tr\left[\mathbf{P}_{H^{(2)}} diag_{1 \le i \le m}\left(c^{-1} \mathbf{f}' \mathbf{P} \mathbf{R}_i \mathbf{P}' \mathbf{f}\right)\right] \right\} \left[c \times tr\left(\mathbf{P}_{H^{(2)}} \mathbf{B}^{-1}\right)\right]^{-1}$ são constantes.

Notando-se agora que $\left(\hat{\mathbf{e}}^{(2)}\right)' \hat{\mathbf{e}}^{(2)} = \left(\mathbf{e}^{(2)}\right)' \mathbf{P}_{H^{(2)}} \mathbf{e}^{(2)}$, $\mathbf{H}^{(2)} = col_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P} \mathbf{X}_i\right)$ e $\mathbf{e}^{(2)} = diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{f}\right) \mathbf{v} + col_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P} \mathbf{u}_{2i}\right) + col_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P} \boldsymbol{\varepsilon}_i\right)$, e definindo-se $\mathbf{C}^{(2)} = diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P}\right)$, tem-se que:

$$\mathbf{H}^{(2)} = diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P}\right) col_{1 \le i \le m}\left(\mathbf{X}_i\right) = \mathbf{C}^{(2)} \mathbf{X}$$

e

$$\begin{aligned}
\mathbf{e}^{(2)} &= diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P}\right) diag_{1 \le i \le m}\left(\mathbf{1}_T\right) \mathbf{v} + diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P}\right) col_{1 \le i \le m}\left(\mathbf{u}_{2i}\right) + \\
&\quad + diag_{1 \le i \le m}\left(c^{-1/2} \mathbf{f}' \mathbf{P}\right) col_{1 \le i \le m}\left(\boldsymbol{\varepsilon}_i\right) \\
&= \mathbf{C}^{(2)} \mathbf{Z}_1 \mathbf{v} + \mathbf{C}^{(2)} \mathbf{u}_2 + \mathbf{C}^{(2)} \boldsymbol{\varepsilon} \\
&= \mathbf{C}^{(2)} \left(\mathbf{Z}_1 \mathbf{v} + \mathbf{u}_2 + \boldsymbol{\varepsilon}\right).
\end{aligned}$$

Pode, finalmente, mostrar-se que $\hat{\mathbf{e}}^{(2)\prime} \hat{\mathbf{e}}^{(2)}$ pode ser apresentada como uma forma quadrática de variável aleatória normalmente distribuída:

$$
\begin{aligned}
\hat{\mathbf{e}}^{(2)\prime}\hat{\mathbf{e}}^{(2)} &= \mathbf{e}^{(2)\prime}\mathbf{P}_{H^{(2)}}\mathbf{e}^{(2)} \\
&= \left(\mathbf{Z}_1\mathbf{v}+\mathbf{u}_2+\boldsymbol{\varepsilon}\right)'\mathbf{C}^{(2)\prime}\left[\mathbf{I}_m-\mathbf{H}^{(2)}\left(\mathbf{H}^{(2)\prime}\mathbf{H}^{(2)}\right)^{-}\mathbf{H}^{(2)\prime}\right]\mathbf{C}^{(2)}\left(\mathbf{Z}_1\mathbf{v}+\mathbf{u}_2+\boldsymbol{\varepsilon}\right) \\
&= \left(\mathbf{Z}_1\mathbf{v}+\mathbf{u}_2+\boldsymbol{\varepsilon}\right)'\mathbf{C}^{(2)\prime}\left[\mathbf{I}_m-\mathbf{C}^{(2)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(2)\prime}\mathbf{C}^{(2)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(2)\prime}\right]\mathbf{C}^{(2)}\left(\mathbf{Z}_1\mathbf{v}+\mathbf{u}_2+\boldsymbol{\varepsilon}\right) \\
&= \mathbf{a}'\mathbf{C}_2\mathbf{a}\,,
\end{aligned}
$$

onde $\mathbf{a}=\mathbf{Z}_1\mathbf{v}+\mathbf{u}_2+\boldsymbol{\varepsilon}\sim N(\mathbf{0};\mathbf{V})$ e $\mathbf{C}_2=\mathbf{C}^{(2)\prime}\left[\mathbf{I}_m-\mathbf{C}^{(2)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(2)\prime}\mathbf{C}^{(2)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(2)\prime}\right]\mathbf{C}^{(2)}$ é uma matriz simétrica, porque $\mathbf{P}_{H^{(2)}}=\mathbf{I}_m-\mathbf{C}^{(2)}\mathbf{X}\left(\mathbf{X}'\mathbf{C}^{(2)\prime}\mathbf{C}^{(2)}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}^{(2)\prime}$ é uma matriz simétrica. Mostrou-se, portanto, que o estimador $\tilde{\sigma}_u^2$ pode ser apresentado como uma forma quadrática de variável aleatória normalmente distribuída:

$$
\tilde{\sigma}_u^2 = k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_5\,. \qquad \blacksquare
$$

Voltando à expressão (5.6.4), se se definir $\mathbf{D}=\overline{\mathbf{V}}(\hat{\boldsymbol{\psi}})=\{d_{kl}\}$, então esta é uma matriz 2×2 simétrica com elementos dados por:

$$
\begin{aligned}
d_{11}=V(\tilde{\sigma}^2) &= Cov\left(k_1\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_2;k_1\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_2\right)= \\
&= Cov\left(k_1\mathbf{a}'\mathbf{C}_1\mathbf{a};k_1\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= k_1^2Cov\left(\mathbf{a}'\mathbf{C}_1\mathbf{a};\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= 2k_1^2tr\left(\mathbf{C}_1\mathbf{V}\mathbf{C}_1\mathbf{V}\right),
\end{aligned}
\qquad (5.6.12)
$$

$$
\begin{aligned}
d_{22}=V(\tilde{\sigma}_u^2) &= Cov\left(k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_5;k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_5\right)= \\
&= Cov\left(k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a};k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= k_3^2Cov\left(\mathbf{a}'\mathbf{C}_2\mathbf{a};\mathbf{a}'\mathbf{C}_2\mathbf{a}\right)+2k_3k_4Cov\left(\mathbf{a}'\mathbf{C}_2\mathbf{a};\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)+k_4^2Cov\left(\mathbf{a}'\mathbf{C}_1\mathbf{a};\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= 2k_3^2tr\left(\mathbf{C}_2\mathbf{V}\mathbf{C}_2\mathbf{V}\right)+4k_3k_4tr\left(\mathbf{C}_1\mathbf{V}\mathbf{C}_2\mathbf{V}\right)+2k_4^2tr\left(\mathbf{C}_1\mathbf{V}\mathbf{C}_1\mathbf{V}\right)
\end{aligned}
\qquad (5.6.13)
$$

e

$$
\begin{aligned}
d_{12}=d_{21}=Cov(\tilde{\sigma}^2;\tilde{\sigma}_u^2) &= Cov\left(k_1\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_2;k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}+k_5\right)= \\
&= Cov\left(k_1\mathbf{a}'\mathbf{C}_1\mathbf{a};k_3\mathbf{a}'\mathbf{C}_2\mathbf{a}+k_4\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= k_1k_3Cov\left(\mathbf{a}'\mathbf{C}_1\mathbf{a};\mathbf{a}'\mathbf{C}_2\mathbf{a}\right)+k_1k_4Cov\left(\mathbf{a}'\mathbf{C}_1\mathbf{a};\mathbf{a}'\mathbf{C}_1\mathbf{a}\right)= \\
&= 2k_1k_3tr\left(\mathbf{C}_1\mathbf{V}\mathbf{C}_2\mathbf{V}\right)+2k_1k_4tr\left(\mathbf{C}_1\mathbf{V}\mathbf{C}_1\mathbf{V}\right).
\end{aligned}
\qquad (5.6.14)
$$

Portanto, combinando as expressões (5.6.2)-(5.6.4), obtém-se uma aproximação analítica de segunda ordem para o EQMP do EBLUP espaciotemporal, dada por:

$$EQMP\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] \approx g_{1it}(\boldsymbol{\psi}) + g_{2it}(\boldsymbol{\psi}) + g_{3it}(\boldsymbol{\psi}). \qquad (5.6.15)$$

De acordo com os resultados de Prasad e Rao (1990), os termos desprezados na aproximação (5.6.15) são de ordem $o\left(m^{-1}\right)$, para $m \to \infty$, desde que se verifiquem as seguintes condições gerais de regularidade: (*i*) os elementos de $\mathbf{X}$ e $\mathbf{Z}$ são uniformemente limitados, tal que $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \left[O(m)\right]_{p\times p}$; (*ii*) m e T são finitos; (*iii*) os elementos de $\mathbf{G}$ e $\mathbf{R}$ são uniformemente limitados e diferenciáveis em relação a $\boldsymbol{\psi}$; e (*iv*) $\hat{\psi}_f = \mathbf{y}'\mathbf{C}_f\mathbf{y}$ é um estimador de ψ_f centrado e invariante a translações, onde $\mathbf{C}_f$ tem a forma $\mathbf{C}_f = diag_{1\le i\le m}\left[O\left(m^{-1}\right)\right]_{T\times T} + \left[O\left(m^{-2}\right)\right]_{mT\times mT}$, f=1, 2.

Tendo em conta os resultados apresentados em Prasad e Rao (1990), tem-se que $E\left[g_{1it}(\hat{\boldsymbol{\psi}})\right] \approx g_{1it}(\boldsymbol{\psi}) - g_{3it}(\boldsymbol{\psi})$, $E\left[g_{2it}(\hat{\boldsymbol{\psi}})\right] \approx g_{2it}(\boldsymbol{\psi})$ e $E\left[g_{3it}(\hat{\boldsymbol{\psi}})\right] \approx g_{3it}(\boldsymbol{\psi})$, sendo os termos desprezados nestas aproximações também de ordem $o\left(m^{-1}\right)$. Propõe-se, então, o seguinte estimador analítico do EQMP do EBLUP espaciotemporal, com correcção de enviesamento até à ordem $o\left(m^{-1}\right)$:

$$eqmp^{A}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) + 2g_{3it}(\hat{\boldsymbol{\psi}}). \qquad (5.6.16)$$

5.7 APROXIMAÇÃO *BOOTSTRAP* DO EQMP DO EBLUP ESPACIOTEMPORAL

Como foi referido várias vezes anteriormente, a terceira parcela da expressão (5.6.1), g_{3it}, que representa a variabilidade adicional presente no EBLUP espaciotemporal devida à estimação das componentes de variância, não pode ser calculada analiticamente de forma exacta, sendo necessário utilizar aproximações. Neste subcapítulo propõe-se uma metodologia *bootstrap* para obter essa aproximação. A metodologia aqui proposta é baseada nos trabalhos de Butar e Lahiri (2003) ao nível da

utilização de um método *bootstrap* robusto (Wu, 1986) no contexto de estimação em pequenos domínios com populações finitas. Assumindo que a estimação é assistida pelo modelo espaciotemporal de nível área (5.2.5), que está disponível um conjunto de dados iniciais provenientes de uma amostra aleatória, que as componentes de variância são estimadas pelo método dos momentos apresentado no subcapítulo 5.5, e que ϕ e ρ são conhecidos, propõe-se o seguinte procedimento *bootstrap*:

1. Calcular as estimativas das componentes de variância pelo método dos momentos, $\hat{\sigma}_u^2$ e $\hat{\sigma}^2$, com base nos dados iniciais, $\mathbf{y}$, e ajustar o modelo (5.2.5), de forma a determinar as estimativas dos efeitos fixos $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}; \hat{\boldsymbol{\psi}})$, com $\hat{\boldsymbol{\psi}} = \left(\hat{\sigma}_u^2, \hat{\sigma}^2\right)'$.

2. Calcular estimativas EBLUP espaciotemporais de θ_{it}, $\hat{\theta}_{it} = \hat{\theta}_{it}\left(\mathbf{y}; \hat{\boldsymbol{\psi}}\right)$, e os dois primeiros termos do seu EQMP, $g_{1it}\left(\hat{\boldsymbol{\psi}}\right)$ e $g_{2it}\left(\hat{\boldsymbol{\psi}}\right)$.

3. Gerar *m* cópias independentes da variável $\mathbf{u}_1^*$, com $\mathbf{u}_1^* \sim N\left(\mathbf{0}; \hat{\sigma}_u^2 \mathbf{I}_m\right)$. Com base nestes valores, gerar o vector aleatório $\mathbf{v}^* = \left(\mathbf{I}_m - \phi \mathbf{W}\right)^{-1} \mathbf{u}_1^*$, admitindo ϕ conhecido.

4. Gerar *mT* cópias independentes da variável ξ^*, com $\xi^* \sim N\left(\mathbf{0}; \hat{\sigma}^2 \mathbf{I}_{mT}\right)$, independentes de $\mathbf{u}_1^*$. Com base nestes valores, gerar o vector aleatório $\mathbf{u}_2^*$, admitindo ρ conhecido.

5. Gerar *mT* cópias independentes da variável $\boldsymbol{\varepsilon}^*$, com $\boldsymbol{\varepsilon}^* \sim N\left(\mathbf{0}; \mathbf{R}\right)$, independentes de $\mathbf{u}_1^*$ e ξ^*.

6. Construir o conjunto de dados *bootstrap* $\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}^*$, onde $\boldsymbol{\upsilon}^* = \begin{bmatrix} \mathbf{v}^{*\prime} & \mathbf{u}_2^{*\prime} \end{bmatrix}'$.

7. Calcular estimativas *bootstrap* das componentes de variância, $\hat{\sigma}_u^{2*}$ e $\hat{\sigma}^{2*}$, com base nos dados *bootstrap*, $\mathbf{y}^*$, e ajustar o modelo (5.2.5) de forma a obter as estimativas *bootstrap* dos efeitos fixos $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}\left(\mathbf{y}; \hat{\boldsymbol{\psi}}^*\right)$, com $\hat{\boldsymbol{\psi}}^* = \left(\hat{\sigma}_u^{2*}, \hat{\sigma}^{2*}\right)'$.

8. Calcular estimativas *bootstrap* do EBLUP espaciotemporal, bem como das duas primeiras componentes do seu EQMP, utilizando as estimativas *bootstrap* das componentes de variância, $\hat{\boldsymbol{\psi}}^*$:

$$\hat{\theta}_{it}^{*}=\hat{\theta}_{it}^{*}(\mathbf{y};\hat{\boldsymbol{\beta}}^{*};\hat{\boldsymbol{\psi}}^{*})=\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}^{*}+\left(\hat{\sigma}_{u}^{2*}\boldsymbol{\varsigma}_{i}'\otimes\mathbf{1}_{T}'+\hat{\sigma}^{2*}\boldsymbol{\zeta}_{it}'\right)\left[\hat{\mathbf{V}}\left(\hat{\boldsymbol{\psi}}^{*}\right)\right]^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{*}\right),$$

$$g_{1it}^{*}=g_{1it}^{*}\left(\mathbf{y};\hat{\boldsymbol{\beta}}^{*};\hat{\boldsymbol{\psi}}^{*}\right)=\hat{\sigma}_{u}^{2*}\varsigma_{ii}+\frac{\hat{\sigma}^{2*}}{1-\rho^{2}}-\left(\hat{\sigma}_{u}^{2*}\boldsymbol{\varsigma}_{i}\otimes\mathbf{1}_{T}+\hat{\sigma}^{2*}\boldsymbol{\zeta}_{it}\right)'\left(\hat{\mathbf{V}}^{*}\right)^{-1}\left(\hat{\sigma}_{u}^{2*}\boldsymbol{\varsigma}_{i}\otimes\mathbf{1}_{T}+\hat{\sigma}^{2*}\boldsymbol{\zeta}_{it}\right),$$

$$g_{2it}^{*}=g_{2it}^{*}\left(\mathbf{y};\hat{\boldsymbol{\beta}}^{*};\hat{\boldsymbol{\psi}}^{*}\right)=\left[\mathbf{x}_{it}-\mathbf{X}'\left(\hat{\mathbf{V}}^{*}\right)^{-1}\left(\hat{\sigma}_{u}^{2*}\boldsymbol{\varsigma}_{i}\otimes\mathbf{1}_{T}+\hat{\sigma}^{2*}\boldsymbol{\zeta}_{it}\right)\right]'\left[\mathbf{X}'\left(\hat{\mathbf{V}}^{*}\right)^{-1}\mathbf{X}\right]^{-1}$$
$$\times\left[\mathbf{x}_{it}-\mathbf{X}'\left(\hat{\mathbf{V}}^{*}\right)^{-1}\left(\hat{\sigma}_{u}^{2*}\boldsymbol{\varsigma}_{i}\otimes\mathbf{1}_{T}+\hat{\sigma}^{2*}\boldsymbol{\zeta}_{it}\right)\right].$$

9. Repetir as etapas 3)-8) *B* vezes. Defina-se $\hat{\sigma}_{u}^{2*(b)}$ e $\hat{\sigma}^{2*(b)}$ como estimativas *bootstrap* dos parâmetros de variância obtidas na *b*-ésima réplica *bootstrap*, $\hat{\boldsymbol{\psi}}^{*(b)}=\left(\hat{\sigma}_{u}^{2*(b)},\hat{\sigma}^{2*(b)}\right)'$; e $\hat{\boldsymbol{\beta}}^{*(b)}$, $\hat{\theta}_{it}^{*(b)}$, $g_{1it}^{*(b)}$ e $g_{2it}^{*(b)}$ como estimativas *bootstrap* de $\boldsymbol{\beta}$, θ_{it}, g_{1it} e g_{2it}, respectivamente, obtidas na *b*-ésima réplica *bootstrap*, *b*=1, …, *B*.

10. Calcular uma estimativa *bootstrap* de g_{3it}, usando a seguinte aproximação de Monte Carlo: $g_{3it}^{*}=B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{it}^{*(b)}-\hat{\theta}_{it}\right)^{2}$.

Uma vez obtidas as estimativas *bootstrap* g_{3it}^{*}, e considerando que $g_{1it}(\hat{\boldsymbol{\psi}})+g_{2it}(\hat{\boldsymbol{\psi}})$ é um estimador enviesado de $g_{1it}(\boldsymbol{\psi})+g_{2it}(\boldsymbol{\psi})$ (Prasad e Rao, 1990), então propõe-se o seguinte estimador *bootstrap* do EQMP do EBLUP espaciotemporal com correcção de enviesamento:

$$eqmp^{B}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right]=2\left[g_{1it}(\hat{\boldsymbol{\psi}})+g_{2it}(\hat{\boldsymbol{\psi}})\right]-B^{-1}\sum_{b=1}^{B}\left[g_{1it}^{*(b)}+g_{2it}^{*(b)}\right]+g_{3it}^{*}. \qquad (5.7.1)$$

5.8 APROXIMAÇÃO *JACKKNIFE* DO EQMP DO EBLUP ESPACIOTEMPORAL

À semelhança do que foi feito no âmbito do modelo de Rao-Yu, nesta secção é proposta uma metodologia alternativa que permite obter uma aproximação do estimador do EQMP do EBLUP espaciotemporal através de um procedimento *jackknife* ponderado.

Tal como foi referido na secção 4.3.7, a metodologia proposta para determinar estimativas para a parcela g_{3it} é baseada no trabalho geral de Jiang *et al.* (2002) e nos desenvolvimentos em série de Taylor de Chen e Lahiri (2008). Importa aqui referir que se decidiu introduzir a metodologia *jackkife* no âmbito do modelo espaciotemporal (5.2.5) proposto, apesar de se admitir neste modelo a existência de associação (espacial) entre os pequenos domínios, não só como mais uma metodologia alternativa de estimação do EQMP, mas sobretudo como forma de testar o desempenho desta metodologia em situações de dependência espacial entre domínios (ou entre observações).

Considerando-se o modelo (5.2.5) e a estimação das componentes de variância pelo método dos momentos, e admitindo-se conhecidos os parâmetros de dependência espacial e correlação temporal, uma aproximação *jackknife* ponderada para o estimador do EQMP do EBLUP espaciotemporal é dada por:

$$\begin{aligned} eqmp^{J}\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] &= g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) - \hat{\mathbf{c}}'_{WJ,t}(\hat{\boldsymbol{\psi}})\nabla g_{1it}(\hat{\boldsymbol{\psi}}) + tr\left[\mathbf{A}_{it}(\hat{\boldsymbol{\psi}})\hat{\mathbf{v}}_{WJ,t}(\hat{\boldsymbol{\psi}})\right] + \\ &\quad + tr\left\{\mathbf{L}_{it}(\hat{\boldsymbol{\psi}})\left[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]\left[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]'\mathbf{L}'_{it}(\hat{\boldsymbol{\psi}})\hat{\mathbf{v}}_{WJ,t}(\hat{\boldsymbol{\psi}})\right\}, \end{aligned} \tag{5.8.1}$$

onde $g_{1it}(\boldsymbol{\psi})$ é dado por (5.6.2); $g_{2it}(\boldsymbol{\psi})$ é dado por (5.6.3), $\mathbf{L}_{it}(\boldsymbol{\psi})$ é uma matriz com elementos dados por (5.6.5) e (5.6.6), e $\mathbf{A}_{it}(\boldsymbol{\psi})$ é uma matriz 2×2 simétrica com elementos (5.6.7)-(5.6.9). Para além disso, $\nabla g_{1it}(\boldsymbol{\psi}) = \left(\frac{\partial g_{1it}}{\partial \sigma^2}, \frac{\partial g_{1it}}{\partial \sigma_u^2}\right)'$ é o gradiente de $g_{1it}(\boldsymbol{\psi})$ em σ^2 e σ_u^2, o qual é um vector de dimensão 2×1 com elementos:

$$\begin{aligned} \frac{\partial g_{1it}}{\partial \sigma^2} &= \frac{1}{1-\rho^2} - \boldsymbol{\zeta}'_{it}\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right) + \\ &\quad + \left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right)'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma^2}\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right) - \left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right)'\mathbf{V}^{-1}\boldsymbol{\zeta}_{it} \\ &= \frac{1}{1-\rho^2} - 2\boldsymbol{\zeta}'_{it}\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right) + \\ &\quad + \left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right)'\mathbf{V}^{-1}(\mathbf{I}_m \otimes \boldsymbol{\Gamma})\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right) \\ &= \frac{1}{1-\rho^2} + \left[\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right)'\mathbf{V}^{-1}(\mathbf{I}_m \otimes \boldsymbol{\Gamma}) - 2\boldsymbol{\zeta}'_{it}\right]\mathbf{V}^{-1}\left(\sigma_u^2\boldsymbol{\varsigma}_i \otimes \mathbf{1}_T + \sigma^2\boldsymbol{\zeta}_{it}\right) \end{aligned} \tag{5.8.2}$$

e

$$\begin{aligned}\frac{\partial g_{1it}}{\partial \sigma_u^2} &= \varsigma_{ii} - \left(\varsigma_i \otimes \mathbf{1}_T\right)' \mathbf{V}^{-1}\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right) + \\ &+ \left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right)' \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_u^2} \mathbf{V}^{-1}\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right) - \left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right)' \mathbf{V}^{-1}\left(\varsigma_i \otimes \mathbf{1}_T\right) \\ &= \varsigma_{ii} - 2\left(\varsigma_i \otimes \mathbf{1}_T\right)' \mathbf{V}^{-1}\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right) + \\ &+ \left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right)' \mathbf{V}^{-1}\left(\mathbf{Z}_1 \mathbf{B}^{-1} \mathbf{Z}_1'\right) \mathbf{V}^{-1}\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right) \\ &= \varsigma_{ii} + \left[\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right)' \mathbf{V}^{-1}\left(\mathbf{Z}_1 \mathbf{B}^{-1} \mathbf{Z}_1'\right) - 2\left(\varsigma_i \otimes \mathbf{1}_T\right)'\right] \mathbf{V}^{-1}\left(\sigma_u^2 \varsigma_i \otimes \mathbf{1}_T + \sigma^2 \zeta_{it}\right). \quad (5.8.3)\end{aligned}$$

Por último, $\hat{\mathbf{c}}_{WJ,t}\left(\hat{\boldsymbol{\psi}}\right) = \sum_{e=1}^{m} w_{et}\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)$ é um estimador *jackknife* ponderado do enviesamento de $\hat{\boldsymbol{\psi}}$, e $\hat{\mathbf{v}}_{WJ,t}\left(\hat{\boldsymbol{\psi}}\right) = \sum_{e=1}^{m} w_{et}\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)\left(\hat{\boldsymbol{\psi}}_{-e} - \hat{\boldsymbol{\psi}}\right)'$ é um estimador *jackknife* ponderado da matriz de covariâncias de $\hat{\boldsymbol{\psi}}$, ambos corrigidos até à ordem $o\left(m^{-1}\right)$, e onde $\hat{\boldsymbol{\psi}}_{-e}$ é o estimador de $\boldsymbol{\psi}$ depois de eliminar as T observações referentes ao e-ésimo pequeno domínio e w_{et} são ponderadores que devem satisfazer a seguinte condição $w_{et} = 1 + o\left(m^{-1}\right)$.

Existem diferentes possibilidades de escolha dos ponderadores, w_{et}, podendo-se, por exemplo, utilizar duas possibilidades usadas por Chen e Lahiri (2008) no contexto do modelo de Fay-Herriot: $w_{1et} = \frac{m-1}{m}$ e $w_{2et} = 1 - \mathbf{x}_{et}'\left(\sum_{i=1}^{m}\sum_{t=1}^{T} \mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1} \mathbf{x}_{et}$. Naturalmente que o uso de diferentes ponderadores resulta em diferentes estimadores *jackknife* do EQMP do EBLUP espaciotemporal.

Finalmente, note-se que apesar de se ter assumido a normalidade dos erros do modelo (5.2.5) aquando da sua especificação, esta hipótese só foi indispensável para derivar uma aproximação analítica do estimador do EQMP do EBLUP, corrigida até à segunda ordem quando $m \to \infty$. Porém, não foi necessário assumir a hipótese da normalidade dos erros na obtenção do BLUP espaciotemporal nem na derivação das componentes de variância, tal como não é fundamental admiti-la nos procedimentos de estimação do EQMP do EBLUP espaciotemporal por métodos de reamostragem. Basta apenas, em

alguns desses casos, que os erros do modelo (5.2.5) estejam simetricamente distribuídos.

5.9 CASO PARTICULAR DO MODELO

Tal como foi referido no subcapítulo 5.2, o modelo espacial de Salvati (4.5.4) pode ser considerado um caso particular do modelo espaciotemporal proposto (5.2.5), quando se considera T=1, ρ=0 e $\sigma^2 = 0$. Salvati (2004), bem como Singh *et al*. (2005), propuseram que a estimação da componente de variância do modelo espacial (4.5.4) seja efectuada através de um dos métodos de verosimilhança. Contudo, no âmbito deste trabalho pretende estimar-se as componentes de variância através de um método que não exija a verificação da hipótese da normalidade dos erros do modelo de estimação em domínios. Desta forma, neste subcapítulo propõe-se uma extensão do método III de Henderson (1953) para a estimação da componente de variância, σ_u^2, no contexto do modelo espacial (4.5.4), com erros associados espacialmente através de um processo SAR, v_i, e erros da sondagem independentes, ε_{it}. Considere-se, por facilidade de exposição, que $\mathbf{Z} = \mathbf{I}$.

À semelhança da metodologia proposta no subcapítulo 5.5, o estimador proposto para a componente de variância é baseado nos resíduos de uma regressão efectuada pelo método dos mínimos quadrados ordinários sobre o modelo (4.5.4) transformado. Assim, transforma-se $\mathbf{y}$ em $\mathbf{z}^{(1)} = (\mathbf{I}_m - \phi\mathbf{W})\mathbf{y}$ de forma a que $\mathbf{u}^{(1)} = (\mathbf{I}_m - \phi\mathbf{W})\mathbf{v}$ tenha média nula e variância $\sigma_u^2\mathbf{I}_m$. Pré-multiplicando o modelo (4.5.4) por $(\mathbf{I}_m - \phi\mathbf{W})$, obtém-se o seguinte modelo transformado:

$$\mathbf{z}^{(1)} = \mathbf{H}^{(1)}\boldsymbol{\beta} + \mathbf{e}^{(1)}, \tag{5.9.1}$$

onde $\mathbf{H}^{(1)} = (\mathbf{I}_m - \phi\mathbf{W})\mathbf{X}$ e $\mathbf{e}^{(1)} = \mathbf{u} + (\mathbf{I}_m - \phi\mathbf{W})\boldsymbol{\varepsilon}$. Note-se que $E(\mathbf{e}^{(1)}) = \mathbf{0}$ e que:

$$
\begin{aligned}
V\left(\mathbf{e}^{(1)}\right) &= E\left\{\left[\mathbf{u}+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\boldsymbol{\varepsilon}\right]\left[\mathbf{u}+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\boldsymbol{\varepsilon}\right]'\right\} \\
&= E\left[\mathbf{u}\mathbf{u}'+\mathbf{u}\boldsymbol{\varepsilon}'\left(\mathbf{I}_m-\phi\mathbf{W}\right)'+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\boldsymbol{\varepsilon}\mathbf{u}+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\left(\mathbf{I}_m-\phi\mathbf{W}\right)'\right] \\
&= E\left(\mathbf{u}\mathbf{u}'\right)+\left(\mathbf{I}_m-\phi\mathbf{W}\right)E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\left(\mathbf{I}_m-\phi\mathbf{W}\right)' \\
&= \sigma_u^2\mathbf{I}_m+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\mathbf{R}\left(\mathbf{I}_m-\phi\mathbf{W}\right)'.
\end{aligned}
$$

Faça-se agora a regressão de $\mathbf{z}^{(1)}$ sobre $\mathbf{H}^{(1)}$ pelo método dos mínimos quadrados ordinários e defina-se $\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}$ como a soma dos quadrados dos resíduos obtidos nessa regressão. A partir destes resíduos, obtém-se o seguinte estimador não enviesado de σ_u^2, dado por:

$$
\tilde{\sigma}_u^2=\frac{1}{m-r\left(\mathbf{H}^{(1)}\right)}\left\{\hat{\mathbf{e}}^{(1)\prime}\hat{\mathbf{e}}^{(1)}-tr\left[\mathbf{P}_{H^{(1)}}\left(\mathbf{I}_m-\phi\mathbf{W}\right)\mathbf{R}\left(\mathbf{I}_m-\phi\mathbf{W}\right)'\right]\right\}, \qquad (5.9.2)
$$

onde $\mathbf{P}_{H^{(1)}}=\mathbf{I}_m-\mathbf{H}^{(1)}\left(\mathbf{H}^{(1)\prime}\mathbf{H}^{(1)}\right)^{-}\mathbf{H}^{(1)\prime}$. A demonstração do não enviesamento de $\tilde{\sigma}_u^2$ é baseada nas propriedades dos resíduos dos mínimos quadrados, pelo que é muito semelhante à demonstração do não enviesamento do estimador (5.5.4). Notando-se que $tr\left(\mathbf{P}_{H^{(1)}}\right)=m-r\left(\mathbf{H}^{(1)}\right)$ e que $\hat{\mathbf{e}}^{(1)}=\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}$, então o valor esperado da soma dos quadrados dos resíduos é dada por:

$$
\begin{aligned}
E\left[\left(\hat{\mathbf{e}}^{(1)}\right)'\hat{\mathbf{e}}^{(1)}\right] &= E\left[\left(\mathbf{e}^{(1)}\right)'\mathbf{P}_{H^{(1)}}\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\right] \\
&= E\left[\left(\mathbf{e}^{(1)}\right)'\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\right] \\
&= E\left\{tr\left[\left(\mathbf{e}^{(1)}\right)'\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\right]\right\} \\
&= tr\left\{E\left[\mathbf{P}_{H^{(1)}}\mathbf{e}^{(1)}\left(\mathbf{e}^{(1)}\right)'\right]\right\} \\
&= tr\left\{\mathbf{P}_{H^{(1)}}E\left[\mathbf{e}^{(1)}\left(\mathbf{e}^{(1)}\right)'\right]\right\} \\
&= tr\left\{\mathbf{P}_{H^{(1)}}\left[\sigma_v^2\mathbf{I}_m+\left(\mathbf{I}_m-\phi\mathbf{W}\right)\mathbf{R}\left(\mathbf{I}_m-\phi\mathbf{W}\right)'\right]\right\} \\
&= tr\left(\sigma_v^2\,\mathbf{P}_{H^{(1)}}\right)+tr\left[\mathbf{P}_{H^{(1)}}\left(\mathbf{I}_m-\phi\mathbf{W}\right)\mathbf{R}\left(\mathbf{I}_m-\phi\mathbf{W}\right)'\right] \\
&= \sigma_v^2\left[m-r\left(\mathbf{H}^{(1)}\right)\right]+tr\left[\mathbf{P}_{H^{(1)}}\left(\mathbf{I}_m-\phi\mathbf{W}\right)\mathbf{R}\left(\mathbf{I}_m-\phi\mathbf{W}\right)'\right].
\end{aligned}
$$

Com base no resultado anterior, facilmente se verifica que $\tilde{\sigma}_u^2$ é um estimador não enviesado de σ_u^2. Pelo facto de $\tilde{\sigma}_u^2$ poder assumir valores negativos, então faz-se a truncagem a zero desse estimador sempre que ele assuma valores negativos:

$\hat{\sigma}_u^2 = \max\{0; \tilde{\sigma}_u^2\}$. Tal como foi referido anteriormente, o estimador truncado $\hat{\sigma}_u^2$ já não é centrado, mas continuará a ser assimptoticamente consistente quando $m \to \infty$.

Note-se que a utilização de um estimador pelo método dos momentos da componentes de variância não tem qualquer influência sobre a expressão do BLUP espacial, nem do seu EQMP (expressões (4.5.7)-(4.5.10)). Contudo, as estimativas obtidas pelo EBLUP espacial (4.5.11) são influenciadas pelo método de estimação da componente de variância, bem como as estimativas do EQMP do EBLUP espacial, particularmente resultante de uma diferente variabilidade devida à estimação da componente de variância, avaliada através da conhecida parcela g_3 (4.5.13). Neste contexto, propõe-se o seguinte estimador analítico do EQMP do EBLUP espacial com correcção de enviesamento:

$$eqmp\left[\hat{\theta}_i(\hat{\sigma}_u^2)\right] = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2), \tag{5.9.3}$$

onde $g_{1i}(\sigma_u^2)$ é dado por (4.5.9), $g_{2i}(\sigma_u^2)$ é dado por (4.5.10) e $g_{3i}\left(\sigma_u^2\right)$ é dado por (4.5.13). Apesar das expressões de $g_{1i}(\sigma_u^2)$ e de $g_{2i}(\sigma_u^2)$ não serem afectadas pelo método de estimação da componente de variância, convém, no entanto, apresentá-las de forma mais detalhada. Assim, as variabilidades devidas à estimação dos efeitos aleatórios e dos efeitos fixos, são dadas, respectivamente, por:

$$\begin{aligned} g_{1i}\left(\sigma_u^2\right) &= \mathbf{m}_i'\left\{\sigma_u^2\mathbf{B}^{-1} - \sigma_u^2\mathbf{B}^{-1}\mathbf{V}^{-1}\sigma_u^2\mathbf{B}^{-1}\right\}\mathbf{m}_i \\ &= \mathbf{m}_i'\sigma_u^2\mathbf{B}^{-1}\mathbf{m}_i - \mathbf{m}_i'\sigma_u^2\mathbf{B}^{-1}\mathbf{V}^{-1}\sigma_u^2\mathbf{B}^{-1}\mathbf{m}_i \\ &= \sigma_u^2\boldsymbol{\varsigma}_i'\mathbf{m}_i - \sigma_u^2\boldsymbol{\varsigma}_i'\mathbf{V}^{-1}\sigma_u^2\boldsymbol{\varsigma}_i \\ &= \sigma_u^2\varsigma_{ii} - \sigma_u^2\boldsymbol{\varsigma}_i'\mathbf{V}^{-1}\sigma_u^2\boldsymbol{\varsigma}_i \end{aligned} \tag{5.9.4}$$

e

$$\begin{aligned} g_{2i}\left(\sigma_u^2\right) &= \left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{B}^{-1}\mathbf{m}_i\right]'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{B}^{-1}\mathbf{m}_i\right] \\ &= \left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varsigma}_i\right]'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{x}_i - \sigma_u^2\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varsigma}_i\right] \end{aligned} \tag{5.9.5}$$

onde $\boldsymbol{\varsigma}_i' = \{\varsigma_{ii'}\}$ é a *i*-ésima linha da matriz $\mathbf{B}^{-1}$. Por último, a variabilidade devida à estimação da componente de variância é dada por:

$$g_{3i}\left(\sigma_u^2\right) = tr\left[\mathbf{L}_i\left(\sigma_u^2\right)\mathbf{V}\left(\sigma_u^2\right)\mathbf{L}_i'\left(\sigma_u^2\right)\overline{\mathbf{V}}\left(\hat{\sigma}_u^2\right)\right], \tag{5.9.6}$$

onde $\mathbf{V}\left(\sigma_u^2\right)$ é dada por (4.5.5),

$$\begin{aligned}\mathbf{L}_i\left(\sigma_u^2\right) &= \frac{\partial \mathbf{b}_i'}{\partial \sigma_u^2} = \frac{\partial\left(\sigma_u^2 \boldsymbol{\varsigma}_i' \mathbf{V}^{-1}\right)}{\partial \sigma_u^2} \\ &= \boldsymbol{\varsigma}_i' \mathbf{V}^{-1} - \sigma_u^2 \boldsymbol{\varsigma}_i' \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_u^2} \mathbf{V}^{-1} \\ &= \boldsymbol{\varsigma}_i' \mathbf{V}^{-1} - \sigma_u^2 \boldsymbol{\varsigma}_i' \mathbf{V}^{-1} \mathbf{B}^{-1} \mathbf{V}^{-1} \\ &= \left(\boldsymbol{\varsigma}_i' - \sigma_u^2 \boldsymbol{\varsigma}_i' \mathbf{V}^{-1} \mathbf{B}^{-1}\right) \mathbf{V}^{-1}\end{aligned}$$

e a variância assimptótica de σ_u^2 é dada por $\overline{\mathbf{V}}\left(\tilde{\sigma}_u^2\right) = 2k_1^2 tr\left(\mathbf{C}_1 \mathbf{V} \mathbf{C}_1 \mathbf{V}\right)$, na qual $k_1 = \left[m - r\left(\mathbf{H}^{(1)}\right)\right]^{-1}$, $\mathbf{V} = \mathbf{R} + \sigma_u^2 \mathbf{B}^{-1}$ e $\mathbf{C}_1 = \mathbf{C}^{(1)'}\left[\mathbf{I}_m - \mathbf{C}^{(1)} \mathbf{X}\left(\mathbf{X}' \mathbf{C}^{(1)'} \mathbf{C}^{(1)} \mathbf{X}\right)^{-} \mathbf{X}' \mathbf{C}^{(1)'}\right] \mathbf{C}^{(1)}$

com $\mathbf{C}^{(1)} = \left(\mathbf{I}_m - \phi \mathbf{W}\right)$. Tal como foi apresentado no subcapítulo 5.6, a avaliação da variância assimptótica de um estimador da componente de variância foi baseada no lema sobre avaliação da covariância entre duas formas quadráticas de variáveis normalmente distribuídas. Notando-se que o estimador (5.9.2) pode ser apresentado como $\tilde{\sigma}_u^2 = k_1 \hat{\mathbf{e}}^{(1)'} \hat{\mathbf{e}}^{(1)} + k_2$, onde $k_1 = \left[m - r\left(\mathbf{H}^{(1)}\right)\right]^{-1}$ e $k_2 = -tr\left[\mathbf{P}_{H^{(1)}}\left(\mathbf{I}_m - \phi \mathbf{W}\right) \mathbf{R} \left(\mathbf{I}_m - \phi \mathbf{W}\right)'\right] \times k_1$, e notando-se que a soma dos quadrados dos resíduos pode ser apresentada como uma forma quadrática:

$$\begin{aligned}\hat{\mathbf{e}}^{(1)'} \hat{\mathbf{e}}^{(1)} &= \mathbf{e}^{(1)'} \mathbf{P}_{H^{(1)}} \mathbf{e}^{(1)} \\ &= \left[\mathbf{u} + \left(\mathbf{I}_m - \phi \mathbf{W}\right) \boldsymbol{\varepsilon}\right]' \mathbf{P}_{H^{(1)}} \left[\mathbf{u} + \left(\mathbf{I}_m - \phi \mathbf{W}\right) \boldsymbol{\varepsilon}\right] \\ &= \left[\left(\mathbf{I}_m - \phi \mathbf{W}\right)\left(\mathbf{v} + \boldsymbol{\varepsilon}\right)\right]' \mathbf{P}_{H^{(1)}} \left[\left(\mathbf{I}_m - \phi \mathbf{W}\right)\left(\mathbf{v} + \boldsymbol{\varepsilon}\right)\right] \\ &= \left(\mathbf{v} + \boldsymbol{\varepsilon}\right)' \mathbf{C}^{(1)'}\left[\mathbf{I}_m - \mathbf{H}^{(1)}\left(\mathbf{H}^{(1)'} \mathbf{H}^{(1)}\right)^{-} \mathbf{H}^{(1)'}\right] \mathbf{C}^{(1)}\left(\mathbf{v} + \boldsymbol{\varepsilon}\right) \\ &= \mathbf{a}' \mathbf{C}^{(1)'}\left[\mathbf{I}_m - \mathbf{C}^{(1)} \mathbf{X}\left(\mathbf{X}' \mathbf{C}^{(1)'} \mathbf{C}^{(1)} \mathbf{X}\right)^{-} \mathbf{X}' \mathbf{C}^{(1)'}\right] \mathbf{C}^{(1)} \mathbf{a}',\end{aligned}$$

onde $\mathbf{a} = \mathbf{v} + \boldsymbol{\varepsilon} \sim N(\mathbf{0}; \mathbf{V})$ e $\mathbf{C}^{(1)} = \left(\mathbf{I}_m - \phi \mathbf{W}\right)$, então a variância assimptótica de σ_u^2 obtém-se de imediato a partir do referido lema.

6. ESTIMAÇÃO COM RESTRIÇÕES

6.1 INTRODUÇÃO

Tal como tem sido referido nos capítulos anteriores, a estimação em pequenos domínios é um procedimento utilizado para produzir estimativas com precisão adequada para pequenas áreas geográficas ou subpopulações, denominadas por pequenos domínios. Contudo, quando se trabalha com estatísticas oficiais ou outros dados publicados, exige-se normalmente que as estimativas produzidas para um determinado nível de agregação coincidam com o resultado do agrupamento das estimativas produzidas para níveis de agregação inferiores (pequenos domínios). Por exemplo, exige-se que a estimativa directa do parâmetro de interesse para toda a população seja igual à soma ponderada das estimativas indirectas desse parâmetro de interesse em pequenos domínios que constituem essa população de interesse. Quando esta consistência interna dos estimadores não se verifica, é necessário que os produtores de estatísticas oficiais procedam a uma calibração das estimativas para pequenos domínios de forma a torná-las coerentes com as estimativas produzidas para um nível de agregação superior, as quais são normalmente mais precisas, pelo facto de serem produzidas com base numa maior dimensão amostral.

Esta calibração assume também um papel muito importante no quadro de uma abordagem de estimação do tipo *model-based* ou *model-assisted*, pois é conveniente que os estimadores dos parâmetros de interesse incorporem mecanismos que possam assegurar alguma protecção contra possíveis falhas ou más especificações dos modelos que os suportam. Esses mecanismos poderão conferir alguma robustez ao estimador ajustado, *i.e.*, poderão reduzir o seu enviesamento e o seu EQM, quando alguns dos pressupostos do modelo subjacente não se verificam. Na verdade, a grande vantagem da

garantia da consistência interna dos estimadores emerge quando o modelo postulado falha, podendo este procedimento assegurar que sejam produzidas estimativas adequadas para um determinado nível de agregação. Por outro lado, quando o modelo postulado se verifica aproximadamente, espera-se que a garantia da consistência interna não deva produzir alterações significativas nas estimativas.

Na primeira parte deste capítulo é efectuada uma revisão bibliográfica sobre a garantia da consistência interna dos estimadores no contexto da estimação em pequenos domínios (subcapítulo 6.2). Apesar de este problema ser mais conhecido e estudado no contexto das séries cronológicas económicas, no âmbito do qual as estimativas mensais ou trimestrais de um determinado parâmetro devem ser ajustadas de forma a serem consistentes com as respectivas estimativas anuais, normalmente mais precisas, ele não será estudado no âmbito desta tese. Este assunto pode ser encontrado, por exemplo, nos trabalhos de Hillmer e Trabelsi (1987) e de Cholette e Dagum (1994). Na segunda parte deste capítulo, *i.e.*, no subcapítulo 6.3, é introduzido um modelo linear misto geral com restrições para estimação em domínios, no âmbito do qual é deduzido o EBLUP com restrições. No final deste subcapítulo, são propostas duas metodologias gerais de estimação do EQMP do EBLUP com restrições baseadas em métodos por reamostragem, válidas para qualquer método de estimação dos parâmetros de interesse em níveis mais agregados. Este capítulo termina com a apresentação de um modelo de estimação em domínios com restrições, utilizando dados espaciais e cronológicos. Este último trabalho é apresentado no subcapítulo 6.4.

6.2 TRABALHOS PRÉVIOS

Os principais trabalhos sobre a garantia da consistência interna dos estimadores, no contexto da estimação em pequenos domínios, baseiam-se na introdução de restrições na estimação. Uma vez que geralmente se pretende que a soma ponderada das estimativas combinadas em pequenos domínios, que se supõem ter sido obtidas através da metodologia EBLUP apresentada no subcapítulo 3.2, seja igual à respectiva estimativa (directa ou indirecta) obtida para um nível de agregação mais elevado (para simplificação, admita-se a população total), então é desejável que os estimadores para

pequenos domínios, $\hat{\theta}_i$, satisfaçam a seguinte restrição (Wang *et al*., 2008):

$$\sum_{i=1}^{m} \omega_i \hat{\theta}_i = \sum_{i=1}^{m} \omega_i y_i \,, \tag{6.2.1}$$

onde y_i é um estimador *design-based* de θ_i e ω_i são pesos amostrais tais que $\sum_{i=1}^{m} \omega_i y_i$ é um estimador consistente no desenho do total (ou média) populacional. Têm sido propostos na literatura vários procedimentos que permitem construir estimadores que satisfazem a restrição (6.2.1). De forma a rever alguns desses procedimentos mais utilizados no contexto de modelos de nível área, considere-se o conhecido modelo de Fay-Herriot (Fay e Herriot, 1979) apresentado no subcapítulo 4.2. Tal como foi apresentado nesse subcapítulo, considere-se que $\tilde{\theta}_i$ e $\hat{\theta}_i$ representam o BLUP e o EBLUP de θ_i, respectivamente.

Pfeffermann e Barnard (1991) propuseram a modificação dos preditores dos parâmetros desse modelo, considerando uma restrição do tipo $\sum_{i=1}^{m} \omega_i \left(\mathbf{x}_i'\boldsymbol{\beta} + u_i\right) = \sum_{i=1}^{m} \omega_i y_i$. O respectivo estimador dos parâmetros de interesse com restrições é dado por:

$$\hat{\theta}_i^{R,PB} = \hat{\theta}_i + \left[V\left(\hat{\theta}\right)\right]^{-1} . \left[Cov\left(\hat{\theta}_i ; \hat{\theta}\right)\right] \left(\sum_{j=1}^{m} \omega_j y_j - \sum_{i=1}^{m} \omega_i \hat{\theta}_i \right), \tag{6.2.2}$$

onde $\hat{\theta} = \sum_{i=1}^{m} \omega_i \hat{\theta}_i$, $Cov\left(\hat{\theta}_i, \hat{\theta}\right) = \omega_i \gamma_i \sigma_{\varepsilon,i}^2 + \sum_{j=1}^{m} \omega_j \left(1-\gamma_i\right)\left(1-\gamma_j\right) \mathbf{x}_i' V\left(\tilde{\boldsymbol{\beta}}\right) \mathbf{x}_j'$ e $V\left(\hat{\theta}\right) = \sum_{i=1}^{m} \omega_i Cov\left(\hat{\theta}_i ; \hat{\theta}\right).$

Isaki *et al.* (2000) impuseram uma restrição através de um procedimento que, aproximadamente, produz os melhores preditores de *m*-1 quantidades não correlacionadas com $\sum_{i=1}^{m} \omega_i y_i$. Segundo Wang *et al.* (2008), o estimador dos parâmetros de interesse com restrições proposto por aqueles autores pode ser apresentado como:

$$\hat{\theta}_i^{R,ITF} = \hat{\theta}_i + \left[\sum_{j=1}^{m} \omega_j^2 \hat{V}(y_j)\right]^{-1} .\omega_i \hat{V}(y_i).\left(\sum_{j=1}^{m} \omega_j y_j - \sum_{i=1}^{m} \omega_i \hat{\theta}_i\right), \qquad (6.2.3)$$

onde $\hat{\theta} = \sum_{i=1}^{m} \omega_i \hat{\theta}_i$ e $\hat{V}(y_i)$ é um estimador de $V(y_i) = \sigma_u^2 + \sigma_{\varepsilon,i}^2$.

No trabalho de Wang *et al*. (2008), no qual foi observado que os estimadores (6.2.2) e (6.2.3) têm a forma geral:

$$\hat{\theta}_i^R = \hat{\theta}_i + a_i.\left(\sum_{j=1}^{m} \omega_j y_j - \sum_{i=1}^{m} \omega_i \hat{\theta}_i\right), \qquad (6.2.4)$$

onde $\sum_{i=1}^{m} \omega_i a_i = 1$, foi também deduzido o "melhor" preditor linear não enviesado (BLUP) de θ_i que satisfaz a restrição (6.2.1). Esse preditor é dado por:

$$\hat{\theta}_i^R = \hat{\theta}_i + \breve{a}_i.\left(\sum_{j=1}^{m} \omega_j y_j - \sum_{i=1}^{m} \omega_i \hat{\theta}_i\right), \qquad (6.2.5)$$

onde $\breve{a}_i = \left(\sum_{i=1}^{m} \varphi_i^{-1} w_i^2\right)^{-1} \varphi_i^{-1} w_i$ e φ_i é uma função das componentes de variância. Considerando $\varphi_i = \omega_i \left[Cov(\hat{\theta}_i; \hat{\theta})\right]^{-1}$ e $\varphi_i = \left[\hat{V}(y_i)\right]^{-1}$ obtêm-se, respectivamente, os estimadores (6.2.2) e (6.2.3), pelo que se verifica, desta forma, que são BLUP.

No contexto de modelos de estimação em pequenos domínios de nível unidade têm também sido propostos procedimentos semelhantes aos acima apresentados. Battese *et al*. (1988) foram pioneiros na introdução de restrições, no âmbito do seu conhecido modelo de nível unidade, de forma a garantir que a soma ponderada das estimativas de colheitas médias para pequenas áreas geográficas fosse igual à estimativa da colheita total para uma área geográfica mais ampla. Coelho (2000) propôs uma metodologia geral que permite a introdução de restrições na estimação sob modelos espaciotemporais de nível unidade. Mais recentemente, Ugarte *et al*. (2009) sistematizaram a metodologia de introdução de restrições sob o modelo de Battese *et al*. (1988), e apresentaram um trabalho inovador ao nível da estimação do EQM do EBLUP com restrições, admitindo que as estimativas para níveis mais agregados são obtidas através de um estimador

sintético pela regressão. Por sua vez, Pfeffermann e Burk (1990) e Pfeffermann e Tiller (2006) impuseram restrições do tipo (6.2.1) em modelos *state space* aplicados à estimação de preços da habitação e à estimação de taxas de desemprego, respectivamente.

6.3 MODELO COM RESTRIÇÕES

6.3.1 Especificação do modelo com restrições

Considere-se um problema de estimação em pequenos domínios no quadro de um modelo linear misto geral, tal como o modelo apresentado no subcapítulo 3.2:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{\varepsilon}, \tag{6.3.1}$$

onde $\boldsymbol{y}$ é um vector $m\times1$ de observações da variável dependente, $\boldsymbol{X} = col_{1\le i\le m}(\boldsymbol{x'}_i)$ é uma matriz $m\times p$ de observações de variáveis independentes com $\boldsymbol{x'}_i = (x_{i1}, \dots, x_{ip})$, $\boldsymbol{\beta}$ é um vector $p\times1$ de efeitos fixos, $\boldsymbol{Z} = \boldsymbol{I}_m$, $\boldsymbol{v}$ é um vector $m\times1$ de efeitos aleatórios e $\boldsymbol{\varepsilon}$ é um vector $m\times1$ de variáveis residuais, os quais são independentemente distribuídos com $\boldsymbol{v} \sim N(\boldsymbol{0}; \boldsymbol{G})$ e $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}; \boldsymbol{R})$. A matriz de covariâncias de $\boldsymbol{y}$ é dada por $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$. Recorde-se que a estimação de um parâmetro de interesse geral num domínio i, $\boldsymbol{\theta_i} = \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{m'_i v}$, é efectuada com base no seguinte BLUE de $\boldsymbol{\beta}$, $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X'}\mathbf{V^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X'}\mathbf{V^{-1}}\mathbf{y}$, e no seguinte BLUP de $\boldsymbol{v}$, $\widetilde{\boldsymbol{v}} = \boldsymbol{GZ'}\mathbf{V^{-1}}(\mathbf{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})$. Com o objectivo de introduzir restrições na estimação, tendo em vista a garantia da consistência interna, é conveniente apresentar o modelo (6.3.1) como:

$$\boldsymbol{y} = \boldsymbol{X^0\xi} + \boldsymbol{\varepsilon}, \tag{6.3.2}$$

onde $\boldsymbol{X^0} = [\boldsymbol{X}\ \boldsymbol{Z}]$ e $\boldsymbol{\xi} = [\boldsymbol{\beta'}\ \boldsymbol{v'}]'$. Assumindo que as componentes de variância são conhecidas, então o preditor dos mínimos quadrados generalizados de $\boldsymbol{\xi}$ é dado por (Henderson, 1975):

$$\widetilde{\boldsymbol{\xi}} = (\boldsymbol{X^{0\prime}}\boldsymbol{V^{-1}}\boldsymbol{X^0})^{-1}\boldsymbol{X^{0\prime}}\mathbf{V^{-1}}\boldsymbol{y} = \begin{bmatrix}(\boldsymbol{X'}\mathbf{V^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X'}\mathbf{V^{-1}}\mathbf{y} \\ \boldsymbol{GZ'}\mathbf{V^{-1}}(\mathbf{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})\end{bmatrix}. \tag{6.3.3}$$

Pfeffermann (1984) mostrou que a matriz de covariâncias dos erros de predição de $\boldsymbol{\xi}$, sob o modelo (6.3.2), $\tilde{\tilde{\boldsymbol{\xi}}} - \boldsymbol{\xi}$, é dada por:

$$V\left(\tilde{\tilde{\boldsymbol{\xi}}} - \boldsymbol{\xi}\right) = E\left[\left(\tilde{\tilde{\boldsymbol{\xi}}} - \boldsymbol{\xi}\right)\left(\tilde{\tilde{\boldsymbol{\xi}}} - \boldsymbol{\xi}\right)'\right] = (\boldsymbol{X^0}'\boldsymbol{V}^{-1}\boldsymbol{X^0})^{-1} = \boldsymbol{C}. \qquad (6.3.4)$$

Posteriormente, McLean e Sanders (1988) mostraram que a matriz de covariâncias dos erros de predição pode igualmente ser apresentada como:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}'\mathbf{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \boldsymbol{X}'\mathbf{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}'_{21} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \qquad (6.3.5)$$

onde $\mathbf{C}_{11} = (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}$, $\mathbf{C}_{21} = -\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}$ e $\mathbf{C}_{22} = (\boldsymbol{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} - \mathbf{C}_{21}\boldsymbol{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}$.

A abordagem de garantia da consistência interna que se vai propor passa por considerar, à semelhança das abordagens apresentadas anteriormente, a existência de áreas geográficas ou subpopulações de maior dimensão que agregam os pequenos domínios de interesse, para as quais se dispõe de estimativas com precisão adequada. As estimativas dos parâmetros de interesse ao nível da região podem ser obtidas de forma directa ou indirecta, sendo frequentemente utilizados os estimadores de Horvitz-Thompson, directo pós-estratificado e sintético pela regressão. É importante observar que as restrições devem reportar-se a um nível de agregação para o qual se disponham de estimadores com um bom nível de precisão. Só neste caso é possível ter a garantia que as modificações impostas aos estimadores são necessárias e que não interferem com as flutuações aleatórias dos dados.

Para tornar a leitura mais simples neste capítulo, considere-se a seguinte notação: *i* representa, tal como anteriormente, os pequenos domínios, *i*=1, …, *m*, *a* representa áreas geográficas mais vastas denominadas por regiões, *a*=1, …, *A*. De forma genérica propõe utilizar-se o termo "região" para referir uma população que agrega um conjunto de pequenos domínios, independentemente na sua natureza, representando-se por $m(a)$ o número total de pequenos domínios contidos na região *a*. Finalmente, $\sum_{i \in a}$ representa a soma sobre todos os pequenos domínios pertencentes à região *a*.

Assume-se, também, que o parâmetro de interesse é uma média, exigindo-se que a média ponderada das estimativas produzidas para os pequenos domínios de uma dada região seja igual à estimativa do referido parâmetro de interesse nessa região. Neste

sentido, os estimadores/preditores dos efeitos fixos e aleatórios do modelo devem ser modificados pelo seguinte conjunto de restrições:

$$\sum_{i\in a} \omega_i \tilde{\theta}_i^R = \hat{\theta}_a \ , \quad a = 1, \dots, A \tag{6.3.6}$$

onde ω_i são ponderadores tal que $\sum_{i\in a} \omega_i = 1$, $\hat{\theta}_a$ é um estimador da média na região *a* e $\tilde{\theta}_i^R = \boldsymbol{k'}_i \widetilde{\boldsymbol{\beta}}^R + \boldsymbol{m'}_i \widetilde{\boldsymbol{v}}^R$ é um estimador combinado modificado da média no *i*-ésimo pequeno domínio, sendo $\widetilde{\boldsymbol{\beta}}^R$ o vector dos estimadores modificados dos efeitos fixos e $\widetilde{\boldsymbol{v}}^R$ o vector dos preditores modificados dos efeitos aleatórios. Apesar da escolha dos ponderadores ω_i ser arbitrária, admite-se que é dada pelo rácio entre o número de observações incluídas no *i*-ésimo pequeno domínio e o número de observações incluídas na *a*-ésima região, $\omega_i = \frac{n_i}{n_a}$. Pretende-se, portanto, substituir o preditor $\tilde{\boldsymbol{\xi}}$ por $\tilde{\boldsymbol{\xi}}^{\boldsymbol{R}} = \left[\widetilde{\boldsymbol{\beta}}^{\boldsymbol{R}\prime} \ \ \widetilde{\boldsymbol{v}}^{\boldsymbol{R}\prime}\right]'$, através da resolução do seguinte problema de minimização:

$$\begin{aligned} &\min_{\xi} (\boldsymbol{y} - \boldsymbol{X^0\xi})' \mathbf{V}^{-1} (\boldsymbol{y} - \boldsymbol{X^0\xi}) \\ &s.a. \quad \boldsymbol{Q\xi} = \boldsymbol{q}. \end{aligned} \tag{6.3.7}$$

Note-se que o conjunto de *A* restrições (6.3.6) pode ser apresentado de forma compacta como:

$$\boldsymbol{Q\xi} = \boldsymbol{q} \ , \tag{6.3.8}$$

onde $\boldsymbol{q} = col_{1\le a\le A}(\hat{\theta}_a)$, $\boldsymbol{Q} = [\boldsymbol{X}^R \ \boldsymbol{Q}_v]$, $\boldsymbol{X}^R = col_{1\le a\le A}(\boldsymbol{\delta'}_a \boldsymbol{\Psi X})$, $\boldsymbol{Q}_v = col_{1\le a\le A}(\boldsymbol{\delta'}_a \boldsymbol{\Psi})$, $\boldsymbol{\Psi} = diag_{1\le i\le m}(\omega_i)$, e $\boldsymbol{\delta'}_a = (\delta_{a1}, \dots, \delta_{am})$ com δ_{ai}=1 se o *i*-ésimo domínio pertence à *a*-ésima região e δ_{ai}=0 em caso contrário. Note-se que $\boldsymbol{\delta'}_a \boldsymbol{\Psi X\beta} + \boldsymbol{\delta'}_a \boldsymbol{\Psi v} = \sum_{i\in a} \omega_i \boldsymbol{x'}_i \boldsymbol{\beta} + \sum_{i\in a} \omega_i v_i$.

A resolução do problema (6.3.7) é semelhante à resolução do conhecido problema da econometria clássica de estimação dos parâmetros de uma regressão sob restrições lineares (Greene, 2003). A solução do problema é dada por:

$$\tilde{\boldsymbol{\xi}}^R = \tilde{\boldsymbol{\xi}} + \boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right), \tag{6.3.9}$$

onde $\boldsymbol{A} = \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}$ com $\boldsymbol{C} = (\boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{X^0})^{-1}$. Os preditores modificados (6.3.9) verificam naturalmente as restrições (6.3.8), uma vez que $\boldsymbol{Q}\tilde{\boldsymbol{\xi}}^R = \boldsymbol{Q}\left[\tilde{\boldsymbol{\xi}} + \boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)\right] = \boldsymbol{Q}\tilde{\boldsymbol{\xi}} + \boldsymbol{QCQ'}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) = \boldsymbol{q}$.

Demonstração: Considere-se a seguinte função Lagrangeana:

$$\mathrm{L}(\boldsymbol{\xi},\boldsymbol{\lambda}) = (\boldsymbol{y}-\boldsymbol{X^0\xi})'\mathbf{V}^{-1}(\boldsymbol{y}-\boldsymbol{X^0\xi}) + 2\boldsymbol{\lambda}'(\boldsymbol{Q\xi}-\boldsymbol{q}),$$

e igualando-se a matrizes nulas as suas derivadas em ordem a $\boldsymbol{\xi}$ e a $\boldsymbol{\lambda}$, tem-se

$$\begin{cases} \frac{\partial \mathrm{L}}{\partial \boldsymbol{\xi}} = -2\boldsymbol{X^{0\prime}}\mathbf{V}^{-1}(\boldsymbol{y}-\boldsymbol{X^0\xi}) + 2\boldsymbol{Q'\lambda} = \mathbf{0} \\ \frac{\partial \mathrm{L}}{\partial \boldsymbol{\lambda}} = 2(\boldsymbol{Q\xi}-\boldsymbol{q}) = \mathbf{0} \end{cases} \Leftrightarrow \begin{cases} \boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{X^0\xi} + \boldsymbol{Q'\lambda} = \boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{y} \\ \boldsymbol{Q\xi} = \boldsymbol{q} \end{cases}.$$

O resultado anterior pode ser reescrito utilizando matrizes por blocos,

$$\underbrace{\begin{bmatrix} \boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{X^0} & \boldsymbol{Q'} \\ \boldsymbol{Q} & \mathbf{0} \end{bmatrix}}_{\boldsymbol{B}} \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{y} \\ \boldsymbol{q} \end{bmatrix} \Leftrightarrow \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\lambda} \end{bmatrix} = \boldsymbol{B}^{-1}\begin{bmatrix} \boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{y} \\ \boldsymbol{q} \end{bmatrix},$$

e calculando-se a matriz inversa de $\boldsymbol{B}$ (veja-se, por exemplo, em Greene (2003) a metodologia de cálculo da matriz inversa de uma matriz por blocos),

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{C}[\boldsymbol{I}-\boldsymbol{Q'}(\boldsymbol{QCQ'})^{-1}\boldsymbol{QC}] & \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1} \\ (\boldsymbol{QCQ'})^{-1}\boldsymbol{QC} & -(\boldsymbol{QCQ'})^{-1} \end{bmatrix},$$

onde $\boldsymbol{C} = (\boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{X^0})^{-1}$, tem-se finalmente que:

$$\begin{aligned} \tilde{\boldsymbol{\xi}}^R &= \boldsymbol{CX^{0\prime}}\mathbf{V}^{-1}\boldsymbol{y} - \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\boldsymbol{QCX^{0\prime}}\mathbf{V}^{-1}\boldsymbol{y} + \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\boldsymbol{q} \\ &= \tilde{\boldsymbol{\xi}} - \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\boldsymbol{Q}\tilde{\boldsymbol{\xi}} + \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\boldsymbol{q} = \tilde{\boldsymbol{\xi}} + \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q}-\boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) \\ &= \tilde{\boldsymbol{\xi}} + \boldsymbol{A}\left(\boldsymbol{q}-\boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) \end{aligned}$$

com $\boldsymbol{A} = \boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}$. ■

Note-se, também, que as expressões (6.3.8) e (6.3.9) podem ser genericamente aplicadas, independentemente do estimador utilizado ao nível da região (naturalmente com as correspondentes expressões para $\boldsymbol{Q}$ e $\boldsymbol{q}$). Contudo, é necessário conhecer-se a forma específica do estimador utilizado ao nível da região, de forma a derivar as respectivas expressões particulares do enviesamento e da variância do preditor (6.3.9).

Lema 6.1: *O valor esperado no modelo do erro de predição* $\left(\tilde{\boldsymbol{\xi}}^{R}-\boldsymbol{\xi}\right)$ *é dado por:*

$$\mathrm{E}\left(\tilde{\boldsymbol{\xi}}^{\mathrm{R}}-\boldsymbol{\xi}\right)=\mathbf{A}\left(\mathbf{q}_{\boldsymbol{\mu}}-\mathbf{X}^{\mathrm{R}}\boldsymbol{\beta}\right), \tag{6.3.10}$$

onde $\boldsymbol{q_{\mu}}=E(\boldsymbol{q})$.

Demonstração:

$$E(\tilde{\boldsymbol{\xi}}^{R}-\boldsymbol{\xi})=E[\tilde{\boldsymbol{\xi}}+\boldsymbol{A}(\boldsymbol{q}-\boldsymbol{Q}\tilde{\boldsymbol{\xi}})-\boldsymbol{\xi}]=E(\tilde{\boldsymbol{\xi}}-\boldsymbol{\xi})+\boldsymbol{A}E(\boldsymbol{q}-\boldsymbol{Q}\tilde{\boldsymbol{\xi}})=$$

$$=\mathbf{0}+\boldsymbol{A}E(\boldsymbol{q})-\boldsymbol{A}[\boldsymbol{X}^{R}\ \boldsymbol{Q}_{v}]E\left(\tilde{\boldsymbol{\xi}}\right)=\boldsymbol{A}\boldsymbol{q}_{\boldsymbol{\mu}}-\boldsymbol{A}\boldsymbol{X}^{R}\boldsymbol{\beta}\ .$$

Note-se que $E\left(\tilde{\boldsymbol{\xi}}\right)=\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{0}\end{bmatrix}$ sob o modelo (6.3.1) e se denominou $\boldsymbol{q_{\mu}}=E(\boldsymbol{q})$ pelo facto de depender do estimador utilizado ao nível da região. ■

Corolário 6.1.1: *Quando é utilizado o estimador directo ao nível da região,* $\hat{\theta}_{a}^{dir}$*, então o enviesamento no modelo do preditor* $\tilde{\boldsymbol{\xi}}^{R}$ *é dado por:*

$$E\left(\tilde{\boldsymbol{\xi}}^{R}-\boldsymbol{\xi}\right)=\boldsymbol{A}(\boldsymbol{\Upsilon}\boldsymbol{X}-\boldsymbol{X}^{R})\boldsymbol{\beta}.$$

Demonstração: O estimador directo da média na *a*-ésima região, sob um plano de sondagem genérico com dimensões populacionais dos domínios desconhecidas, é dado por:

$$\hat{\theta}_{a}^{dir}=\sum_{j\in s_a}\pi_j^{-1}y_j/\sum_{j\in s_a}\pi_j^{-1}=\hat{\tau}_a/\widehat{N}_a,$$

onde π_j é a probabilidade de inclusão da *j*-ésima observação y_j, *j*=1, …, N. Notando-se que o estimador $\hat{\tau}_a$ pode ser apresentado como:

$$\hat{\tau}_a=\sum_{i\in a}\sum_{j\in s_{ai}}\pi_j^{-1}y_j=\sum_{i\in a}\hat{\tau}_{ai}=\sum_{i\in a}\widehat{N}_i\hat{\theta}_i$$

e que do ponto de vista do modelo $E\left(\hat{\theta}_i\right)=E(\boldsymbol{x'}_i\boldsymbol{\beta}+\upsilon_i+\boldsymbol{\varepsilon}_i)=\boldsymbol{x'}_i\boldsymbol{\beta}$, $E\left(\widehat{N}_i\right)=E\left(\sum_{j\in s_{ai}}\pi_j^{-1}\right)=\sum_{j\in s_{ai}}\pi_j^{-1}=\pi_i$ e $E\left(\widehat{N}_a\right)=E\left(\sum_{j\in s_a}\pi_j^{-1}\right)=\sum_{j\in s_a}\pi_j^{-1}=\pi_a$, uma vez que as probabilidades de inclusão são constantes neste contexto, então:

$$E\left(\hat{\theta}_{a}^{dir}\right)=\sum_{i\in a}\pi_i\,(\boldsymbol{x'}_i\boldsymbol{\beta})/\pi_a=\frac{1}{\pi_a}\boldsymbol{\delta'}_a\boldsymbol{\Pi}\boldsymbol{X}\boldsymbol{\beta}\ ,$$

onde $\boldsymbol{\Pi} = diag_{1\leq i\leq n}(\pi_i)$. Tem-se, então, que $E(\boldsymbol{q}) = \boldsymbol{\Upsilon X \beta}$, onde $\boldsymbol{\Upsilon} = col_{1\leq a\leq A}\left(\frac{1}{\pi_a}\boldsymbol{\delta'}_a\boldsymbol{\Pi}\right)$, pelo que de imediato se verifica que $E\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right) = \boldsymbol{A}(\boldsymbol{\Upsilon X \beta} - \boldsymbol{X}^R\,\boldsymbol{\beta})$. ■

Observação: O enviesamento do preditor $\tilde{\boldsymbol{\xi}}^R$ quando é utilizado o estimador directo pode ser negligenciável. Note-se que:

$$\boldsymbol{\Upsilon X} - \boldsymbol{X}^R = col_{1\leq a\leq A}\left(\frac{1}{\pi_a}\boldsymbol{\delta'}_a\boldsymbol{\Pi}\right)\boldsymbol{X} - col_{1\leq a\leq A}(\boldsymbol{\delta'}_a\boldsymbol{\Psi X}) =$$

$$= col_{1\leq a\leq A}\left(\frac{1}{\pi_a}\boldsymbol{\delta'}_a\boldsymbol{\Pi} - \boldsymbol{\delta'}_a\boldsymbol{\Psi}\right)\boldsymbol{X}$$

$$= col_{1\leq a\leq A}\left\{\boldsymbol{\delta'}_a\left(\frac{1}{\pi_a}\boldsymbol{\Pi} - \boldsymbol{\Psi}\right)\right\}\boldsymbol{X}$$

$$= col_{1\leq a\leq A}\left\{\boldsymbol{\delta'}_a diag_{1\leq i\leq n}\left(\frac{\pi_i}{\pi_a} - \omega_i\right)\right\}\boldsymbol{X}\,,$$

pelo que a diferença entre $\omega_i = \frac{n_i}{n_a}$ e $\frac{\pi_i}{\pi_a} = \frac{\widehat{N}_i}{\widehat{N}_a}$ tenderá a ser próxima de zero. No caso particular de uma sondagem aleatória simples sem reposição tem-se que $\widehat{N}_i = n_i\frac{N}{n}$ e $\widehat{N}_a = n_a\frac{N}{n}$, pelo que se verifica de imediato que o preditor $\tilde{\boldsymbol{\xi}}^R$ é centrado no modelo. É, ainda, de salientar que sendo arbitrária a escolha dos pesos ω_i, ela pode ser efectuada de forma a tornar o preditor $\tilde{\boldsymbol{\xi}}^R$ centrado no modelo num contexto de um qualquer outro plano de sondagem, caso seja conveniente.

Corolário 6.1.2: *O preditor $\tilde{\boldsymbol{\xi}}^R$ é centrado no modelo quando é utilizado o estimador sintético pela regressão ao nível da região.*

Demonstração: O estimador sintético pela regressão da média na *a*-ésima região é dado por $\hat{\theta}_a^{SR} = \boldsymbol{x'}_a\tilde{\boldsymbol{\beta}}$, onde $\boldsymbol{x'}_a = \sum_{i\in a}\omega_i\,\boldsymbol{x'}_i = \boldsymbol{\delta'}_a\boldsymbol{\Psi X}$. Tem-se então que $E(\boldsymbol{q}) = E\left(\boldsymbol{X}^R\,\tilde{\boldsymbol{\beta}}\right) = \boldsymbol{X}^R\,\boldsymbol{\beta}$, pelo que é garantido de imediato o não enviesamento de $\tilde{\boldsymbol{\xi}}^R$ uma vez que $E(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}) = \boldsymbol{0}$. ■

Lema 6.2: *Quando é utilizado um estimador directo ao nível da região e admitindo-se que o preditor* $\tilde{\boldsymbol{\xi}}^R$ *é centrado no modelo, então a covariância no modelo do erro de predição* $\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right)$ *é dada por:*

$$V\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right) = \boldsymbol{C} + \boldsymbol{M}\boldsymbol{A}' + \boldsymbol{A}\boldsymbol{M}' + \boldsymbol{A}\boldsymbol{N}\boldsymbol{A}', \tag{6.3.11}$$

onde $\boldsymbol{M} = \mathbf{C}_{.1}(\boldsymbol{\Upsilon}\mathbf{X} - \boldsymbol{X}^R)'$ *com* $\mathbf{C}_{.1} = [\mathbf{C'}_{11} \quad \mathbf{C'}_{21}]'$ *e* $\boldsymbol{N} = (\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)' + (\boldsymbol{\Upsilon}\boldsymbol{V} - \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}')\boldsymbol{\Upsilon}' - (\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{\Upsilon}(\boldsymbol{Z}\boldsymbol{G} + \boldsymbol{X}\mathbf{C}'_{21})\boldsymbol{Q}'_v - (\boldsymbol{X}^R\mathbf{C}_{11} + \boldsymbol{Q}_v\mathbf{C}_{21})\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'(\boldsymbol{I} - \mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}')\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}'_v$.

Demonstração:

Antes de se avançar para o cálculo da covariância no modelo do erro de predição de $\boldsymbol{\xi}$, é necessário observar que o BLUE de $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$, e o BLUP de $\boldsymbol{v}$, $\tilde{\boldsymbol{v}}$, podem ser reescritos como:

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$= (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\varepsilon})$$

$$= (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{e}$$

$$= \boldsymbol{\beta} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}\,,$$

$$\tilde{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)$$

$$= \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\varepsilon}) - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}(\boldsymbol{\beta} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon})$$

$$= \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}$$

$$= (\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v} + (\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}\,.$$

A covariância no modelo do erro de predição de $\boldsymbol{\xi}$ é dada por uma soma de quatro parcelas:

$$V\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right) = E\left[\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right)\left(\tilde{\boldsymbol{\xi}}^R - \boldsymbol{\xi}\right)'\right] = E\left\{\left[\tilde{\boldsymbol{\xi}} + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}) - \boldsymbol{\xi}\right]\left[\tilde{\boldsymbol{\xi}} + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}) - \boldsymbol{\xi}\right]'\right\}$$

$$= E\left\{\left[\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})\right]\left[\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})\right]'\right\}$$

$$= E\left[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})' + (\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A}' + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})' + \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A}'\right]$$

$$= E\left[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})'\right] + E\left[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A}'\right] + E\left[\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})'\right] + E\left[\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A}'\right] = \boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2 + \boldsymbol{\zeta}_3 + \boldsymbol{\zeta}_4$$

A primeira parcela da covariância no modelo do erro de predição de $\boldsymbol{\xi}$ é a covariância do erro de previsão de $\boldsymbol{\xi}$ sob o modelo sem restrições, dada por (6.3.4), logo $\boldsymbol{\zeta}_1 = \boldsymbol{C}$.

A segunda parcela da covariância no modelo do erro de predição de $\boldsymbol{\xi}$ é dada por:

$$\boldsymbol{\zeta}_2 = E\left[(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A}'\right] = E\left\{\begin{bmatrix}\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{v}} - \boldsymbol{v}\end{bmatrix}(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})'\right\}\boldsymbol{A}'$$

$$= E\left\{\begin{bmatrix}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})' \\ (\tilde{\boldsymbol{v}} - \boldsymbol{v})(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})'\end{bmatrix}\right\}\boldsymbol{A}'$$

Em seguida irá calcular-se separadamente o valor esperado de cada um dos dois blocos da matriz. Para tal vai utilizar-se a formulação alternativa do BLUE $\tilde{\boldsymbol{\beta}}$ e do BLUP $\tilde{\boldsymbol{v}}$, bem como os resultados obtidos na demonstração do Corolário 6.1.1 para notar que, do ponto de vista do modelo, se tem $\boldsymbol{q} = \boldsymbol{\Upsilon}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\varepsilon})$.

$$E\left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})'\right] =$$

$$\begin{aligned}= E\{&[\boldsymbol{\beta} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} - \boldsymbol{\beta}][\boldsymbol{\Upsilon}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Upsilon}\boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\Upsilon}\boldsymbol{\varepsilon} - \boldsymbol{X}^R\boldsymbol{\beta} \\ &- \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} - \boldsymbol{Q}_v(\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v} \\ &- \boldsymbol{Q}_v(\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1} - \boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}]'\}\end{aligned}$$

$$\begin{aligned}= E\{&[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v} + \mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}][(\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)\boldsymbol{\beta} \\ &+ (\boldsymbol{\Upsilon}\boldsymbol{Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v} \\ &+ (\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}]'\}\end{aligned}$$

$$\begin{aligned}= E[&\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v}\boldsymbol{\beta}'(\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)'] + E[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v}\boldsymbol{v}'(\boldsymbol{\Upsilon}\boldsymbol{Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \\ &\boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'] + E[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{v}\boldsymbol{\varepsilon}'(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \\ &\boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})'] + E[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\beta}'(\boldsymbol{\Upsilon}\boldsymbol{X} - \boldsymbol{X}^R)'] +\end{aligned}$$

$$E[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\upsilon}'(\boldsymbol{\Upsilon}\boldsymbol{Z}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'] +$$
$$E[\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\boldsymbol{\Upsilon}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})']$$

E recordando que os erros do modelo, $\boldsymbol{\upsilon}$ e $\boldsymbol{\varepsilon}$, satisfazem $E(\boldsymbol{\upsilon})=\mathbf{0}$, $E(\boldsymbol{\varepsilon})=\mathbf{0}$, $E(\boldsymbol{\upsilon}\boldsymbol{\upsilon}')=\boldsymbol{G}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')=\boldsymbol{R}$ e $E(\boldsymbol{\upsilon}\boldsymbol{\varepsilon}')=\mathbf{0}$, e que $\boldsymbol{V}=\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R}$, tem-se que:

$$E\left[(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\boldsymbol{q}-\boldsymbol{X}^{R}\widetilde{\boldsymbol{\beta}}-\boldsymbol{Q}_{\upsilon}\widetilde{\boldsymbol{\upsilon}})'\right]=$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\Upsilon}\boldsymbol{Z}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'+$$
$$\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}(\boldsymbol{\Upsilon}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})'$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Upsilon}'-\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'}-\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+$$
$$\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\boldsymbol{\Upsilon}'-$$
$$\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'}-\ \mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+$$
$$\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R})\boldsymbol{\Upsilon}'-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R})\boldsymbol{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'}-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+$$
$$\boldsymbol{R})\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R})\boldsymbol{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}'-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'}-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}'-\mathbf{C}_{11}\mathbf{C}_{\mathbf{11}}^{-1}\mathbf{C}_{11}\boldsymbol{X}^{R'}-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+\mathbf{C}_{11}\mathbf{C}_{\mathbf{11}}^{-1}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'$$

$$=\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}'-\mathbf{C}_{11}\boldsymbol{X}^{R'}-\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'+\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}_{\upsilon}'$$

$$=\mathbf{C}_{11}(\boldsymbol{\Upsilon}\mathbf{X}-\boldsymbol{X}^{R})'.$$

Seguindo o mesmo raciocínio, tem-se que:

$$E\left[(\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon})(\boldsymbol{q}-\boldsymbol{X}^{R}\widetilde{\boldsymbol{\beta}}-\boldsymbol{Q}_{\upsilon}\widetilde{\boldsymbol{\upsilon}})'\right]=$$

$$=E\{[(\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}-\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}-\boldsymbol{I})\boldsymbol{\upsilon}$$
$$+(\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}-\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}][(\boldsymbol{\Upsilon}\boldsymbol{X}-\boldsymbol{X}^{R})\boldsymbol{\beta}$$
$$+(\boldsymbol{\Upsilon}\boldsymbol{Z}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{Z}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{\upsilon}$$
$$+(\boldsymbol{\Upsilon}-\boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}-\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}+\boldsymbol{Q}_{\upsilon}\boldsymbol{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}]'\}$$

$$
\begin{aligned}
&= E[(\boldsymbol{GZ'V^{-1}Z - GZ'V^{-1}XC_{11}X'V^{-1}Z - I})\boldsymbol{v\beta'}(\boldsymbol{\Upsilon X - X^R})'] \\
&\quad + E[(\boldsymbol{GZ'V^{-1}Z - GZ'V^{-1}XC_{11}X'V^{-1}Z - I})\boldsymbol{vv'}(\boldsymbol{\Upsilon Z - X^RC_{11}X'V^{-1}Z} \\
&\quad \boldsymbol{- Q_v GZ'V^{-1}Z + Q_v GZ'V^{-1}XC_{11}X'V^{-1}Z})'] \\
&\quad + E[(\boldsymbol{GZ'V^{-1}Z - GZ'V^{-1}XC_{11}X'V^{-1}Z - I})\boldsymbol{v\varepsilon'}(\boldsymbol{\Upsilon - X^RC_{11}X'V^{-1} - Q_v GZ'V^{-1}} \\
&\quad \boldsymbol{+ Q_v GZ'V^{-1}XC_{11}X'V^{-1}})'] + E[(\boldsymbol{GZ'V^{-1} - GZ'V^{-1}XC_{11}X'V^{-1}})\boldsymbol{\varepsilon\beta'}(\boldsymbol{\Upsilon X - X^R})'] \\
&\quad + E[(\boldsymbol{GZ'V^{-1} - GZ'V^{-1}XC_{11}X'V^{-1}})\boldsymbol{\varepsilon v'}(\boldsymbol{\Upsilon Z - X^RC_{11}X'V^{-1}Z - Q_v GZ'V^{-1}Z} \\
&\quad \boldsymbol{+ Q_v GZ'V^{-1}XC_{11}X'V^{-1}Z})'] \\
&\quad + E[(\boldsymbol{GZ'V^{-1} - GZ'V^{-1}XC_{11}X'V^{-1}})\boldsymbol{\varepsilon\varepsilon'}(\boldsymbol{\Upsilon - X^RC_{11}X'V^{-1} - Q_v GZ'V^{-1}} \\
&\quad \boldsymbol{+ Q_v GZ'V^{-1}XC_{11}X'V^{-1}})'] \\
&= (\boldsymbol{GZ'V^{-1}Z - GZ'V^{-1}XC_{11}X'V^{-1}Z} \\
&\quad \boldsymbol{- I})\boldsymbol{G}(\boldsymbol{\Upsilon Z - X^RC_{11}X'V^{-1}Z - Q_v GZ'V^{-1}Z + Q_v GZ'V^{-1}XC_{11}X'V^{-1}Z})' \\
&\quad + (\boldsymbol{GZ'V^{-1} - GZ'V^{-1}XC_{11}X'V^{-1}})\boldsymbol{R}(\boldsymbol{\Upsilon - X^RC_{11}X'V^{-1} - Q_v GZ'V^{-1}} \\
&\quad \boldsymbol{+ Q_v GZ'V^{-1}XC_{11}X'V^{-1}})' \\
&= \boldsymbol{GZ'V^{-1}ZGZ'\Upsilon' - GZ'V^{-1}ZGZ'V^{-1}XC_{11}X^{R'} - GZ'V^{-1}ZGZ'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{+ GZ'V^{-1}ZGZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'V^{-1}ZGZ'\Upsilon'} \\
&\quad \boldsymbol{+ GZ'V^{-1}XC_{11}X'V^{-1}ZGZ'V^{-1}XC_{11}X^{R'} + GZ'V^{-1}XC_{11}X'V^{-1}ZGZ'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'V^{-1}ZGZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'\Upsilon' + GZ'V^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{+ GZ'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v' + GZ'V^{-1}R\Upsilon'} \\
&\quad \boldsymbol{- GZ'V^{-1}RV^{-1}XC_{11}X^{R'} - GZ'V^{-1}RV^{-1}ZGQ_v' + GZ'V^{-1}RV^{-1}XC_{11}X'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- \ GZ'V^{-1}XC_{11}X'V^{-1}R\Upsilon' + GZ'V^{-1}XC_{11}X'V^{-1}RV^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{+ GZ'V^{-1}XC_{11}X'V^{-1}RV^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'V^{-1}RV^{-1}XC_{11}X'V^{-1}ZGQ_v'} \\
&= \boldsymbol{GZ'V^{-1}(ZGZ' + R)\Upsilon' - GZ'V^{-1}(ZGZ' + R)V^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{- GZ'V^{-1}(ZGZ' + R)V^{-1}ZGQ_v' + GZ'V^{-1}(ZGZ' + R)V^{-1}XC_{11}X'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'V^{-1}(ZGZ' + R)\Upsilon'} \\
&\quad \boldsymbol{+ GZ'V^{-1}XC_{11}X'V^{-1}(ZGZ' + R)V^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{+ GZ'V^{-1}XC_{11}X'V^{-1}(ZGZ' + R)V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'V^{-1}(ZGZ' + R)V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'\Upsilon'} \\
&\quad \boldsymbol{+ GZ'V^{-1}XC_{11}X^{R'} + GZ'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v'}
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{GZ'\Upsilon' - GZ'V^{-1}XC_{11}X^{R'} - GZ'V^{-1}ZGQ_v' + GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'\Upsilon' + GZ'V^{-1}XC_{11}X'V^{-1}XC_{11}X^{R'} + GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v'} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'\Upsilon' + GZ'V^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{+ GZ'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v'}
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'\Upsilon' + GZ'V^{-1}XC_{11}C_{11}^{-1}C_{11}X^{R'}} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}C_{11}^{-1}C_{11}X'V^{-1}ZGQ_v'}
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v' - GZ'V^{-1}XC_{11}X'\Upsilon' + GZ'V^{-1}XC_{11}X^{R'}} \\
&\quad \boldsymbol{- GZ'V^{-1}XC_{11}X'V^{-1}ZGQ_v'}
\end{aligned}
$$

$$
= \boldsymbol{-GZ'V^{-1}XC_{11}(\Upsilon X - X^R)'}.
$$

Portanto,

$$
\boldsymbol{\zeta_2} = E\left\{\begin{bmatrix}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})' \\ (\tilde{\boldsymbol{v}} - \boldsymbol{v})(\boldsymbol{q} - \boldsymbol{X}^R\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}})'\end{bmatrix}\right\}\boldsymbol{A'} = \begin{bmatrix}\mathbf{C}_{11}(\Upsilon X - X^R)' \\ -\boldsymbol{GZ'V^{-1}XC}_{11}(\Upsilon X - X^R)'\end{bmatrix}\boldsymbol{A'}
$$

$$
= \begin{bmatrix}\mathbf{C}_{11} \\ -\boldsymbol{GZ'V^{-1}XC}_{11}\end{bmatrix}(\Upsilon X - X^R)'\boldsymbol{A'} = \begin{bmatrix}\mathbf{C}_{11} \\ \mathbf{C}_{21}\end{bmatrix}(\Upsilon X - X^R)'\boldsymbol{A'}
$$

$$
= \boldsymbol{MA'},
$$

onde $\boldsymbol{M} = \mathbf{C}_{.1}(\Upsilon X - X^R)'$ com $\mathbf{C}_{.1} = [\mathbf{C'}_{11} \quad \mathbf{C'}_{21}]'$. A terceira parcela da covariância em estudo é a transposta da segunda parcela, pelo que $\boldsymbol{\zeta_3} = \boldsymbol{AM'}$.

Por último, a quarta parcela dessa covariância é dada por:

$$
\boldsymbol{\zeta_4} = E\left[\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\boldsymbol{A'}\right] = \boldsymbol{A}E\left[(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}})'\right]\boldsymbol{A'}\ .
$$

E utilizando os resultados obtidos no cálculo de $\boldsymbol{\zeta_2}$, relativamente a $\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}$, tem-se que:

$$
\begin{aligned}
\boldsymbol{\zeta_4} &= \boldsymbol{A}E\{[(\Upsilon X - X^R)\boldsymbol{\beta}} \\
&\quad \boldsymbol{+ (\Upsilon Z - X^R C_{11}X'V^{-1}Z - Q_v GZ'V^{-1}Z + Q_v GZ'V^{-1}XC_{11}X'V^{-1}Z)v} \\
&\quad \boldsymbol{+ (\Upsilon - X^R C_{11}X'V^{-1} - Q_v GZ'V^{-1} + Q_v GZ'V^{-1}XC_{11}X'V^{-1})\varepsilon][(\Upsilon X - X^R)\beta} \\
&\quad \boldsymbol{+ (\Upsilon Z - X^R C_{11}X'V^{-1}Z - Q_v GZ'V^{-1}Z + Q_v GZ'V^{-1}XC_{11}X'V^{-1}Z)v} \\
&\quad \boldsymbol{+ (\Upsilon - X^R C_{11}X'V^{-1} - Q_v GZ'V^{-1} + Q_v GZ'V^{-1}XC_{11}X'V^{-1})\varepsilon]'\}A'}
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\{E[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)'] \\
&\quad + E[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)\boldsymbol{\beta v}'(\boldsymbol{\Upsilon Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} \\
&\quad + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'] \\
&\quad + E[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)\boldsymbol{\beta\varepsilon}'(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})'] \\
&\quad + E[(\boldsymbol{\Upsilon Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v\beta}'(\boldsymbol{\Upsilon X} \\
&\quad - \boldsymbol{X}^R)'] \\
&\quad + E[(\boldsymbol{\Upsilon Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{vv}'(\boldsymbol{\Upsilon Z} \\
&\quad - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'] \\
&\quad + E[(\boldsymbol{\Upsilon Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v\varepsilon}'(\boldsymbol{\Upsilon} \\
&\quad - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})'] \\
&\quad + E[(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)'] \\
&\quad + E[(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon v}'(\boldsymbol{\Upsilon Z} \\
&\quad - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})'] \\
&\quad + E[(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon\varepsilon}'(\boldsymbol{\Upsilon} \\
&\quad - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})']\}\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^R)' \\
&\quad + (\boldsymbol{\Upsilon Z} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{G}(\boldsymbol{\Upsilon Z} \\
&\quad - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})' \\
&\quad + (\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{R}(\boldsymbol{\Upsilon} - \boldsymbol{X}^R\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1} \\
&\quad - \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})']\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon ZGZ}'\boldsymbol{\Upsilon}' - \boldsymbol{\Upsilon ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{\Upsilon ZGZ}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{\Upsilon ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\boldsymbol{\Upsilon}' \\
&\quad + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\boldsymbol{\Upsilon}' \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\boldsymbol{\Upsilon}' \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{\Upsilon R\Upsilon}' - \boldsymbol{\Upsilon R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{\Upsilon R}\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{\Upsilon R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R\Upsilon}' \\
&\quad + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{R\Upsilon}' \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R\Upsilon}' \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{R}\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon}\big]\,\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon}(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{\Upsilon}' - \boldsymbol{\Upsilon}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{\Upsilon}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} + \boldsymbol{\Upsilon}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{\Upsilon}' + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{\Upsilon}' \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{\Upsilon}' \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}(\boldsymbol{ZGZ}' + \boldsymbol{R})\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{ZGQ}'_{\upsilon}\big]\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon V \Upsilon}' - \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{\Upsilon Z G Q}'_{\upsilon} \\
&\quad + \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\boldsymbol{\Upsilon}' \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon}\big]\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon V \Upsilon}' - \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{\Upsilon Z G Q}'_{\upsilon} \\
&\quad + \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{X}^{R}\mathbf{C}_{11}\mathbf{C}_{11}^{-1}\mathbf{C}_{11}\boldsymbol{X}^{R'} + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\mathbf{C}_{11}^{-1}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\boldsymbol{\Upsilon}' + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\mathbf{C}_{11}^{-1}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\mathbf{C}_{11}^{-1}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon}\big]\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon V \Upsilon}' - \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{\Upsilon Z G Q}'_{\upsilon} \\
&\quad + \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\boldsymbol{\Upsilon}' + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}'\big]\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + \boldsymbol{\Upsilon V \Upsilon}' - \boldsymbol{\Upsilon X}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{\Upsilon Z G Q}'_{\upsilon} - \boldsymbol{\Upsilon X}\mathbf{C}'_{21}\boldsymbol{Q}'_{\upsilon} \\
&\quad - \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{X}^{R}\mathbf{C}_{11}\boldsymbol{X}^{R'} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\boldsymbol{\Upsilon}' + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} \\
&\quad - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon} - \boldsymbol{Q}_{\upsilon}\mathbf{C}_{21}\boldsymbol{X}'\boldsymbol{\Upsilon}'\big]\boldsymbol{A}'
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{A}\big[(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + (\boldsymbol{\Upsilon V} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}')\boldsymbol{\Upsilon}' - (\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\mathbf{C}_{11}\boldsymbol{X}^{R'} \\
&\quad - \boldsymbol{\Upsilon}(\boldsymbol{Z G} + \boldsymbol{X}\mathbf{C}'_{21})\boldsymbol{Q}'_{\upsilon} - (\boldsymbol{X}^{R}\mathbf{C}_{11} + \boldsymbol{Q}_{\upsilon}\mathbf{C}_{21})\boldsymbol{X}'\boldsymbol{\Upsilon}' \\
&\quad + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'(\boldsymbol{I} - \mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}')\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon}\big]\boldsymbol{A}'
\end{aligned}
$$

$$= \boldsymbol{A N A}' ,$$

onde

$$
\begin{aligned}
\boldsymbol{N} = {} & (\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\boldsymbol{\beta\beta}'(\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})' + (\boldsymbol{\Upsilon V} - \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}')\boldsymbol{\Upsilon}' - (\boldsymbol{\Upsilon X} - \boldsymbol{X}^{R})\mathbf{C}_{11}\boldsymbol{X}^{R'} - \\
& \boldsymbol{\Upsilon}(\boldsymbol{Z G} + \boldsymbol{X}\mathbf{C}'_{21})\boldsymbol{Q}'_{\upsilon} - (\boldsymbol{X}^{R}\mathbf{C}_{11} + \boldsymbol{Q}_{\upsilon}\mathbf{C}_{21})\boldsymbol{X}'\boldsymbol{\Upsilon}' + \boldsymbol{Q}_{\upsilon}\boldsymbol{G Z}'(\boldsymbol{I} - \mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}')\mathbf{V}^{-1}\boldsymbol{Z G Q}'_{\upsilon}.
\end{aligned}
$$

Somando as expressões $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_3$ e $\boldsymbol{\zeta}_4$, tem-se finalmente que:

$V\left(\tilde{\xi}^R - \xi\right) = \boldsymbol{C} + \boldsymbol{MA}' + \boldsymbol{AM}' + \boldsymbol{ANA}'.$ ■

Lema 6.3: *Quando é utilizado um estimador sintético pela regressão ao nível da região, então a covariância no modelo do erro de predição* $\left(\tilde{\xi}^R - \xi\right)$ *é dada por:*

$$V\left(\tilde{\xi}^R - \xi\right) = \boldsymbol{C} + \boldsymbol{ANA}', \tag{6.3.12}$$

onde $\boldsymbol{N} = \boldsymbol{Q}_v\boldsymbol{GZ}'(\boldsymbol{I} - \mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}')\mathbf{V}^{-1}\boldsymbol{ZGQ}'_v$.

Demonstração: A demonstração deste lema é apresentada de forma resumida pelo facto de ser semelhante à do lema 6.2, sendo apenas diferente o estimador utilizado para estimação ao nível das regiões. Assim, tem-se que o vector $\boldsymbol{q}$, utilizado nas parcelas $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_3$ e $\boldsymbol{\zeta}_4$ da variância no modelo do erro de predição de $\boldsymbol{\xi}$, é agora dado por:

$$\boldsymbol{q} = \boldsymbol{X}^{\boldsymbol{R}}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}^{\boldsymbol{R}}\boldsymbol{\beta} + \boldsymbol{X}^{\boldsymbol{R}}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Zv} + \boldsymbol{X}^{\boldsymbol{R}}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}.$$

Desta forma, a expressão $\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$ simplifica-se para:

$$\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}} = \boldsymbol{X}^{\boldsymbol{R}}\boldsymbol{\beta} - \boldsymbol{X}^{\boldsymbol{R}}\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}} =$$

$$= (-\boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{Z} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{Z})\boldsymbol{v} + (-\boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1} + \boldsymbol{Q}_v\boldsymbol{GZ}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}'\mathbf{V}^{-1})\boldsymbol{\varepsilon}\ .$$

Como consequência tem-se que $\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}_3 = \mathbf{0}$ e que

$$\boldsymbol{\zeta}_4 = \boldsymbol{AQ}_v\boldsymbol{GZ}'(\boldsymbol{I} - \mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}\boldsymbol{X}')\mathbf{V}^{-1}\boldsymbol{ZGQ}'_v\boldsymbol{A}' = \boldsymbol{ANA}'\ .$$

De forma sintética, mostrou-se que a covariância no modelo do erro de predição $\left(\tilde{\xi}^R - \xi\right)$ quando é utilizado um estimador sintético pela regressão ao nível da região simplifica-se para:

$$V\left(\tilde{\xi}^R - \xi\right) = \boldsymbol{C} + \boldsymbol{ANA}'.$$

Observação: Um resultado equivalente ao do lema 6.3 para o caso de modelos de nível unidade pode ser encontrado em Ugarte *et al*. (2009).

6.3.2 O BLUP com restrições

6.3.2.1 Modelo geral com *A* restrições

O preditor da média no *i*-ésimo pequeno domínio modificado pelas *A* restrições é, então, dado por:

$$\tilde{\theta}_i^R = \boldsymbol{k'}_i\tilde{\boldsymbol{\beta}}^R + \boldsymbol{m'}_i\tilde{\boldsymbol{v}}^R = \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}}^R = \tag{6.3.13}$$

$$= \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} + \boldsymbol{l'}_i\boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \tilde{\theta}_i + [\boldsymbol{k'}_i\ \boldsymbol{m'}_i]\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C'}_{21}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}^{R\,\prime}\\ \boldsymbol{Q}_v{}'\end{bmatrix}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \tilde{\theta}_i + [\boldsymbol{k'}_i\ \boldsymbol{m'}_i]\begin{bmatrix}\mathbf{C}_{11}\boldsymbol{X}^{R\,\prime} + \mathbf{C'}_{21}\boldsymbol{Q}_v{}'\\ \mathbf{C}_{21}\boldsymbol{X}^{R\,\prime} + \mathbf{C}_{22}\boldsymbol{Q}_v{}'\end{bmatrix}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \tilde{\theta}_i + \left(\boldsymbol{k'}_i\mathbf{C}_{11}\boldsymbol{X}^{R\,\prime} + \boldsymbol{k'}_i\mathbf{C'}_{21}\boldsymbol{Q}_v' + \boldsymbol{m'}_i\mathbf{C}_{21}\boldsymbol{X}^{R\,\prime} + \boldsymbol{m'}_i\mathbf{C}_{22}\boldsymbol{Q}_v{}'\right)(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

onde $\boldsymbol{l'}_i = [\boldsymbol{k'}_i\ \ \boldsymbol{m'}_i]$. O preditor (6.3.13) satisfaz a exigência da garantia da consistência interna, isto é, a agregação das estimativas produzidas para os pequenos domínios de uma dada região com base neste preditor iguala a estimativa do parâmetro de interesse nessa região. É, ainda, de notar que o preditor (6.3.13) é dado pela soma de duas parcelas. Uma dessas parcelas é o preditor sem restrições. A outra parcela, que pode ser entendida como um factor de correcção, é definida como uma proporção das diferenças entre as estimativas do parâmetro de interesse ao nível de uma dada região e a média ponderada das estimativas produzidas para os pequenos domínios dessa região. Desta forma, quando se assume a existência de *A* restrições, esse factor de correcção do preditor associado a um domínio particular, pode não só ter em conta os erros de predição da região em que está contido, mas também de outras regiões.

Note-se, ainda, que o preditor (6.3.13) também pode ser dado por:

$$\tilde{\theta}_i^R = \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} + \boldsymbol{l'}_i\boldsymbol{CQ'}(\boldsymbol{QCQ'})^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) =$$

$$= [\boldsymbol{k'}_i \quad \boldsymbol{m'}_i]\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{v}}\end{bmatrix} + \boldsymbol{l'}_i\boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \boldsymbol{k'}_i\tilde{\boldsymbol{\beta}} + \boldsymbol{m'}_i\tilde{\boldsymbol{v}} + \boldsymbol{l'}_i\boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \boldsymbol{k'}_i\tilde{\boldsymbol{\beta}} + \boldsymbol{m'}_i\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right) + \boldsymbol{l'}_i\boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

$$= \boldsymbol{k'}_i\tilde{\boldsymbol{\beta}} + \boldsymbol{b}'\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right) + \boldsymbol{l'}_i\boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right).$$

Lema 6.4: *O preditor $\tilde{\theta}_i^R$ é o BLUP que satisfaz as restrições (6.3.6).*

Demonstração: Notando-se que:

$$\boldsymbol{Q}\tilde{\boldsymbol{\xi}} = \boldsymbol{X}^R\,\tilde{\boldsymbol{\beta}} + \boldsymbol{Q}_v\tilde{\boldsymbol{v}} = col_{1\le a\le A}\left(\boldsymbol{\delta'}_a\boldsymbol{\Psi}\boldsymbol{X}\,\tilde{\boldsymbol{\beta}}\right) + col_{1\le a\le A}\left(\boldsymbol{\delta'}_a\,\boldsymbol{\Psi}\tilde{\boldsymbol{v}}\right) =$$

$$= col_{1\le a\le A}\left(\sum_{i\in a}\omega_i\,\boldsymbol{x'}_i\tilde{\boldsymbol{\beta}} + \sum_{i\in a}\omega_i\,\tilde{v}_i\right) = col_{1\le a\le A}\left(\tilde{\theta}_a\right) = \tilde{\boldsymbol{\Theta}}$$

é o vector dos BLUP dos parâmetros de interesse ao nível da região, $\theta_a,\ a = 1, \dots, A$, sob o modelo (6.3.1), ou seja, sob o modelo sem restrições. Notando-se também que a covariância entre os BLUP dos parâmetros de interesse ao nível do pequeno domínio, $\tilde{\theta}_i$, e ao nível da região, $\tilde{\theta}_a$, é dada por:

$$Cov\left(\tilde{\theta}_i; \tilde{\boldsymbol{\Theta}}\right) = E\left[\left(\tilde{\theta}_i - \theta_i\right)\left(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right)'\right] = E\left[\left(\boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} - \boldsymbol{l'}_i\boldsymbol{\xi}\right)\left(\boldsymbol{Q}\tilde{\boldsymbol{\xi}} - \boldsymbol{Q}\boldsymbol{\xi}\right)'\right]$$

$$= E\left[\boldsymbol{l'}_i\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)'\boldsymbol{Q}'\right]$$

$$= \boldsymbol{l'}_i E\left[\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)'\right]\boldsymbol{Q}' = \boldsymbol{l'}_i\boldsymbol{C}\,\boldsymbol{Q}'.$$

Observando-se, ainda, que covariância do erro de predição dos parâmetros de interesse ao nível da região, $\tilde{\theta}_a$, é dado por:

$$V\left(\tilde{\boldsymbol{\Theta}}\right) = E\left[\left(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right)\left(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right)'\right] = E\left[\left(\boldsymbol{Q}\tilde{\boldsymbol{\xi}} - \boldsymbol{Q}\boldsymbol{\xi}\right)\left(\boldsymbol{Q}\tilde{\boldsymbol{\xi}} - \boldsymbol{Q}\boldsymbol{\xi}\right)'\right]$$

$$= E\left[\boldsymbol{Q}\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)'\boldsymbol{Q}'\right]$$

$$= \boldsymbol{Q}E\left[\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)'\right]\boldsymbol{Q}' = \boldsymbol{QCQ'}.$$

Verifica-se, então, que:

$$\tilde{\theta}_i^R = \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} + \boldsymbol{l'}_i\boldsymbol{C}\boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}')^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) =$$

$$= \tilde{\theta}_i \ + Cov\left(\tilde{\theta}_i; \widetilde{\boldsymbol{\Theta}}\right) \times \left[V\left(\widetilde{\boldsymbol{\Theta}}\right)\right]^{-1} \times \left(\boldsymbol{q} - \widetilde{\boldsymbol{\Theta}}\right).$$

Para verificar que $\tilde{\theta}_i^R$, dado pela expressão (6.3.13), é o BLUP de $\tilde{\theta}_i$ basta apenas confirmar que verifica a condição (6.2.5) devida a Wang *et al*. (2008). Considerando-se que $\varphi_i = \omega_i\left[Cov\left(\tilde{\theta}_i; \widetilde{\boldsymbol{\Theta}}\right)\right]^{-1}$, notando-se que $\tilde{\theta}_a = \sum_{i\in a} \omega_i \, \tilde{\theta}_i$ e generalizando-se o resultado de Wang *et al.* (2008) para um caso com *A* restrições como o que tem vindo a ser apresentado, tem-se:

$$\breve{\alpha}_i = \varphi_i^{-1}\omega_i \left[col_{1\le a\le A}\left(\sum_{i\in a} \varphi_i^{-1}\omega_i^2\right)\right]^{-1}$$

$$= Cov\left(\tilde{\theta}_i; \widetilde{\boldsymbol{\Theta}}\right)\left\{col_{1\le a\le A}\left[\sum_{i\in a} \omega_i Cov\left(\tilde{\theta}_i; \widetilde{\boldsymbol{\Theta}}\right)\right]\right\}^{-1}$$

$$= Cov\left(\tilde{\theta}_i; \widetilde{\boldsymbol{\Theta}}\right)\left[V\left(\widetilde{\boldsymbol{\Theta}}\right)\right]^{-1}.$$

Os dois últimos resultados provam o lema 6.4. ■

6.3.2.2 Modelo geral com uma restrição

Quando se considera apenas uma única restrição no processo de estimação (por exemplo, quando se pretende que a soma ponderada das estimativas combinadas para pequenos domínios, $\tilde{\theta}_i^R$, iguale a respectiva estimativa obtida para a população total, $\hat{\theta}$),

$$\sum_{i=1}^{m} \omega_i \tilde{\theta}_i^R = \hat{\theta}\,, \tag{6.3.14}$$

onde ω_i são ponderadores tal que $\sum_{i=1}^{m} \omega_i = 1$, então o estimador da média no *i*-ésimo pequeno domínio modificado pela restrição é, então, dado por:

$$\tilde{\theta}_i^R = \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} + \boldsymbol{l'}_i\boldsymbol{C}\boldsymbol{Q}'(\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}')^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) \tag{6.3.15}$$

$$= \tilde{\theta}_i \ + [\boldsymbol{k'}_i \, \boldsymbol{m'}_i]\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C'}_{21}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}^{R\,\prime}\\ \boldsymbol{Q}_v{}'\end{bmatrix}(\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}')^{-1}\left(\hat{\theta} - \boldsymbol{X}^R\,\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\tilde{\boldsymbol{v}}\right)$$

$$\begin{aligned}= \tilde{\theta}_i \ &+ [\boldsymbol{x'}_i\mathbf{C}_{11}(\textstyle\sum_{i=1}^{m}\omega_i\boldsymbol{x}_i) + \boldsymbol{x'}_i\mathbf{C'}_{21}(\sum_{i=1}^{m}\omega_i\mathbf{1}_m) + \boldsymbol{m'}_i\mathbf{C}_{21}(\sum_{i=1}^{m}\omega_i\boldsymbol{x}_i) +\\ &\boldsymbol{m'}_i\mathbf{C}_{22}(\textstyle\sum_{i=1}^{m}\omega_i\mathbf{1}_m)][(\sum_{i=1}^{m}\omega_i\boldsymbol{x'}_i)\mathbf{C}_{11}(\sum_{i=1}^{m}\omega_i\boldsymbol{x}_i) +\\ &(\textstyle\sum_{i=1}^{m}\omega_i\boldsymbol{x'}_i)\mathbf{C'}_{21}(\sum_{i=1}^{m}\omega_i\mathbf{1}_m) + (\sum_{i=1}^{m}\omega_i\mathbf{1'}_m)\mathbf{C}_{21}(\sum_{i=1}^{m}\omega_i\boldsymbol{x}_i) +\\ &(\textstyle\sum_{i=1}^{m}\omega_i\mathbf{1'}_m)\mathbf{C}_{22}(\sum_{i=1}^{m}\omega_i\mathbf{1}_m)]^{-1}\left(\hat{\theta} - \sum_{i=1}^{m}\omega_i\boldsymbol{x'}_i\,\tilde{\boldsymbol{\beta}} - \sum_{i=1}^{m}\omega_i\tilde{v}_i\right).\end{aligned}$$

Note-se que, neste caso particular, se tem $\boldsymbol{\delta'}_a = \mathbf{1}_m$ (1×m), $\boldsymbol{q} = \theta$ (1×1), $\boldsymbol{X}^R = \sum_{i=1}^m \omega_i \boldsymbol{x'}_i$ (1×p), $\boldsymbol{Q}_v = \sum_{i=1}^m \omega_i \mathbf{1'}_m$ (m×m) e $\widetilde{\boldsymbol{\Theta}} = \tilde{\theta}$ (1×1). Relembre-se, ainda, que $\boldsymbol{k'}_i = \boldsymbol{x'}_i$, $\boldsymbol{m'}_i$ é a i-ésima linha de $\boldsymbol{I}_m$ e $\boldsymbol{C} = (\boldsymbol{X^{0\prime}}\mathbf{V}^{-1}\boldsymbol{X^0})^{-1}$. É, ainda, de notar que o preditor (6.3.15) é dado pela soma de duas parcelas. Uma dessas parcelas é o preditor sem restrições. A outra parcela é definida como uma proporção da diferença entre a estimativa do parâmetro de interesse ao nível de uma dada região e a média ponderada das estimativas produzidas para os pequenos domínios dessa região.

Naturalmente que o estimador (6.3.15) é igualmente BLUP uma vez que pode igualmente ser apresentado na seguinte forma:

$$\tilde{\theta}_i^R = \tilde{\theta}_i \ + Cov\left(\tilde{\theta}_i; \tilde{\theta}\right) \times \left[V\left(\tilde{\theta}\right)\right]^{-1} \times \left(\hat{\theta} - \tilde{\theta}\right),$$

onde, como foi visto acima:

$$Cov\left(\tilde{\theta}_i; \tilde{\theta}\right) = \boldsymbol{l'}_i \boldsymbol{C}\, \boldsymbol{Q'} =$$

$$= \boldsymbol{x'}_i \mathbf{C}_{11} \left(\sum_{i=1}^m \omega_i \boldsymbol{x}_i\right) + \boldsymbol{x'}_i \mathbf{C'}_{21} \left(\sum_{i=1}^m \omega_i \mathbf{1}_m\right) +$$

$$+ \boldsymbol{m'}_i \mathbf{C}_{21} \left(\sum_{i=1}^m \omega_i \boldsymbol{x}_i\right) + \boldsymbol{m'}_i \mathbf{C}_{22} \left(\sum_{i=1}^m \omega_i \mathbf{1}_m\right)$$

e

$$V\left(\tilde{\theta}\right) = \boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q'} =$$

$$= \left(\sum_{i=1}^m \omega_i \boldsymbol{x'}_i\right) \mathbf{C}_{11} \left(\sum_{i=1}^m \omega_i \boldsymbol{x}_i\right) + \left(\sum_{i=1}^m \omega_i \boldsymbol{x'}_i\right) \mathbf{C'}_{21} \left(\sum_{i=1}^m \omega_i \mathbf{1}_m\right) +$$

$$+ \left(\sum_{i=1}^m \omega_i \mathbf{1'}_m\right) \mathbf{C}_{21} \left(\sum_{i=1}^m \omega_i \boldsymbol{x}_i\right) + \left(\sum_{i=1}^m \omega_i \mathbf{1'}_n\right) \mathbf{C}_{22} \left(\sum_{i=1}^m \omega_i \mathbf{1}_m\right).$$

6.3.3 O EBLUP com restrições

A derivação do preditor (6.3.13) assume que as componentes de variância são conhecidas. Contudo, tal como foi referido anteriormente, esta situação raramente acontece quando se trabalha com problemas reais. Desta forma, é necessário substituir os parâmetros de variância desconhecidos pelas respectivas estimativas, obtendo-se desta forma o preditor empírico (EBLUP) do parâmetro de interesse:

$$\hat{\theta}_i^R = \boldsymbol{k'}_i\hat{\boldsymbol{\beta}}^R + \boldsymbol{m'}_i\hat{\boldsymbol{v}}^R = \boldsymbol{l'}_i\hat{\boldsymbol{\xi}}^R = \tag{6.3.16}$$

$$= \hat{\theta}_i + \left(\boldsymbol{k'}_i\hat{\mathbf{C}}_{11}\boldsymbol{X}^{R\,\prime} + \boldsymbol{k'}_i\hat{\mathbf{C}}'_{21}\boldsymbol{Q}'_v + \boldsymbol{m'}_i\hat{\mathbf{C}}_{21}\boldsymbol{X}^{R\,\prime} + \boldsymbol{m'}_i\hat{\mathbf{C}}_{22}\boldsymbol{Q}_v{}'\right)\left(\boldsymbol{Q}\hat{\boldsymbol{C}}\boldsymbol{Q}'\right)^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\hat{\boldsymbol{\xi}}\right).$$

6.3.4 O EQMP do BLUP com restrições

Tal como foi referido na secção 3.2.5, o EQMP de um BLUP avalia a variabilidade devida à estimação dos efeitos aleatórios e dos efeitos fixos. À semelhança do caso sem restrições, para se deduzir o EQMP do BLUP com restrições é conveniente exprimir o BLUP em função do respectivo melhor preditor (BP).

Lema 6.5: *O BLUP (6.3.13) pode ser escrito como* $\tilde{\theta}_i^R = \breve{\theta}_i^R + \boldsymbol{f'}_i\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$*, onde* $\breve{\theta}_i^R$ *é o melhor preditor modificado pelas restrições de* θ_i *e* $\boldsymbol{f'}_i = \boldsymbol{d'}_i + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)$.

Demonstração:

$$\tilde{\theta}_i^R = \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}}^R + \breve{\theta}_i^R - \breve{\theta}_i^R =$$

$$= \boldsymbol{l'}_i\tilde{\boldsymbol{\xi}} + \boldsymbol{l'}_i\boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) + \boldsymbol{l'}_i\boldsymbol{\xi} + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\xi}) - \boldsymbol{l'}_i\boldsymbol{\xi} - \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\xi})$$

$$= \boldsymbol{l'}_i\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) + \boldsymbol{l'}_i\boldsymbol{\xi} + \boldsymbol{l'}_i\boldsymbol{A}\left[\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) + (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\xi}) - (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\xi})\right].$$

Desenvolvendo, em primeiro lugar, as duas primeiras parcelas da expressão anterior, obtém-se o BLUP de θ_i do caso sem restrições:

$$\boldsymbol{l'}_i\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) + \boldsymbol{l'}_i\boldsymbol{\xi} = [\boldsymbol{k'}_i\ \boldsymbol{m'}_i]\begin{bmatrix}\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{v}} - \boldsymbol{v}\end{bmatrix} + [\boldsymbol{k'}_i\ \boldsymbol{m'}_i]\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{v}\end{bmatrix} =$$

$$= \boldsymbol{k'}_i\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{m'}_i(\tilde{\boldsymbol{v}} - \boldsymbol{v}) + \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{m'}_i\boldsymbol{v}'$$

$$= \boldsymbol{k'}_i\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{m'}_i\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right) - \boldsymbol{m'}_i\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{k'}_i\boldsymbol{\beta}$$
$$+ \boldsymbol{m'}_i\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

E definindo $\boldsymbol{b'}_i = \boldsymbol{m'}_i\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}$ e $\boldsymbol{d'}_i = \boldsymbol{k'}_i - \boldsymbol{b'}_i\boldsymbol{X}$, tem-se:

$$\boldsymbol{l'}_i\left(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) + \boldsymbol{l'}_i\boldsymbol{\xi} =$$
$$= \boldsymbol{k'}_i\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{b'}_i\boldsymbol{y} - \boldsymbol{b'}_i\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{b'}_i\boldsymbol{y} + \boldsymbol{b'}_i\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$= \boldsymbol{k'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{b'}_i\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta})$$

$$= \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + (\boldsymbol{k'}_i - \boldsymbol{b'}_i\boldsymbol{X})\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

$$= \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{d'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),$$

que é o BLUP de θ_i, conforme apresentado em Rao (2003, p. 98). Notando que $\boldsymbol{q} - \boldsymbol{Q\xi} = \boldsymbol{q} - [\boldsymbol{X}^R\ \boldsymbol{Q}_v]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{v}\end{bmatrix} = \boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta} - \boldsymbol{Q}_v\boldsymbol{v}$ e que $\widetilde{\boldsymbol{v}} = \boldsymbol{GZ'V^{-1}}\left(\mathbf{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right)$, tem-se que:

$$\begin{aligned}
&\boldsymbol{l'}_i\boldsymbol{A}\left[\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right) + (\boldsymbol{q} - \boldsymbol{Q\xi}) - (\boldsymbol{q} - \boldsymbol{Q\xi})\right] \\
&\quad = \boldsymbol{l'}_i\boldsymbol{A}\left[\boldsymbol{q} - \boldsymbol{X}^R\widetilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right) - (\boldsymbol{q} - \boldsymbol{Q\xi}) + (\boldsymbol{q} - \boldsymbol{Q\xi})\right] \\
&\quad = \boldsymbol{l'}_i\boldsymbol{A}\left[\boldsymbol{q} - \boldsymbol{X}^R\widetilde{\boldsymbol{\beta}} - \boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{y} + \boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \boldsymbol{q} + \boldsymbol{X}^R\boldsymbol{\beta}\right. \\
&\qquad \left. + \boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) + (\boldsymbol{q} - \boldsymbol{Q\xi})\right] \\
&\quad = \boldsymbol{l'}_i\boldsymbol{A}\left[-\boldsymbol{X}^R\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + (\boldsymbol{q} - \boldsymbol{Q\xi})\right] \\
&\quad = \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).
\end{aligned}$$

Somando as duas expressões, tem-se finalmente que o BLUP (6.3.13) é dado por:

$$\begin{aligned}
\tilde{\theta}_i^R &= \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{d'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi}) \\
&\quad + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&= \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi}) + [\boldsymbol{d'}_i + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)]\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&= \breve{\theta}_i^R + \boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)
\end{aligned}$$

onde $\breve{\theta}_i^R = \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi})$ é o melhor preditor modificado pelas restrições de $\theta_i = \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{m'}_i\boldsymbol{v}$ e $\boldsymbol{f'}_i = \boldsymbol{d'}_i + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)$. ■

O EQMP do BLUP com restrições pode ser, então, deduzido com base nos resultados do Lema 6.5. Notando que $E\left[\boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] = \boldsymbol{f'}_iE\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \mathbf{0}$, tem-se que:

$$EQMP\left(\tilde{\theta}_i^R\right) = EQMP\left[\breve{\theta}_i^R + \boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]$$

$$= EQMP\left(\breve{\theta}_i^R\right) + V\left[\boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] + 2Cov\left[\breve{\theta}_i^R; \boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]. \qquad (6.3.17)$$

Seguidamente irá calcular-se separadamente cada um destes termos. Antes de se calcular $EQMP\left(\breve{\theta}_i^R\right) = E\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\breve{\theta}_i^R - \theta_i\right)'\right]$, a qual é a primeira parcela da expressão (6.3.17), vai reescrever-se $\breve{\theta}_i^R - \theta_i$ como:

$$\breve{\theta}_i^R - \theta_i = \boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi}) - \boldsymbol{k'}_i\boldsymbol{\beta} - \boldsymbol{m'}_i\boldsymbol{v} =$$

$$= \boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta} - \boldsymbol{Q}_v\boldsymbol{v}) - \boldsymbol{m'}_i\boldsymbol{v}$$

$$= \boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) + \boldsymbol{l'}_i\boldsymbol{Aq} - \boldsymbol{l'}_i\boldsymbol{AX}^R\boldsymbol{\beta} - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{v} - \boldsymbol{m'}_i\boldsymbol{v}$$

$$= \boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) + \boldsymbol{l'}_i\boldsymbol{Aq} - \boldsymbol{l'}_i\boldsymbol{AX}^R\boldsymbol{\beta} - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) - \boldsymbol{m'}_i\boldsymbol{v}$$

$$= \boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta}) - \boldsymbol{m'}_i\boldsymbol{v}$$

$$= (\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1})(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{m'}_i\boldsymbol{v} + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})$$

$$= \boldsymbol{c'}_{1i}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{m'}_i\boldsymbol{v} + \boldsymbol{c'}_{2i},$$

onde $\boldsymbol{c'}_{1i} = \boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}$ e $\boldsymbol{c'}_{2i} = \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})$ são constantes. Tem-se, então, que:

$$EQMP\left(\breve{\theta}_i^R\right) = E\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\breve{\theta}_i^R - \theta_i\right)'\right] =$$

$$= E\{[\boldsymbol{c'}_{1i}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{m'}_i\boldsymbol{v} + \boldsymbol{c'}_{2i}][\boldsymbol{c'}_{1i}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{m'}_i\boldsymbol{v} + \boldsymbol{c'}_{2i}]'\}$$

$$\begin{aligned} = {} & \boldsymbol{c'}_{1i}E[(\boldsymbol{Zv} + \boldsymbol{\varepsilon})(\boldsymbol{Zv} + \boldsymbol{\varepsilon})']\boldsymbol{c}_{1i} - \boldsymbol{c'}_{1i}E[(\boldsymbol{Zv} + \boldsymbol{\varepsilon})\boldsymbol{v}']\boldsymbol{m}_i \\ & + \boldsymbol{c'}_{1i}E[(\boldsymbol{Zv} + \boldsymbol{\varepsilon})]\boldsymbol{c}_{2i} - \boldsymbol{m'}_iE[\boldsymbol{v}(\boldsymbol{Zv} + \boldsymbol{\varepsilon})']\boldsymbol{c}_{1i} + \boldsymbol{m'}_iE[\boldsymbol{vv}']\boldsymbol{m}_i \\ & - \boldsymbol{m'}_iE[\boldsymbol{v}]\boldsymbol{c}_{2i} + \boldsymbol{c'}_{2i}E[(\boldsymbol{Zv} + \boldsymbol{\varepsilon})']\boldsymbol{c}_{1i} - \boldsymbol{c'}_{2i}E[\boldsymbol{v}']\boldsymbol{m}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}\,. \end{aligned}$$

E recordando novamente que os vectores aleatórios $\boldsymbol{v}$ e $\boldsymbol{\varepsilon}$ satisfazem $E(\boldsymbol{v}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $E(\boldsymbol{vv'}) = \boldsymbol{G}$, $E(\boldsymbol{\varepsilon\varepsilon'}) = \boldsymbol{R}$ e $E(\boldsymbol{v\varepsilon'}) = \boldsymbol{0}$, e que $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ e $\boldsymbol{b'}_i = \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}$, tem-se:

$$EQMP\left(\breve{\theta}_i^R\right) = E\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\breve{\theta}_i^R - \theta_i\right)'\right] =$$

$$\begin{aligned} = {} & \boldsymbol{c'}_1E[\boldsymbol{Zvv'Z'} + \boldsymbol{Zv\varepsilon'} + \boldsymbol{\varepsilon v'Z'} + \boldsymbol{\varepsilon\varepsilon'}]\boldsymbol{c}_{1i} - \boldsymbol{c'}_{1i}E[\boldsymbol{Zvv'} + \boldsymbol{\varepsilon v'}]\boldsymbol{m}_i \\ & - \boldsymbol{m'}_iE[\boldsymbol{vv'Z'} + \boldsymbol{v\varepsilon'}]\boldsymbol{c}_{1i} + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i} \end{aligned}$$

$$= \boldsymbol{c'}_{1i}(\boldsymbol{ZGZ'} + \boldsymbol{R})\boldsymbol{c}_{1i} - \boldsymbol{c'}_{1i}\boldsymbol{ZGm}_i - \boldsymbol{m'}_i\boldsymbol{GZ'c}_{1i} + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}$$

$$\begin{aligned}= {} & (\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1})\boldsymbol{V}(\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1})' \\ & - (\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1})\boldsymbol{ZGm}_i - \boldsymbol{m'}_i\boldsymbol{GZ'}(\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1})' \\ & + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}\end{aligned}$$

$$\begin{aligned}= {} & \boldsymbol{b'}_i\boldsymbol{Vb}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{Vb}_i - \boldsymbol{b'}_i\boldsymbol{VV}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i \\ & + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{VV}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i - \boldsymbol{b'}_i\boldsymbol{ZGm}_i + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i \\ & - \boldsymbol{m'}_i\boldsymbol{GZ'b}_i + \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}\end{aligned}$$

$$\begin{aligned}= {} & \boldsymbol{b'}_i\boldsymbol{Vb}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'b}_i - \boldsymbol{b'}_i\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i \\ & - \boldsymbol{b'}_i\boldsymbol{ZGm}_i + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i - \boldsymbol{m'}_i\boldsymbol{GZ'b}_i + \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i \\ & + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}\end{aligned}$$

$$\begin{aligned}= {} & \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{VV}^{-1}\boldsymbol{ZGm}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i - \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i \\ & + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i - \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i \\ & - \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i + \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}\end{aligned}$$

$$= \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i - \boldsymbol{m'}_i\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGm}_i + \boldsymbol{m'}_i\boldsymbol{Gm}_i + \boldsymbol{c'}_{2i}\boldsymbol{c}_{2i}$$

$$\begin{aligned}= {} & \boldsymbol{m'}_i(\boldsymbol{G} - \boldsymbol{GZ'V}^{-1}\boldsymbol{ZG})\boldsymbol{m}_i + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i \\ & + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})'\boldsymbol{A'l}_i\end{aligned}$$

$$= g_{1i}(\boldsymbol{\psi}) + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{ZGQ'}_v\boldsymbol{A'l}_i + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})'\boldsymbol{A'l}_i$$

uma vez que $g_{1i}(\boldsymbol{\psi}) = \boldsymbol{m'}_i(\boldsymbol{G} - \boldsymbol{GZ'V}^{-1}\boldsymbol{ZG})\boldsymbol{m}_i$, tal como definido em (3.2.11).

A segunda parcela da expressão (6.3.17) é dada por:

$$V[\boldsymbol{f'}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] = E\{[\boldsymbol{f'}_i(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{0}][\boldsymbol{f'}_i(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{0}]'\} =$$

$$= \boldsymbol{f'}_i E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})']\boldsymbol{f}_i = \boldsymbol{f'}_i V(\tilde{\boldsymbol{\beta}})\boldsymbol{f}_i =$$

$$\begin{aligned}= {} & [\boldsymbol{d'}_i + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)](\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}[\boldsymbol{d'}_i \\ & + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)]'\end{aligned}$$

$$= \boldsymbol{d'}_i(\boldsymbol{X'V^{-1}X})^{-1}\,\boldsymbol{d}_i + \boldsymbol{d'}_i(\boldsymbol{X'V^{-1}X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)'\boldsymbol{A'l}_i$$
$$+ \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{d}_i$$
$$+ \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)(\boldsymbol{X'V^{-1}X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)'\boldsymbol{A'l}_i$$

$$= g_{2i}(\boldsymbol{\psi}) + \boldsymbol{d'}_i(\boldsymbol{X'V^{-1}X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)'\boldsymbol{A'l}_i$$
$$+ \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{d}_i$$
$$+ \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)(\boldsymbol{X'V^{-1}X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ'V^{-1}X} - \boldsymbol{X}^R)'\boldsymbol{A'l}_i\,,$$

uma vez que $g_{2i}(\boldsymbol{\psi}) = \boldsymbol{d'}_i(\boldsymbol{X'V^{-1}X})^{-1}\,\boldsymbol{d}_i$**,** tal como definido em (3.2.12).

Notando-se que o melhor preditor é não enviesado[57] (McCulloch e Searle, 2001), tem-se que a terceira parcela da expressão (6.3.17) é dada por:

$$Cov\left[\breve{\theta}_i^R, \boldsymbol{f'}_i(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] = E\left\{\left(\breve{\theta}_i^R - \theta_i^R\right)\left[\boldsymbol{f'}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{0}\right]'\right\} =$$

$$= E\left\{\left[\boldsymbol{k'}_i\boldsymbol{\beta} + \boldsymbol{b'}_i(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi}) - \boldsymbol{k'}_i\boldsymbol{\beta} - \boldsymbol{m'}_i\boldsymbol{v} - \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{Q\xi})\right]\left(\widetilde{\boldsymbol{\beta}}\right.\right.$$
$$\left.\left. - \boldsymbol{\beta}\right)'\boldsymbol{f}_i\right\}$$

$$= E\left\{\left[\boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) + \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta} - \boldsymbol{Q}_v\boldsymbol{v}) - \boldsymbol{m'}_i\boldsymbol{v} - \boldsymbol{l'}_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta} - \boldsymbol{Q}_v\boldsymbol{v})\right]\left(\widetilde{\boldsymbol{\beta}}\right.\right.$$
$$\left.\left. - \boldsymbol{\beta}\right)'\boldsymbol{f}_i\right\}$$

$$= E\left\{\left[\boldsymbol{b'}_i(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V^{-1}}(\boldsymbol{y} - \boldsymbol{X\beta}) - \boldsymbol{m'}_i\boldsymbol{v} + \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{v}\right]\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{f}_i\right\}$$

$$= E\left\{\left[(\boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V^{-1}})(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - (\boldsymbol{m'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v)\boldsymbol{v}\right]\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{f}_i\right\}$$

$$= E\left\{\left[\boldsymbol{c'}_{1i}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{c'}_{3i}\boldsymbol{v}\right]\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{f}_i\right\}$$

onde $\boldsymbol{c'}_{1i} = \boldsymbol{b'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v\boldsymbol{GZ'V^{-1}}$ e $\boldsymbol{c'}_{3i} = \boldsymbol{m'}_i - \boldsymbol{l'}_i\boldsymbol{AQ}_v$. Note-se, também, que a expressão $\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{f}_i$ pode ser reescrita como:

$$\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{f}_i = \left[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}y} - \boldsymbol{\beta}\right]'\boldsymbol{f}_i =$$

$$= \left[(\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}}(\boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}\right]'\boldsymbol{f}_i =$$

$$= \left[\boldsymbol{\beta} + (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}\right]'\boldsymbol{f}_i =$$

[57] No âmbito da predição, o não enviesamento significa que o valor esperado do preditor é igual ao valor esperado da variável aleatória objecto de predição.

$$= (\boldsymbol{Zv} + \boldsymbol{\varepsilon})'[(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}]'\boldsymbol{f}_i$$

$$= (\boldsymbol{Zv} + \boldsymbol{\varepsilon})'\boldsymbol{h}_i$$

onde $\boldsymbol{h}_i = [(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}]'\boldsymbol{f}_i$ é constante. Portanto,

$$Cov\left[\breve{\theta}_i^R; \boldsymbol{f}'_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] = E\{[\boldsymbol{c}'_{1i}(\boldsymbol{Zv} + \boldsymbol{\varepsilon}) - \boldsymbol{c}'_{3i}\boldsymbol{v}](\boldsymbol{Zv} + \boldsymbol{\varepsilon})'\boldsymbol{h}_i\} =$$

$$= \boldsymbol{c}'_{1i}E[(\boldsymbol{Zv} + \boldsymbol{\varepsilon})(\boldsymbol{Zv} + \boldsymbol{\varepsilon})']\boldsymbol{h}_i - \boldsymbol{c}'_{3i}E[\boldsymbol{v}(\boldsymbol{Zv} + \boldsymbol{\varepsilon})']\boldsymbol{h}_i$$

$$= \boldsymbol{c}'_{1i}E(\boldsymbol{Zvv}'\boldsymbol{Z}' + \boldsymbol{Zve\varepsilon}' + \boldsymbol{\varepsilon v}'\boldsymbol{Z}' + \boldsymbol{\varepsilon\varepsilon}')\boldsymbol{h}_i$$
$$- \boldsymbol{c}'_{3i}E(\boldsymbol{vv}'\boldsymbol{Z}' + \boldsymbol{v\varepsilon}')\boldsymbol{h}_i$$

$$= \boldsymbol{c}'_{1i}(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{h}_i - \boldsymbol{c}'_{3i}(\boldsymbol{GZ}')\boldsymbol{h}_i$$

$$= (\boldsymbol{c}'_{1i}\boldsymbol{V} - \boldsymbol{c}'_{3i}\boldsymbol{GZ}')\boldsymbol{h}_i$$

$$= [(\boldsymbol{b}'_i - \boldsymbol{l}'_i\boldsymbol{AQ}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1})\boldsymbol{V} - (\boldsymbol{m}'_i - \boldsymbol{l}'_i\boldsymbol{AQ}_v)\boldsymbol{GZ}']\boldsymbol{h}_i$$

$$= [\boldsymbol{b}'_i\boldsymbol{V} - \boldsymbol{l}'_i\boldsymbol{AQ}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{V} - \boldsymbol{m}'_i\boldsymbol{GZ}' + \boldsymbol{l}'_i\boldsymbol{AQ}_v\boldsymbol{GZ}']\boldsymbol{h}_i$$

$$= [\boldsymbol{m}'_i\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{V} - \boldsymbol{m}'_i\boldsymbol{GZ}']\boldsymbol{h}_i$$

$$= \boldsymbol{0}\,.$$

Demonstrou-se assim, que à semelhança do que se verifica para o caso da predição em modelos lineares mistos sem restrições (conforme apresentado no subcapítulo 3.2), o EQMP do BLUP com restrições é também dado pela soma de duas componentes:

$$EQMP\left(\tilde{\theta}_i^R\right) = EQMP\left(\breve{\theta}_i^R\right) + V\left[\boldsymbol{f}'_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] = g_{1i}^R(\boldsymbol{\psi}) + g_{2i}^R(\boldsymbol{\psi}), \qquad (6.3.18)$$

onde

$$g_{1i}^R(\boldsymbol{\psi}) = g_{1i}(\boldsymbol{\psi}) + \boldsymbol{l}'_i\boldsymbol{AQ}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{ZGQ}'_v\boldsymbol{A}'\boldsymbol{l}_i + \boldsymbol{l}'_i\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^R\boldsymbol{\beta})'\boldsymbol{A}'\boldsymbol{l}_i \quad (6.3.19)$$

e

$$g_{2i}^R(\boldsymbol{\psi}) = g_{2i}(\boldsymbol{\psi}) + \boldsymbol{d}'_i(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)'\boldsymbol{A}'\boldsymbol{l}_i + \boldsymbol{l}'_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}_i + \boldsymbol{l}'_i\boldsymbol{A}(\boldsymbol{Q}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{Q}_v\boldsymbol{GZ}'\boldsymbol{V}^{-1}\boldsymbol{X} - \boldsymbol{X}^R)'\boldsymbol{A}'\boldsymbol{l}_i$$

(6.3.20)

Verifica-se, então, que cada uma destas componentes do EQMP do BLUP com restrições é igual à respectiva componente do EQMP do BLUP sem restrições, mais um acréscimo devido à variabilidade gerada pela introdução das restrições na estimação.

6.3.5 O EQMP do EBLUP com restrições

Uma medida da incerteza associada ao EBLUP com restrições, $\hat{\theta}_i^R$, é dada pelo seu EQMP. Kackar e Harville (1984) mostraram que o EQMP do EBLUP pode ser decomposto na soma de três componentes, tal como apresentado na expressão (3.2.8). No caso do EBLUP com restrições, o seu EQMP toma a seguinte forma:

$$EQMP\left(\hat{\theta}_i^R\right) = E\left(\tilde{\theta}_i^R - \theta_i\right)^2 + E\left(\hat{\theta}_i^R - \tilde{\theta}_i^R\right)^2 + 2E\left[\left(\tilde{\theta}_i^R - \theta_i\right)\left(\hat{\theta}_i^R - \tilde{\theta}_i^R\right)\right], \quad (6.3.21)$$

onde $\tilde{\theta}_i^R = \tilde{\theta}_i^R\left(\boldsymbol{y}; \tilde{\boldsymbol{\beta}}, \boldsymbol{\psi}\right)$ é o BLUP com restrições de θ_i e $\hat{\theta}_i^R = \hat{\theta}_i^R\left(\boldsymbol{y}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right)$ é o EBLUP com restrições de θ_i.

Sob condições de normalidade dos efeitos aleatórios e dos erros da sondagem no modelo linear misto sem restrições, e assumindo que as componentes de variância são funções ímpares e invariantes a translações, então o valor esperado do produto cruzado, $E\left[\left(\tilde{\theta}_i - \theta_i\right)\left(\hat{\theta}_i - \tilde{\theta}_i\right)\right]$, é nulo (veja-se, por exemplo, Kackar e Harville (1984), Harville (1985) e Rao (2003)). Contudo, no âmbito do modelo linear misto com restrições, o valor esperado do produto cruzado, $E\left[\left(\tilde{\theta}_i^R - \theta_i\right)\left(\hat{\theta}_i^R - \tilde{\theta}_i^R\right)\right]$, pode não ser nulo, nem negligenciável. Para além disso, esse termo do EQMP parece intratável.

Tal como foi referido anteriormente, também não é tarefa fácil a estimação da variabilidade presente no EBLUP resultante da estimação das componentes de variância, $E\left(\hat{\theta}_i^R - \tilde{\theta}_i^R\right)^2$. Naturalmente que esta tarefa se complica ainda mais no contexto da estimação com restrições. Nestas condições, será certamente muito difícil (ou mesmo impossível) obter uma aproximação analítica para o EQMP do EBLUP modificado pela introdução de restrições.

Dada a complexidade subjacente à estimação das duas últimas componentes do EQMP segundo a decomposição (6.3.21), propõe-se a utilização de outra decomposição do EQMP do EBLUP, utilizada, por exemplo, em Jiang *et al.* (2002) e em Pfeffermann e

Tiller (2005). No caso do EBLUP modificado pela introdução de restrições, essa decomposição do EQMP toma a seguinte forma:

$$EQMP\left(\hat{\theta}_i^R\right) = E\left(\breve{\theta}_i^R - \theta_i\right)^2 + E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2 + 2E\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)\right], \quad (6.3.22)$$

onde $\breve{\theta}_i^R = E(\theta_i|\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\psi})$ é o melhor preditor com restrições (ou BP com restrições) de θ_i e $\hat{\theta}_i^R = E\left(\theta_i|\boldsymbol{y};\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\psi}}\right)$ é o melhor preditor empírico com restrições (ou EBLUP com restrições) de θ_i. Mas, neste caso tem-se que (Pfeffermann e Tiller, 2005):

$$E\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)\right] = E_y\left\{E_\theta\left[\left(\breve{\theta}_i^R - \theta_i\right)\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)\right]|\boldsymbol{y}\right\} = 0, \quad (6.3.23)$$

uma vez que $\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)$ é uma quantidade fixa quando condicionada por $\boldsymbol{y}$, e $\breve{\theta}_i^R = E(\theta_i|\boldsymbol{y})$, tal como apresentado em (3.2.3). Para além disso, a primeira componente de (6.3.22), $E\left(\breve{\theta}_i^R - \theta_i\right)^2$, que mede a variabilidade devida à estimação dos efeitos aleatórios no modelo linear misto com restrições, é dada pela expressão (6.3.19). Por outras palavras, pode afirmar-se que esta componente mede o EQMP do EBLUP quando os hiper-parâmetros do modelo, $\boldsymbol{\delta} = [\boldsymbol{\beta}' \quad \boldsymbol{\psi}']'$**,** são conhecidos. Por sua vez, a segunda componente de (6.3.22), $E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2$, que quantifica a variabilidade adicional do EQMP devido à estimação desses hiper-parâmetros, parece intratável. Desta forma, a decomposição (6.3.22) pode então ser reescrita como:

$$EQMP\left(\hat{\theta}_i^R\right) = g_{1i}^R(\boldsymbol{\psi}) + E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2. \quad (6.3.24)$$

Também neste caso, será certamente muito difícil (ou mesmo impossível) obter uma aproximação analítica para o EQMP do EBLUP com restrições, devido à dificuldade inerente à estimação de $E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2$, pelo que se propõe a utilização de métodos por reamostragem. Para além disso, os métodos por reamostragem são robustos à violação da normalidade dos erros da sondagem e dos efeitos aleatórios do modelo.

6.3.6 Aproximação *bootstrap* do EQMP do EBLUP com restrições

Em primeiro lugar, propõe-se um método *bootstrap* robusto (Wu, 1986) no contexto de populações finitas. À semelhança do que foi feito por Pfeffermann e Tiller (2001), a

versão do método *bootstrap* aqui proposto é baseada na decomposição do estimador *jackknife* proposto por Jiang *et al.* (2002). Propõe-se que a estimação das duas componentes de (6.3.24) seja efectuada segundo o seguinte procedimento *bootstrap*:

1. Calcular as estimativas das componentes de variância pelo método dos momentos, $\widehat{\boldsymbol{\psi}}$, com base nos dados iniciais, $\boldsymbol{y}$, e ajustar o modelo (6.3.1) de forma a determinar as estimativas dos efeitos fixos e dos efeitos aleatórios sem restrições, $\hat{\boldsymbol{\xi}} = \left[\widehat{\boldsymbol{\beta}}' \;\; \widehat{\boldsymbol{v}}'\right]'$.

2. Determinar as estimativas dos efeitos fixos e dos efeitos aleatórios com restrições, $\hat{\boldsymbol{\xi}}^R = \left[\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}\prime} \;\; \widehat{\boldsymbol{v}}^{\boldsymbol{R}\prime}\right]'$ onde $\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} = \widehat{\boldsymbol{\beta}}^{\boldsymbol{R}}\left(\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right)$ e $\widehat{\boldsymbol{v}}^{\boldsymbol{R}} = \widehat{\boldsymbol{v}}^{\boldsymbol{R}}\left(\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right)$.

3. Calcular as estimativas EBLUP do parâmetro de interesse com restrições, $\hat{\theta}_i^R\left(\hat{\boldsymbol{\xi}}^R\right)$, e do primeiro termo do EQMP, $g_{1i}^R\left(\hat{\boldsymbol{\xi}}^R\right)$, onde $\hat{\boldsymbol{\xi}}^R = \left[\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}\prime} \;\; \widehat{\boldsymbol{v}}^{\boldsymbol{R}\prime}\right]'$.

4. Gerar o vector aleatório $\boldsymbol{v}^*$, com $\boldsymbol{v}^* \sim N\left(\boldsymbol{0}; \widehat{\boldsymbol{G}}\right)$.

5. Gerar o vector aleatório $\boldsymbol{\varepsilon}^*$, com $\boldsymbol{\varepsilon}^* \sim N\left(\boldsymbol{0}; \widehat{\boldsymbol{R}}\right)$, independente de $\boldsymbol{v}^*$.

6. Construir o conjunto de dados *bootstrap* $\boldsymbol{y}^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{v}^* + \boldsymbol{\varepsilon}^*$.

7. Calcular estimativas *bootstrap* das componentes de variância, $\widehat{\boldsymbol{\psi}}^*$, com base nos dados *bootstrap*, $\mathbf{y}^*$, e ajustar o modelo (6.3.1) de forma a obter as estimativas *bootstrap* dos efeitos fixos e dos efeitos aleatórios sem restrições, $\hat{\boldsymbol{\xi}}^* = \left[\widehat{\boldsymbol{\beta}}'^* \;\; \widehat{\boldsymbol{v}}'^*\right]'$.

8. Determinar as estimativas dos efeitos fixos e dos efeitos aleatórios com restrições, $\hat{\boldsymbol{\xi}}^{R*} = \left[\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}*\prime} \;\; \widehat{\boldsymbol{v}}^{\boldsymbol{R}*\prime}\right]'$ onde $\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}*} = \widehat{\boldsymbol{\beta}}^{\boldsymbol{R}*}\left(\boldsymbol{y}^*, \widehat{\boldsymbol{\psi}}^*\right)$ e $\widehat{\boldsymbol{v}}^{\boldsymbol{R}*} = \widehat{\boldsymbol{v}}^{\boldsymbol{R}*}\left(\boldsymbol{y}^*, \widehat{\boldsymbol{\psi}}^*\right)$.

9. Calcular estimativas *bootstrap* do EBLUP com restrições e do primeiro termo do seu EQMP, com base nos dados *bootstrap*, $\mathbf{y}^*$, e utilizando estimativas *bootstrap* das componentes de variância, $\widehat{\boldsymbol{\psi}}^*$: $\hat{\theta}_i^R\left(\hat{\boldsymbol{\xi}}^{R*}\right) = \boldsymbol{l}'_i\left[\tilde{\boldsymbol{\xi}}^* + \boldsymbol{A}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}^*\right)\right]$ e $g_{1i}^R\left(\hat{\boldsymbol{\xi}}^{R*}\right)$.

10. Repetir as etapas 4)-9) B vezes. Defina-se, por conveniência, $\widehat{\boldsymbol{\delta}}^{*(\boldsymbol{b})} = \left[\widehat{\boldsymbol{\beta}}'^{*(\boldsymbol{b})} \;\; \widehat{\boldsymbol{\psi}}'^{*(\boldsymbol{b})}\right]'$ como o vector das estimativas *bootstrap* dos hiper-parâmetros obtidas na b-ésima réplica *bootstrap*; e $\hat{\theta}_i^R\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right)$ como a estimativa *bootstrap* do EBLUP com restrições de θ_i obtida na b-ésima réplica *bootstrap*, $b = 1, \ldots, B$.

11. Calcular uma estimativa *bootstrap* de $M_{2i} = E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2$, usando a seguinte aproximação de Monte Carlo:

$$\widehat{M}_{2i}^B = B^{-1} \sum_{b=1}^{B} \left[\hat{\theta}_i^R\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right) - \hat{\theta}_i^R\left(\widehat{\boldsymbol{\delta}}\right)\right]^2. \qquad (6.3.25)$$

Uma vez obtida esta estimativa *bootstrap*, propõe-se o seguinte estimador *bootstrap* com correcção de enviesamento para o EQMP do EBLUP com restrições:

$$eqmp^B\left(\hat{\theta}_i^R\right) = g_{1i}^R\left(\widehat{\boldsymbol{\delta}}\right) - B^{-1} \sum_{b=1}^{B} \left[g_{1i}^R\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right) - g_{1i}^R\left(\widehat{\boldsymbol{\delta}}\right)\right] + \widehat{M}_{2i}^B. \qquad (6.3.26)$$

Note-se que é necessário fazer uma correcção de enviesamento, porque $g_{1i}^R\left(\widehat{\boldsymbol{\delta}}\right)$ é um estimador enviesado de $g_{1i}^R(\boldsymbol{\delta})$.

6.3.7 Aproximação *jackknife* do EQMP do EBLUP com restrições

Uma abordagem alternativa para a estimação do EQMP do EBLUP com restrições consiste no uso de um procedimento *jackknife*, baseado nos trabalhos de Jiang *et al.* (2002). Os passos para a estimação das duas componentes de (6.3.24), segundo este procedimento *jackknife*, são os seguintes:

1. Calcular as estimativas das componentes de variância pelo método dos momentos, $\widehat{\boldsymbol{\psi}}_{-\boldsymbol{e}}$, depois de eliminada a *e*-ésima observação $(y_e, \boldsymbol{x}'_e)$ do conjunto de dados iniciais, $\{(y_i, \boldsymbol{x}'_i); i = 1, \ldots, m\}$.

2. Ajustar o modelo (6.3.1) de forma a determinar as estimativas dos efeitos fixos e dos efeitos aleatórios sem restrições, $\widehat{\boldsymbol{\xi}}_{-\boldsymbol{e}} = \left[\widehat{\boldsymbol{\beta}}'_{-\boldsymbol{e}}\ \widehat{\boldsymbol{v}}'_{-\boldsymbol{e}}\right]'$, depois de eliminada a *e*-ésima observação $(y_e, \boldsymbol{x}'_e)$ do conjunto de dados iniciais, $\{(y_i, \boldsymbol{x}'_i); i = 1, \ldots, m\}$.

3. Determinar as estimativas dos efeitos fixos com restrições, $\widehat{\boldsymbol{\beta}}_{-e}^R = \widehat{\boldsymbol{\beta}}^{\boldsymbol{R}}\left(\boldsymbol{y}_{-\boldsymbol{e}}, \widehat{\boldsymbol{\psi}}_{-\boldsymbol{e}}\right)$.

4. Calcular as estimativas EBLUP do parâmetro de interesse com restrições, $\hat{\theta}_i^R\left(\widehat{\boldsymbol{\delta}}_{-e}\right)$, e do primeiro termo do EQMP, $g_{1i}^R\left(\widehat{\boldsymbol{\delta}}_{-e}\right)$, onde $\widehat{\boldsymbol{\delta}}_{-e} = \left[\widehat{\boldsymbol{\beta}}'_{-\boldsymbol{e}}\ \ \widehat{\boldsymbol{\psi}}'_{-\boldsymbol{e}}\right]'$.

5. Repetir as etapas 1)-4) m vezes, de forma a obter m estimativas para $\hat{\theta}_i^R(\widehat{\boldsymbol{\delta}}_{-e})$ e para $g_{1i}^R(\widehat{\boldsymbol{\delta}}_{-e})$, $e = 1, \ldots, m$.

6. Calcular uma estimativa *jackknife* de $M_{2i} = E\left(\hat{\theta}_i^R - \breve{\theta}_i^R\right)^2$, usando o seguinte estimador:

$$\widehat{M}_{2i}^J = \frac{m-1}{m}\sum_{e=1}^{m}\left[\hat{\theta}_i^R(\widehat{\boldsymbol{\delta}}_{-e}) - \hat{\theta}_i^R(\widehat{\boldsymbol{\delta}})\right]^2. \qquad (6.3.27)$$

Uma vez obtidas esta estimativa *jackknife*, propõe-se a utilização do seguinte estimador *jackknife* com correcção de enviesamento para o EQMP do EBLUP com restrições:

$$eqmp^J(\hat{\theta}_i^R) = g_{1i}^R(\widehat{\boldsymbol{\delta}}) - \frac{m-1}{m}\sum_{e=1}^{m}\left[g_{1i}^R(\widehat{\boldsymbol{\delta}}_{-e}) - g_{1i}^R(\widehat{\boldsymbol{\delta}})\right] + \widehat{M}_{2i}^J, \qquad (6.3.28)$$

onde $g_{1i}^R(\boldsymbol{\delta})$ é dado por (6.3.19). De acordo com os resultados gerais de Jiang *et al.* (2002), o estimador (6.3.28) é aproximadamente não enviesado até à segunda ordem.

6.4 MODELO COM RESTRIÇÕES COM DADOS ESPACIAIS E CRONOLÓGICOS

Este subcapítulo é dedicado à apresentação de um modelo de estimação em pequenos domínios com dados espaciais e cronológicos, considerando restrições na estimação, de forma a garantir a consistência interna das estimativas. Decidiu apresentar-se este modelo de forma resumida, pelo facto de constituir uma extensão do trabalho desenvolvido no capítulo quinto, adicionado da estimação com restrições apresentado no subcapítulo 6.3.

Considere-se o modelo de estimação em pequenos domínios com erros correlacionados espacialmente e temporalmente, especificado no subcapítulo 5.2, dado por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \qquad (6.4.1)$$

onde $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{I}_{mT}]$, $\mathbf{Z}_1 = \mathbf{I}_m \otimes \mathbf{1}_T$ e $\boldsymbol{\upsilon} = [\mathbf{v}' \quad \mathbf{u}_2']'$, e no qual se assume que os termos de erro $\mathbf{v} = (\mathbf{I}_m - \phi\mathbf{W})^{-1}\mathbf{u}_1$, $\mathbf{u}_2$ e $\boldsymbol{\varepsilon}$ são mutuamente independentes. Assume-se também

que $\mathbf{u}_1 \overset{iid}{\sim} N\left(\mathbf{0};\sigma_u^2\mathbf{I}_m\right)$, $\mathbf{u}_2 \sim N\left(\mathbf{0};\sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma}\right)$ e $\boldsymbol{\varepsilon} \overset{iid}{\sim} N\left(\mathbf{0};\mathbf{R}\right)$, onde $\boldsymbol{\Gamma} = \{\gamma_{rs}\}$ é uma matriz $T \times T$ com elementos $\gamma_{rs} = \rho^{|r-s|}/\left(1-\rho^2\right)$, r, s=1, ..., T e $\mathbf{R} = diag_{1\le i\le m}\left(\mathbf{R}_i\right)$ com $\mathbf{R}_i = diag_{1\le t\le T}\left(\sigma_{it}^2\right)$. Relembre-se ainda que a matriz de covariâncias de $\boldsymbol{\upsilon}$ é dada por $\mathbf{G} = diag_{1\le k\le 2}\left(\mathbf{G}_k\right)$, onde $\mathbf{G}_1 = E\left(\mathbf{v}\mathbf{v}'\right) = \sigma_u^2\mathbf{B}^{-1}$ e $\mathbf{G}_2 = E\left(\mathbf{u}_2\mathbf{u}_2'\right) = \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma}$, e que a matriz de covariâncias de $\mathbf{y}$ é dada por $\mathbf{V} = \mathbf{R} + \mathbf{Z}_1\sigma_u^2\mathbf{B}^{-1}\mathbf{Z}_1' + \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma}$, onde $\mathbf{B} = \left(\mathbf{I}_m - \phi\mathbf{W}\right)'\left(\mathbf{I}_m - \phi\mathbf{W}\right)$.

Admita-se agora que no âmbito deste modelo é definido um conjunto de *A* restrições em cada período de tempo, exigindo-se que a média ponderada das estimativas indirectas (baseadas no modelo) produzidas para os pequenos domínios de uma dada região seja igual à estimativa directa do parâmetro de interesse nessa região. As *AT* restrições são definidas como:

$$\boldsymbol{Q}\boldsymbol{\xi} = \boldsymbol{q}\,, \tag{6.4.2}$$

onde $\boldsymbol{q} = col_{1\le a\le A;1\le t\le T}(y_{at})$ (*AT*×1), $\boldsymbol{\xi} = [\boldsymbol{\beta}'\ \boldsymbol{\upsilon}']'$ [(*p*+*m*+*mT*)×1] e $\boldsymbol{Q} = [\boldsymbol{X}^R\ \boldsymbol{Q}_\upsilon]$ [*AT*×(*p*+*m*+*mT*)]. Para além disso, tem-se que $\boldsymbol{X}^R = col_{1\le a\le A;1\le t\le T}(\boldsymbol{\delta}'_{1at}\boldsymbol{\Psi}_1\boldsymbol{X})$ (*AT*×*p*), $\boldsymbol{\Psi}_1 = diag_{1\le i\le m;1\le t\le T}(\omega_{it})$ (*mT*×*mT*) onde $\omega_{it} = \frac{n_{it}}{n_{at}}$, e $\boldsymbol{\delta}_{1at} = col_{1\le i\le m;1\le t\le T}\,(\delta_{1ait}\,)$ (*mT*×1), onde δ_{1ait}=1 se o *i*-ésimo domínio referente ao período *t* pertence à *a*-ésima região associada ao período *t* e δ_{1ait}=0 em caso contrário. Tem-se, ainda, que $\boldsymbol{Q}_\upsilon = col_{1\le a\le A;1\le t\le T}\left(\boldsymbol{\delta}'_{2at}\boldsymbol{\Psi}_2\right)$ [*AT*×(*m*+*mT*)], $\boldsymbol{\Psi}_2 = diag_{0\le k\le 1}(\boldsymbol{\Psi}_k)$ [(*m*+*mT*)×(*m*+*mT*)] onde $\boldsymbol{\Psi}_0 = diag_{1\le i\le m}(\bar{\omega}_i)$ (*m*×*m*) com $\bar{\omega}_i = \frac{\bar{n}_i}{\bar{n}_a}$, e $\boldsymbol{\delta}_{2at} = col_{0\le k\le 1}\left(\boldsymbol{\delta}_{kat}\,\right)$ [(*m*+*mT*)×1] onde $\boldsymbol{\delta}_{0at} = col_{1\le i\le m}\,(\delta_{0ai}\,)$ (*m*×1), no qual δ_{0ai}=1 se o *i*-ésimo domínio pertence à *a*-ésima região e δ_{0ai}=0 em caso contrário.

O estimador BLUP espaciotemporal com restrições da média no *i*-ésimo pequeno domínio no período *t*, é dado por:

$$\tilde{\theta}_{it}^R = \boldsymbol{x}'_{it}\tilde{\boldsymbol{\beta}} + \left(\sigma_u^2\boldsymbol{\zeta}'_i \otimes \mathbf{1}'_t + \sigma^2\boldsymbol{\zeta}'_{it}\right)\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right) \tag{6.4.3}$$

$$+\left(\boldsymbol{x}'_{it}\mathbf{C}_{11}\boldsymbol{X}^{R\,\prime} + \boldsymbol{x}'_{it}\mathbf{C}'_{21}\boldsymbol{Q}'_\upsilon + \boldsymbol{m}'_{it}\mathbf{C}_{21}\boldsymbol{X}^{R\,\prime} + \boldsymbol{m}'_{it}\mathbf{C}_{22}\boldsymbol{Q}_\upsilon{}'\right)(\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}')^{-1}\left(\boldsymbol{q} - \boldsymbol{Q}\tilde{\boldsymbol{\xi}}\right)$$

onde $\boldsymbol{\varsigma}_i' = \{\varsigma_{ii'}\}$ é a *i*-ésima linha da matriz $\mathbf{B}^{-1}$; $\boldsymbol{\zeta}_{it}'$ é um vector linha de dimensão $1\times mT$ com *m* blocos, no qual cada bloco é formado por um vector linha *T*-dimensional, sendo o *i*-ésimo bloco formado pela *t*-ésima linha da matriz $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}_t$, e os restantes blocos por vectores nulos $\mathbf{0}_{1\times T}$, *i*,*i*'=1, …, *m*; $\mathbf{C}_{11} = (\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}$; $\mathbf{C}_{21} = -\mathbf{G}\boldsymbol{Z}'\mathbf{V}^{-1}\boldsymbol{X}\mathbf{C}_{11}$ e $\mathbf{C}_{22} = (\boldsymbol{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} - \mathbf{C}_{21}\boldsymbol{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}$. O estimador EBLUP espaciotemporal com restrições do parâmetro de interesse, $\hat{\theta}_{it}^{R}$, obtém-se quando se substitui em (6.4.3) os parâmetros de variância desconhecidos pelas respectivas estimativas, dadas por (5.5.4) e por (5.5.7).

A medição da incerteza do EBLUP espaciotemporal com restrições deve ser efectuada através da utilização de uma aproximação por reamostragem do EQMP (metodologias *bootstrap* ou *jackknife* apresentadas nas secções 6.3.6 ou 6.3.7, respectivamente). Se se decidir utilizar uma metodologia *bootstrap*, então um estimador *bootstrap* do EQMP do EBLUP espaciotemporal com restrições é dado por:

$$eqmp^{B}\left(\hat{\theta}_{it}^{R}\right) = g_{1it}^{R}\left(\widehat{\boldsymbol{\delta}}\right) - B^{-1}\sum_{b=1}^{B}\left[g_{1it}^{R}\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right) - g_{1it}^{R}\left(\widehat{\boldsymbol{\delta}}\right)\right] + \widehat{M}_{2it}^{B}, \qquad (6.4.4)$$

onde $g_{1it}^{R}(\boldsymbol{\delta}) = g_{1it}(\boldsymbol{\delta}) + \boldsymbol{l}'_{it}\boldsymbol{A}\boldsymbol{Q}_{v}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Q}'_{v}\boldsymbol{A}'\boldsymbol{l}_{it} + \boldsymbol{l}'_{it}\boldsymbol{A}(\boldsymbol{q} - \boldsymbol{X}^{R}\boldsymbol{\beta})(\boldsymbol{q} - \boldsymbol{X}^{R}\boldsymbol{\beta})'\boldsymbol{A}'\boldsymbol{l}_{it}$, $\widehat{M}_{2it}^{B} = B^{-1}\sum_{b=1}^{B}\left[\hat{\theta}_{it}^{R}\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right) - \hat{\theta}_{it}^{R}\left(\widehat{\boldsymbol{\delta}}\right)\right]^{2}$ e $g_{1it}(\boldsymbol{\delta})$ é dada por (5.6.2).

7. ESTUDO EMPÍRICO – ESTIMAÇÃO EM PEQUENOS DOMÍNIOS DO PREÇO MÉDIO DE TRANSACÇÃO DA HABITAÇÃO

7.1 INTRODUÇÃO

Relembre-se que o último grande objectivo desta tese consiste na avaliação, por simulação de Monte Carlo, da qualidade:

a) dos estimadores propostos para estimar o preço médio de transacção da habitação, relativamente a diversos outros estimadores directos e indirectos habitualmente utilizados na estimação desse tipo de parâmetros de interesse em pequenos domínios;

b) dos estimadores do EQMP propostos, utilizados para avaliar a incerteza dos estimadores combinados dos parâmetros de interesse sem restrições.

A avaliação da qualidade desses estimadores é efectuada através da realização de estudos empíricos por simulação de diferentes naturezas:

a) é efectuado um estudo por simulação do tipo *design-based* para avaliar o desempenho dos estimadores dos parâmetros de interesse. Este tipo de estudo empírico é baseado numa pseudo-população finita (fixa) gerada a partir de uma amostra aleatória de dados reais, sendo avaliada a qualidade dos estimadores através de um conjunto de medidas de enviesamento, de precisão e de eficiência num contexto de uma população real e de um método de amostragem realista. Neste estudo por simulação, a avaliação das propriedades dos estimadores é efectuada sob uma perspectiva de amostragem repetida, sendo as estimativas obtidas em cada

amostra extraída da pseudo-população comparadas com os respectivos parâmetros de interesse dessa pseudo-população.

b) é efectuado um estudo por simulação do tipo *model-based* para avaliar o desempenho dos estimadores do EQMP dos estimadores combinados dos parâmetros de interesse sem restrições. Este tipo de estudo empírico é baseado num modelo de superpopulação utilizado para gerar uma população artificial, sendo igualmente avaliada a qualidade dos estimadores através de um conjunto de medidas de precisão e de enviesamento. Neste caso, a avaliação das propriedades dos estimadores é efectuada pela comparação das estimativas do EQMP, obtidas em cada conjunto de dados gerado pelo modelo de superpopulação postulado, com as respectivas aproximações aos verdadeiros valores do EQMP dos EBLUP, calculadas com base num elevado conjunto de dados gerados por esse modelo.

Dada a necessidade de se efectuarem estudos empíricos independentes para se alcançarem os objectivos propostos, é conveniente revelar-se e justificar-se a ordem pela qual esses estudos vão ser apresentados, definindo-se desta forma a estrutura deste capítulo.

O primeiro estudo empírico, apresentado no subcapítulo 7.2, é um estudo por simulação do tipo *design-based* para avaliar o desempenho dos estimadores propostos para estimar o preço médio de transacção da habitação (EBLUP espaciotemporal sem e com restrições), relativamente a diversos outros estimadores directos e indirectos habitualmente utilizados na estimação desse tipo de parâmetros de interesse em pequenos domínios. O estimador EBLUP espaciotemporal é assistido pelo modelo proposto no quinto capítulo, enquanto o estimador EBLUP espaciotemporal com restrições foi deduzido no sexto capítulo. Uma vez que os principais objectivos desta investigação consistem em propor estimadores dos parâmetros de interesse, alternativos aos estimadores tradicionais, e em introduzir restrições na estimação, então parece natural que a avaliação do desempenho dos estimadores propostos seja feita antes da avaliação das medidas da incerteza associadas a esses novos estimadores.

Uma vez avaliado o mérito relativo dos estimadores propostos para estimar o preço médio de transacção da habitação, do ponto de vista da amostragem repetida, é altura de se avaliar o mérito relativo de estimadores *model-based* do EQMP dos EBLUP. Assim,

dedica-se o subcapítulo 7.3 à apresentação de um segundo estudo, por simulação do tipo *model-based*, para avaliar o desempenho dos estimadores propostos do EQMP de estimadores combinados sem restrições. Este estudo empírico divide-se em duas partes.

Na primeira parte, é avaliado o desempenho dos estimadores por reamostragem (*jackknife* e *bootstrap*) do EQMP do EBLUP temporal, face ao estimador analítico desse EQPM, no contexto do modelo seccional e cronológico de Rao-Yu. Esses estimadores por reamostragem foram propostos nas secções 4.3.6 e 4.3.7. Decidiu-se que esta seria a primeira parte do estudo empírico a ser apresentado por três razões. Em primeiro lugar, porque se está a trabalhar no âmbito de um modelo conhecido na literatura (modelo de Rao-Yu). Em segundo lugar, porque se está a avaliar o desempenho de duas novas metodologias por reamostragem para medição da incerteza associada ao EBLUP temporal, face a uma metodologia estabelecida na literatura (metodologia delta apresentada por Rao e Yu, 1994). E por último, porque a verificação da adequação e qualidade de metodologias por reamostragem para medição da incerteza associada aos EBLUP, como alternativa a um estimador delta baseado em longos desenvolvimentos analíticos, poder constituir uma alternativa promissora no contexto de modelos longitudinais de estimação em pequenos domínios mais complexos[58], no âmbito dos quais é geralmente impossível deduzir uma aproximação analítica do EQMP do EBLUP. Esta primeira parte do estudo empírico do tipo *model-based* será apresentada na secção 7.3.2.

Na segunda parte daquele estudo empírico, é avaliado o desempenho dos estimadores propostos (analítico, *jackknife* e *bootstrap*) para medição da incerteza associada ao EBLUP espaciotemporal sem restrições. Este é o trabalho apresentado na secção 7.3.3. A avaliação do desempenho dos estimadores por reamostragem do EQMP do EBLUP espaciotemporal com restrições não fez parte dos objectivos deste trabalho.

Por último, no subcapítulo 7.4 são apresentadas as estimativas do preço médio de transacção da habitação para diferentes níveis de agregação geográfica, bem como as respectivas medidas de precisão.

As tarefas de manipulação de dados e de cálculo foram realizadas no programa estatístico SAS, versões 9.1 e 8.0. Estas versões do SAS estavam instaladas,

[58] O modelo espaciotemporal com restrições, proposto no subcapítulo 6.4, é um destes casos.

respectivamente, em dois computadores com as seguintes características: processador *Intel Core 2 Duo CPU T7300* de 2,00 *Gigahertz*, 2 *Gigabytes* de memória RAM, disco rígido de 150 *Gigabytes* e sistema operativo *Microsoft Windows XP Professional*; e processador *Intel Pentium 4* de 1,80 *Gigahertz*, 224 *Megabytes* de memória RAM, disco rígido de 18,6 *Gigabytes* e sistema operativo *Microsoft Windows XP Professional.*

7.2 AVALIAÇÃO DO DESEMPENHO DOS ESTIMADORES PROPOSTOS PARA ESTIMAR O PREÇO MÉDIO DE TRANSACÇÃO DA HABITAÇÃO

7.2.1 Introdução

O objectivo deste subcapítulo consiste, portanto, na apresentação do estudo empírico do tipo *design-based*, baseado numa simulação de Monte Carlo sobre uma pseudo-população. Este estudo tem como objectivo a avaliação do desempenho dos estimadores propostos para estimar o preço médio de transacção da habitação (EBLUP espaciotemporal sem e com restrições), relativamente a diversos outros estimadores directos e indirectos habitualmente utilizados na estimação desse tipo de parâmetros de interesse em pequenos domínios. Neste estudo empírico, as propriedades estatísticas dos estimadores pontuais são avaliadas do ponto de vista *design-based*, não existindo qualquer dependência das hipóteses dos modelos postulados.

Uma vez que este estudo empírico constitui um dos objectos principais deste trabalho, e pelo facto deste subcapítulo ser extenso, decidiu apresentar-se aqui a sua estrutura. Assim, na secção 7.2.2 são apresentados os inquéritos que forneceram os dados para este estudo e descritas algumas tarefas de preparação e exploração de dados. A secção 7.2.3 é dedicada à apresentação do desenho do estudo empírico do tipo *design-based* e a secção 7.2.4 à apresentação das expressões dos estimadores utilizados neste estudo. O diagnóstico dos modelos que assistem a estimação é efectuado na secção 7.2.5. Na secção 7.2.6 encontram-se as medidas utilizadas para avaliar as propriedades dos estimadores. Finalmente, na secção 7.2.7 são apresentados, analisados e discutidos os resultados obtidos no referido estudo empírico por simulação de Monte Carlo.

7.2.2 Inquéritos e preparação de dados

7.2.2.1 Inquérito aos Preços Médios de Transacção na Habitação

O Inquérito aos Preços Médios de Transacção na Habitação (IPTH) (anexo 1) era um inquérito da responsabilidade do INE, realizado trimestralmente a uma amostra de empresas de mediação imobiliária, doravante designadas apenas por empresas, sediadas em Portugal continental. A informação recolhida no âmbito deste inquérito referia-se exclusivamente a transacções de habitações (apartamentos[59] ou moradias independentes[60]) e de terrenos, destinados à habitação. As variáveis recolhidas em cada uma das rubricas eram as seguintes:

- Identificação da empresa: concelho, número de identificação de pessoa colectiva, identificação dos estabelecimentos;

- Registo de transacções de terrenos para construção da habitação: localização – concelho/freguesia, terreno com alvará de loteamento/estudo de viabilidade, registo respeitante a contrato-promessa de compra e venda, área total de construção autorizada e da qual mais de 70% é destinada à habitação, número de fogos que se previa construir para apartamentos e/ou moradias, valor da transacção;

- Registo de transacções de habitações: localização – concelho/freguesia, registo respeitante contrato-promessa de compra e venda, ano de conclusão da obra de construção ou reconstrução do edifício, área útil[61], valor da transacção, natureza do

[59] Segundo o INE, um apartamento é uma unidade de alojamento inserida num edifício de construção permanente, com mais de um fogo, cuja entrada principal dá para uma escada, corredor ou pátio (INE, 2001*b*).

[60] Segundo o INE, uma moradia independente é um edifício isolado, geminado ou em fila a que corresponde apenas uma unidade de alojamento familiar e cuja entrada principal dá, geralmente, para uma rua, pátio ou para um terreno circundante do edifício (INE, 2001*b*).

[61] Segundo o INE, a área útil é a soma das áreas de todos os compartimentos da habitação, incluindo vestíbulos, circulações interiores, instalações sanitárias, arrumos, outros compartimentos de função similar e armários nas paredes, e mede-se pelo perímetro interior das paredes que limitam o fogo, descontando encalços até 30 centímetros, paredes interiores, divisórias e condutas (INE, 2001*b*).

alojamento (apartamento, moradia independente), tipologia[62] do apartamento ou da moradia independente, fase da obra (em projecto ou construção, pronto a habitar).

O IPTH era um inquérito longitudinal com rotação[63], realizado trimestralmente por amostragem aleatória por conglomerados, às empresas sediadas em Portugal continental. A população era constituída por todas as empresas sediadas em Portugal continental que transaccionavam prédios urbanos, sendo a empresa a unidade estatística de inquirição. A base de sondagem existente era formada por 4.671 empresas. A recolha de informação referente a este inquérito era feita através de suporte informático, tendo sido concebida uma aplicação para ser instalada nas empresas inquiridas. Contudo, para as empresas que não pretendessem responder através de suporte informático, existia a possibilidade de responderem por via postal.

Para efeitos de selecção da amostra, a população de empresas foi estratificada pelo cruzamento das seguintes variáveis: *valor de volume de negócios*[64], *NUTSIII*[65] e *concelho* (apenas para as áreas metropolitanas de Lisboa e do Porto). A descrição dos escalões definidos pelas variáveis de estratificação encontra-se no apêndice 1. A estratificação da população resultante do cruzamento das três variáveis anteriores produziu 477 estratos, dos quais foi seleccionada uma amostra aleatória simples de empresas (unidades primárias), sendo observadas todas as transacções de habitações e de terrenos, destinados à habitação (unidades secundárias), realizadas por cada uma dessas empresas. Porém, alguns estratos foram inquiridos de forma exaustiva. Os estratos inquiridos exaustivamente foram os caracterizados por um escalão de valor de volume de negócios superior ou igual a sete. Conclui-se então, no que se refere ao plano de sondagem, que o IPTH correspondia a um inquérito longitudinal com rotação, no qual foi utilizada uma amostragem aleatória por conglomerados previamente estratificados, em cada vaga.

[62] Segundo o INE, a tipologia dos fogos (T0, T1, T2, T3, ...) corresponde à classificação do fogo segundo o número de quartos de dormir (INE, 2001*b*).

[63] Uma exposição detalhada sobre inquéritos repetidos no tempo pode ser encontrada, por exemplo, em Binder (1998).

[64] Não existiam empresas de mediação imobiliária sediadas em Portugal continental com valor de volume de negócios superior ao definido no nono escalão.

[65] A variável de localização do imóvel transaccionado ao nível de NUTSIII é obtida a partir da sua localização ao nível de *concelho*.

O facto da estratificação utilizada no IPTH ser muito fina (477 estratos) numa população com 4.671 unidades estatísticas, levou a que muitos estratos tivessem dimensão populacional nula ou muito pequena[66]. Esta situação, associada à existência de não respostas, conduziu a um elevado número de estratos com dimensão amostral[67] nula. No apêndice 2 são apresentadas as dimensões populacional e amostrais e as taxas de sondagem em cada estrato, ao longo dos sete trimestres. De forma resumida, na tabela 7.2.1 é apresentada a dimensão amostral real de unidades primárias e de unidades secundárias, em cada trimestre.

Tabela 7.2.1: Número de estratos com dimensão amostral não nula e dimensão amostral real de unidades primárias e de unidades secundárias, por trimestre

Trimestre	N.º de estratos com dimensão amostral não nula	Dimensão amostral de unidades primárias	Dimensão amostral de unidades secundárias
1	121	486	2.548
2	123	498	2.367
3	122	478	2.226
4	115	414	1.681
5	120	441	1.699
6	122	424	1.739
7	122	467	1.954

Uma vez que se está a trabalhar com dados recolhidos através de um plano de sondagem que levanta algumas dificuldades ao nível da estimação dos parâmetros de interesse em pequenos domínios, o qual é da responsabilidade do INE, e que o tratamento de não respostas não cabe no âmbito deste trabalho, procedeu-se a uma redefinição da estratificação da população-alvo. Desta forma, admite-se que a população de empresas foi estratificada pelo cruzamento das seguintes variáveis: *NUTSIII* e *concelho* (apenas para as áreas metropolitanas de Lisboa e do Porto), tendo resultado em 51 estratos. Esta redefinição usada no âmbito deste trabalho, de forma a evitar a existência de estratos com dimensão amostral nula, é independente do problema que se pretende resolver, ou seja, da produção de estimativas com precisão aceitável do preço médio de transacção da habitação em pequenos domínios.

A utilização de apenas duas variáveis de estratificação, no âmbito deste estudo

[66] Existiam 243 estratos com uma dimensão populacional nula, o que corresponde a 51% do total de estratos definidos no plano de sondagem do IPTH.

[67] Só 151 estratos apresentavam dimensão amostral não nula em pelo menos uma vaga do IPTH.

empírico, justifica-se pelo facto da utilização do elevado número de estratos existente no verdadeiro plano de sondagem do IPTH, conduzir a um grande número de estratos sem nenhumas unidades estatísticas primárias pertencentes à amostra. Por esta razão, e pelo facto de não se pretender fazer inferência estatística ao nível de cada estrato definido pelo plano de sondagem adoptado no IPTH, não havia nenhum motivo que conduzisse à utilização de uma estratificação tão fina. Convém, ainda, notar que o objectivo definido no início deste estudo assenta na avaliação das qualidades dos estimadores propostos para estimar o preço médio de transacção da habitação ao nível de NUTSIII, domínios estes que apesar de coincidirem com 23 dos 51 estratos definidos pelo plano de sondagem utilizado na simulação, podem conter unidades secundárias com diferentes probabilidades de inclusão. Esta situação resulta do facto dos domínios serem cruzados com os estratos, ou seja, devido ao facto de empresas sediadas num particular estrato (NUTSIII), poderem efectuar transacções em diferentes domínios de interesse (NUTSIII).

Neste estudo empírico, foram utilizados os dados resultantes deste inquérito correspondentes aos registos de transacções de habitações nas variáveis *valor da transacção*, *área útil* e *concelho*. Desta forma, foram consideradas todas as transacções de habitações independentemente da sua natureza, tipologia, fase da obra ou ano de conclusão da obra de construção ou reconstrução do edifício.

7.2.2.2 Inquérito aos Preços de Avaliação Bancária na Habitação

O Inquérito aos Preços de Avaliação Bancária na Habitação[68] (IABH) (*vide* um excerto no anexo 2) é um inquérito postal, também da responsabilidade do INE, realizado mensalmente, de forma exaustiva, ao universo de instituições bancárias que intervêm no mercado de crédito à habitação em Portugal continental. Este inquérito tem como unidades estatísticas de inquirição as instituições bancárias. Para cada unidade estatística de inquirição são observados todos os alojamentos avaliados no âmbito dos processos de análise para concessão de crédito à habitação. A informação recolhida por este inquérito reporta-se ao conjunto de avaliações realizadas em Portugal continental, independentemente de ter havido, ou não, lugar à aprovação do crédito, ou do destino que posteriormente vai ser dado à habitação.

[68] Actualmente denominado por Inquérito à Avaliação Bancária na Habitação.

No âmbito do IABH, é recolhida informação sobre: dados gerais da avaliação, valor da avaliação, caracterização da habitação (natureza, tipologia, ano de conclusão da obra de construção ou reconstrução do edifício, área), e localização (concelho).

Actualmente, o INE publica trimestralmente os valores médios de avaliação bancária da habitação ao nível de Portugal continental, NUTSII, NUTSIII, concelhos das áreas metropolitanas de Lisboa e do Porto, e de zonas urbanas[69] dos concelhos de Lisboa e do Porto. O INE também publica os valores médios de avaliação bancária da habitação por natureza dos alojamentos ao nível de Portugal continental, NUTSII, NUTSIII e concelhos das áreas metropolitanas. O INE publica ainda, com a mesma periodicidade, os valores médios de avaliação bancária por natureza e tipologia dos alojamentos, ao nível de Portugal continental, NUTSII e áreas metropolitanas de Lisboa e do Porto.

Neste estudo empírico, foram utilizados os dados resultantes do IABH correspondentes às variáveis *valor da avaliação*, *área* e *concelho*. Também neste caso, foram consideradas todas as avaliações de habitações independentemente da sua natureza, tipologia ou ano de conclusão da obra de construção ou reconstrução do edifício.

7.2.2.3 Dados

No âmbito da aplicação prática foram utilizados dados reais obtidos através do IPTH realizado nos quatro trimestres de 2002 e nos três primeiros trimestres de 2003, e através do IABH realizado no último trimestre de 2001, nos quatro trimestres de 2002 e nos três primeiros trimestres de 2003. Foi com base nestes dados de painel disponibilizados pela Direcção Regional do Norte do INE, que se realizou este estudo. Não é possível apresentar nenhum excerto desses dados porque se encontram protegidos pelo segredo estatístico, nos termos do n.º 2 do artigo 5.º da Lei de Bases do Sistema Estatístico Nacional (Lei n.º 6/89 de 15 de Abril) (Assembleia da República, 1989).

Os parâmetros de interesse neste estudo são o preço médio de transacção da habitação, por metro quadrado, em cada NUTSIII. Estes parâmetros de interesse serão estimados, de forma directa ou indirecta, através dos dados recolhidos pelo IPTH. Por sua vez, a única variável auxiliar utilizada nos modelos de estimação em domínios de nível área é

[69] Segundo o INE, as zonas urbanas são formadas por um conjunto de freguesias de um dado concelho (INE, 1998).

o preço médio de avaliação bancária da habitação referente à mesma unidade. Os dados utilizados para calcular esse preço médio são recolhidos pelo IABH.

No âmbito dos modelos espaciais de estimação em domínios, a matriz de pesos espaciais utilizada foi formada com base em dados referentes à vizinhança por adjacência das NUTSIII de Portugal continental[70]. A estrutura de vizinhança foi definida da seguinte forma: $w_{ij}=1$, se a fronteira da NUTSIII *i* partilha pelo menos um ponto comum com a fronteira da NUTSIII *j*, e $w_{ij}=0$ em caso contrário, *i*,*j*=1, …, 28. Com base nesta estrutura, definiu-se uma matriz de pesos espaciais, $\mathbf{W}=\{w_{ij}^{*}\}$, com pesos estandardizados por linhas da forma $w_{ij}^{*}=w_{ij}/w_{i\bullet}$. Apesar desta escolha levar a que a matriz **W** não seja simétrica, ela garante a consistência interna. A matriz **W** encontra-se no apêndice 3. Este tipo de matriz de pesos foi utilizado por Singh *et al*. (2005), Petrucci e Salvati (2004*a*, 2004*b*, 2006), Chandra *et al*. (2007*a*, 2007*b*), Pratesi e Salvati (2004, 2005, 2008), entre outros, no contexto da estimação em pequenos domínios.

7.2.2.4 Trabalho preliminar

Antes da aplicação dos estimadores apresentados na secção 7.2.4, aos dados provenientes do IPTH e do IABH, foi efectuado algum trabalho de preparação dos dados, que consistiu na extracção de *outliers*[71]. Neste trabalho de extracção de *outliers*, foram excluídos separadamente os *outliers* presentes nos dados provenientes do IPTH e nos dados provenientes do IABH. Em ambos os casos, em primeiro lugar foram excluídos os *outliers* presentes nos valores das áreas de cada habitação e em segundo lugar foram excluídos os *outliers* presentes nos preços de transacção/avaliação por metro quadrado da habitação.

No que se refere à extracção dos *outliers* presentes nos valores das áreas, foram considerados *outliers* os registos com áreas inferiores ao limite inferior, ou superiores ao limite superior, definido para cada uma das tipologias dos fogos. Para cada uma

[70] O mapa das NUTSIII de Portugal continental pode ser encontrado no anexo 3.

[71] *Outlier* traduz-se para português por valor aberrante. Nesta tese, decidiu utilizar-se o termo em língua inglesa, por assim ser sugerido no Glossário Estatístico Inglês-Português da SPE e ABE (2007).

dessas tipologias, o limite inferior foi fixado como a área mínima aceitável definida pelo INE. Depois de extraídos os registos considerados *outliers* na aba esquerda da distribuição dos valores das áreas das habitações, o limite superior foi fixado para cada uma das tipologias através da seguinte expressão, $Q_3 + 3(Q_3 - Q_1)$, onde Q_1 e Q_3 representam os primeiro e terceiro quartis das áreas das habitações, respectivamente (*vide* apêndice 4).

Quanto à extracção dos *outliers* presentes nos preços de transacção e nos preços de avaliação bancária das habitações por metro quadrado, foram também considerados *outliers* os registos com preços por metro quadrado inferiores ao limite inferior, ou superiores ao limite superior, definido para cada uma das tipologias dos fogos. Para cada uma dessas tipologias, o limite inferior foi fixado como o preço por metro quadrado mínimo aceitável tido como referência pelo INE, actualizado anualmente de acordo com o aumento dos preços médios de avaliação bancária da habitação. Depois de extraídos os registos considerados *outliers* na aba esquerda das distribuições dos preços de transacção e de avaliação bancária por metro quadrado, o limite superior foi fixado para cada uma das tipologias através da seguinte expressão, $Q_3 + 6(Q_3 - Q_1)$ (*vide* apêndice 4). Foi-se menos exigente na identificação de *outliers* na aba direita das distribuições dos preços de transacção e de avaliação bancária do que na distribuição das áreas das habitações, porque nas distribuições de preços os *outliers* podem ser mais severos uma vez que dependem das condições de mercado, o que torna a amplitude de preços aceitável muito grande.

O limite inferior do preço médio de transacção por metro quadrado é sempre inferior ao limite inferior do preço médio de avaliação bancária por metro quadrado para todas as tipologias, pelo facto de se observar em Portugal continental uma sobreavaliação bancária do preço das habitações por metro quadrado, para ser possível aos compradores contraírem um maior montante de crédito, se for esse o caso. Em Portugal, o montante máximo de crédito que um comprador pode contrair para a compra de uma habitação depende, em geral, da idade do comprador, da duração do contrato de empréstimo à habitação, da modalidade do crédito, e do valor da avaliação bancária da habitação, quando este valor é inferior ou igual ao valor da escritura pública de aquisição. Note-se que estas condições podem sofrer pequenas alterações consoante a instituição bancária a que o comprador recorrer para a obtenção de crédito.

7.2.2.5 Análise exploratória de dados

Todos os modelos de estimação em domínios apresentados na revisão bibliográfica deste estudo, os quais são casos particulares do modelo linear misto, poderão ser utilizados para modelar os dados sobre preços de transacção da habitação. Contudo, alguns desses modelos irão permitir estimar o preço médio de transacção da habitação com melhores níveis de precisão do que outros. Existirão, eventualmente, até melhores modelos do ponto de vista da qualidade da estimação dos parâmetros de interesse do que os modelos que foram revistos no quarto capítulo. Desta forma, foi realizada uma análise exploratória de dados antes de se ter desenvolvido o quinto capítulo, no qual é proposto um novo modelo de estimação em pequenos domínios, tendo como objectivo a estimação do preço médio de transacção da habitação em Portugal.

Assim, nesta subsecção é apresentada uma análise exploratória dos dados amostrais disponíveis, com base no plano de sondagem que suporta a estimação dos parâmetros de interesse, tendo em consideração as recomendações de Littell *et al.* (2000) para a modelação de um determinado fenómeno através de um modelo linear misto. Segundo estes autores, a modelação de um fenómeno através de um modelo linear misto deve ser feita em quatro etapas, designadamente:

1. Modelação da estrutura média, através da especificação dos efeitos fixos;
2. Modelação dos efeitos aleatórios, através da especificação da estrutura de covariâncias;
3. Ajustamento do modelo, tendo em consideração essa estrutura de covariâncias;
4. Realização de inferências estatísticas baseadas nos resultados da etapa anterior.

Esta análise exploratória teve como objectivo a identificação de factores que possam ter um papel explicativo no comportamento médio da variável de interesse e a caracterização da estrutura de covariâncias subjacente ao quadro de dados. Por outras palavras, esta análise teve como objectivo identificar os traços principais que um modelo de estimação em pequenos domínios deve ter, de forma a assistir a estimação do preço médio de transacção da habitação em Portugal com os melhores níveis de precisão possíveis.

A identificação de factores que possam ter um papel explicativo no comportamento da variável de interesse consiste em especificar a estrutura dos efeitos fixos mais adequada ao quadro de dados. Por sua vez, a modelação da estrutura de covariâncias subjacente ao quadro de dados consiste em estudar a existência de autocorrelação temporal entre as observações referentes a um domínio particular (NUTSIII), a associação espacial entre as observações referentes a um determinado período de tempo, bem como a existência de heterogeneidade nas variâncias das observações relativas a diferentes períodos de tempo.

Com o objectivo de se diagnosticar possíveis especificações para a componente fixa e aleatória dos modelos de efeitos mistos, uma abordagem frequentemente recomendada na literatura (Verbeke e Molenbergs, 2000) é a seguinte: em primeiro lugar, ajustar modelos de regressão linear ao conjunto de dados disponível tendo em conta todas as variáveis explicativas, e em seguida, analisar os respectivos resíduos como forma de identificar possíveis especificações para os efeitos aleatórios. Nas referidas regressões lineares com termo independente, efectuadas no âmbito deste estudo, foi considerada como variável dependente as estimativas directas do preço médio de transacção em cada "domínio-período de tempo", e como variável auxiliar o preço médio de avaliação bancária da habitação referente à mesma unidade.

A. Selecção dos efeitos fixos

A primeira análise de dados consiste, portanto, em ajustar o seguinte modelo de regressão linear simples pelo método dos mínimos quadrados ordinários, o qual pode descrever a relação existente entre o preço médio de transacção da habitação e o preço médio de avaliação bancária da habitação:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}, \qquad (7.2.1)$$

onde $\varepsilon_{it} \sim iid\left(0;\sigma^2\right)$, i=1, ..., 28; t=1, ..., 7.

O coeficiente de determinação da regressão é aproximadamente igual a 0,65, o qual indica uma boa capacidade explicativa da variável auxiliar. O teste ao poder explicativo dos efeitos fixos é baseado nos seguintes ensaios de hipóteses: $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$, k=0, 1. Em ambos os ensaios (para k=0 e para k=1) a hipótese nula é

rejeitada, pelo que se conclui que os dados fornecem uma evidência clara à existência daqueles dois efeitos fixos (tabela 7.2.2).

Tabela 7.2.2: Teste à significância estatística dos efeitos fixos do modelo (7.2.1)

Efeito	***gl***	t_{obs}	***p***
β_0	194	-3,386	0,0008
β_1	194	19,052	0,0001

A natureza longitudinal do quadro de dados em estudo pode sugerir que a relação existente entre a variável de interesse e a variável explicativa seja descrita através de um modelo que apresenta diferentes declives para diferentes períodos de tempo:

$$y_{it} = \beta_{0t} + \beta_{1t} x_{it} + \varepsilon_{it}, \tag{7.2.2}$$

onde $\varepsilon_{it} \sim iid\left(0;\sigma^2\right)$, i=1, ..., 28; t=1, ..., 7.

Sendo desejável o ajustamento do modelo mais parcimonioso aos dados disponíveis, então deve testar-se a igualdade nos termos independentes e dos declives para os sete períodos de tempo. Os ensaios de hipóteses são, respectivamente, os seguintes: $H_0 : \beta_{k1} = ... = \beta_{k7}$ contra $H_1 : \exists \beta_{kt} \neq \beta_{kt'}, t, t' = 1,...,7$, k=0, 1. O teste F à igualdade dos termos independentes do modelo, cujos resultados são apresentados na tabela 7.2.3, indica que existe uma forte evidência estatística (valor-p=0,9395) para não se rejeitar a hipótese nula. Da mesma forma, o teste F à igualdade dos parâmetros β_{1t}, t=1, ..., 7, conduz à não rejeição da hipótese nula (valor-p=0,9323), pelo que se conclui pela igualdade dos declives para os sete períodos de tempo. Desta forma, um modelo com termo independente e declive comum para os vários períodos de tempo deverá ser o mais adequado para descrever a relação entre o preço médio de transacção e o preço médio de avaliação bancária da habitação.

Tabela 7.2.3: Teste à significância estatística dos efeitos fixos do modelo (7.2.2)

Efeito	***gl* (numerador)**	***gl* (denominador)**	F_{obs}	***p***
Tempo	6	182	0,29	0,9395
X	1	182	364,74	0,0001
X * tempo	6	182	0,31	0,9323

B. Teste aos efeitos aleatórios

Uma vez efectuado um estudo à especificação dos efeitos fixos, deve verificar-se a relevância dos efeitos aleatórios, a qual constitui uma questão fundamental no âmbito de modelos lineares mistos. A significância estatística da componente de variância dos efeitos aleatórios de domínio, no âmbito de um modelo linear misto, pode ser testada através de um teste *F* exacto proposto por Demidenko (2004, p. 137), cujo ensaio de hipóteses é o seguinte: $H_0 : \sigma_v^2 = 0$ contra $H_1 : \sigma_v^2 > 0$. A estatística-teste é dada por:

$$\frac{\left(S_{MQO} - S_{\min}\right)/(k-p)}{S_{\min}/(n-k)} \sim F_{(k-p;n-k)}, \qquad (7.2.3)$$

onde S_{MQO} é a soma dos quadrados dos resíduos sob a hipótese nula, ou seja, obtidos pelo método dos mínimos quadrados ordinários num modelo sem efeitos aleatórios (7.2.1), $S_{\min}$ é a soma dos quadrados dos resíduos resultantes da estimação de um modelo linear misto básico (4.2.3), ou seja, na presença de efeitos aleatórios, $k = r([\mathbf{X}\ \mathbf{Z}])$, p é o número de variáveis auxiliares (neste caso p=2), e n é a dimensão amostral (neste caso corresponde ao número de domínios de interesse em cada período de tempo, n=28).

Os resultados apresentados na tabela 7.2.4 confirmam a significância estatística da componente de variância dos efeitos aleatórios de domínio, no âmbito do modelo (4.2.3), para cada um dos períodos de tempo.

Tabela 7.2.4: Teste à significância estatística da componente de variância dos efeitos aleatórios do modelo (4.2.3), para cada período de tempo

t	***gl* (numerador)**	***gl* (denominador)**	F_{obs}	***p***
1	25	1	25.829.737	0,0000
2	25	1	2.254.889	0,0000
3	25	1	19.835.660	0,0000
4	25	1	12.870.782	0,0000
5	25	1	19.124.021	0,0000
6	25	1	38.787.322	0,0000
7	25	1	23.258.611	0,0000

C. Selecção da estrutura de covariância dos efeitos aleatórios

Uma vez confirmada a significância estatística da componente de variância dos efeitos aleatórios de domínio, é então altura de se despoletar uma análise de dados no sentido de identificar que configuração da estrutura de covariância poderá ser a mais adequada, ou seja, que matriz **G** deverá ser usada. A flexibilidade oferecida pelo procedimento MIXED do programa SAS na utilização e selecção de entre um vasto leque de estruturas de covariância, permite que seja incorporada na estimação do modelo a estrutura de covariância mais adequada aos dados disponíveis e aos objectivos do estudo, qualquer que ela seja. Uma descrição das estruturas de covariância mais utilizadas pode ser encontrada em Wolfinger (1993, 1996).

Dada a natureza seccional e cronológica do quadro de dados, é plausível que as observações de um dado domínio de interesse referentes a diferentes períodos de tempo, estejam correlacionadas temporalmente. Para além disso, é também aceitável que as observações de um dado domínio apresentem alguma associação com as observações de um domínio vizinho, dado que no contexto do problema prático em estudo as fronteiras dos domínios são arbitrárias.

Em primeiro lugar, vai efectuar-se um diagnóstico à existência de correlação temporal com base nos resíduos do modelo (7.2.1). Um bom instrumento que ajuda na decisão sobre que efeitos aleatórios incluir no modelo é um gráfico dos resíduos dos mínimos quadrados ao longo do tempo (Verbeke e Molenberghs, 2000). Na figura 7.2.1 são apresentadas as caixas de bigodes desses resíduos por período de tempo. Decidiu também analisar-se as caixas de bigodes desses resíduos por domínio, as quais estão apresentadas na figura 7.2.2.

Figura 7.2.1: Diagrama em caixa de bigodes dos resíduos relativos aos sete trimestres

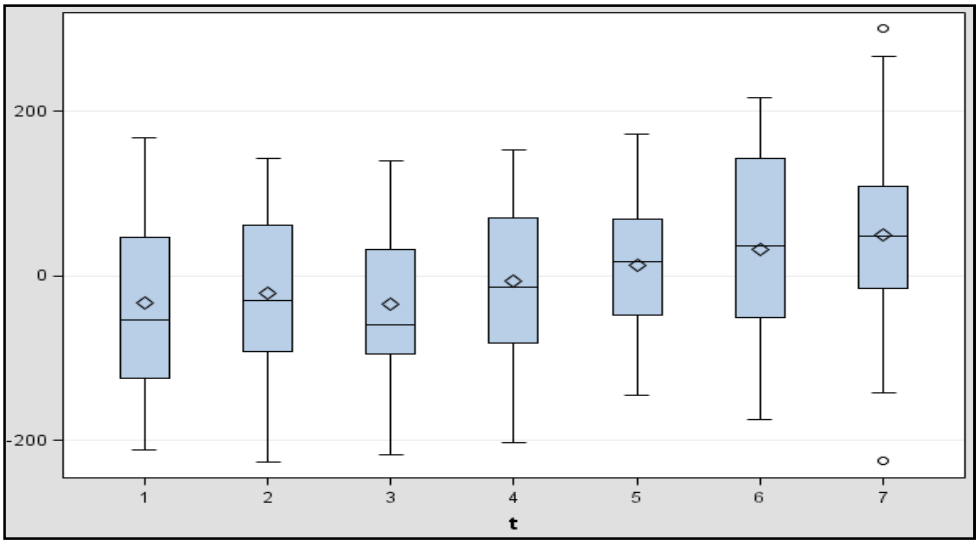
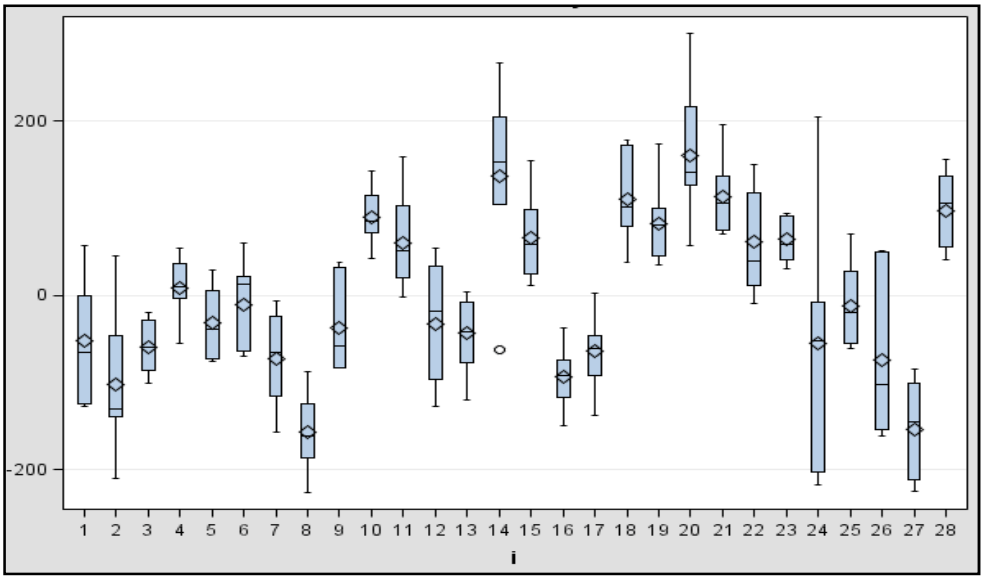


Figura 7.2.2: Diagrama em caixa de bigodes dos resíduos relativos às 28 NUTSIII



Na figura 7.2.1 pode observar-se uma tendência crescente no valor médio dos resíduos, bem como alguma homogeneidade na variabilidade dos resíduos ao longo do tempo. Esta figura parece então sugerir a existência de correlação temporal intra-domínio, bem como uma eventual homogeneidade nas variâncias dos resíduos para diferentes períodos de tempo, pelo que se propõe a utilização de modelos que envolvam efeitos aleatórios específicos de domínio-tempo com uma estrutura autoregressiva. A figura 7.2.2 evidencia uma clara heterogeneidade das variâncias dos resíduos para diferentes domínios. Não sendo parcimonioso incluir nos modelos efeitos aleatórios específicos de domínio com variâncias heterogéneas, então admite-se que:

(*i*) uma parte desta informação é devida ao erro de amostragem que varia ao longo dos domínios, sendo incluída no modelo através das variâncias dos erros da sondagem. Conforme se pode observar na tabela 7.4.2 da secção 7.4.2, onde são apresentadas as estimativas dos CV associadas ao estimador directo, os domínios que apresentam intervalos de variação dos resíduos com maior amplitude tendem a apresentar maiores estimativas dos respectivos CV, devido às mais elevadas variâncias;

(*ii*) outra parte dessa informação é incorporada no modelo através dos efeitos aleatórios específicos de domínio-tempo;

(*iii*) a parte remanescente será eventualmente incorporada no modelo através de efeitos aleatórios de domínio com uma estrutura de covariância espacial. É de salientar que se pode identificar na figura 7.2.2 um padrão espacial[72] nos resíduos, sendo mais pronunciado nos resíduos relativos às regiões norte (domínios 1 a 8), grande Lisboa e vale do Tejo (domínios 19 a 23), e Alentejo (domínios 24 a 27).

Note-se que a existência de autocorrelação entre observações referentes a um domínio particular é uma situação esperada em dados seccionais e cronológicos de preços da habitação. Seria, portanto, de esperar que duas observações relativas à mesma NUTSIII, referentes a períodos de tempo adjacentes, estivessem mais correlacionadas do que duas observações relativas ao mesmo domínio, mas referentes a períodos de tempo mais afastados. Esta evidência foi corroborada pelo teste de Durbin-Watson, o qual indicou a presença de autocorrelação nos resíduos[73], uma vez que o valor da estatística-teste obtida (DW=0,789) é inferior ao menor valor crítico (d_L). O coeficiente de autocorrelação de primeira ordem, calculado com base nos resíduos do método dos mínimos quadrados ordinários, foi de 0,37.

Note-se ainda que, a natureza longitudinal das estimativas directas do parâmetro de interesse (observações) pode levar, não só a que a correlação entre pares de observações relativas a um domínio particular dependa da amplitude do intervalo de tempo entre observações, mas também que essas observações apresentem diferentes variâncias para diferentes períodos de tempo. A este propósito, Littell *et al.* (2006) referem que a não

[72] O padrão espacial é identificado através da proximidade das caixas de bigodes.

[73] Os valores críticos do teste de Durbin-Watson para n=200 e um regressor excluindo a constante são d_L=1,758 e d_U=1,778 (nível de significância de 5%).

consideração dessa heterogeneidade nas variâncias aquando da estimação de um modelo quando ela existe, pode conduzir a inferências ineficientes ao nível dos efeitos fixos do modelo. Desta forma, decidiu realizar-se um teste de hipóteses à igualdade das variâncias do valor absoluto dos resíduos relativos a diferentes períodos de tempo. As hipóteses subjacentes ao teste de Levene são as seguintes: $H_0 : \sigma_1^2 = ... = \sigma_7^2$ contra $H_1 : \exists \sigma_t^2 \neq \sigma_{t'}^2 \, , t, t' = 1,...,7$. Os resultados do teste de Levene para a homogeneidade das variâncias são apresentados na tabela 7.2.5.

Tabela 7.2.5: Resultados do teste de Levene à homogeneidade das variâncias

Origem da variação	***gl***	**Soma de quadrados**	**Médias quadráticas**	F_{obs}	***p***
Entre períodos	6	9.319	1.553	0,1730	0,9838
Dentro dos períodos	189	1.696.533	8.976		
Total	195	1.705.852			

Uma vez que o teste de Levene apresenta um valor-*p*=0,9838, então conclui-se que não existe evidência estatística para a rejeição da hipótese nula da igualdade das variâncias. Desta forma, parece clara a necessidade de incorporar no modelo de estimação em domínios efeitos aleatórios específicos de domínio-tempo com uma estrutura de covariância autorregressiva, mas com variâncias homogéneas[74].

É agora altura de se efectuar um diagnóstico à existência de associação espacial, com base nos resíduos do modelo (7.2.1). Com o objectivo de se detectar um padrão espacial nos dados, foram calculadas estatísticas globais de associação espacial (*I* de Moran e *c* de Geary) e efectuados os respectivos testes de existência de associação espacial, bem como realizados testes de diagnóstico de associação espacial incorporada na estrutura de erro, modelados através de processos do tipo SAR (conforme subsecção 3.4.3.4).

As estatísticas globais de associação espacial, apresentadas na tabela 7.2.6, evidenciam uma associação espacial significativa e positiva no preço médio de transacção da habitação, em todos os períodos de tempo. Na tabela 7.2.6 são também apresentadas as estatísticas-teste e os respectivos valores-*p*, obtidos através da aproximação à distribuição normal e através de métodos de Monte Carlo.

[74] Note-se que, no âmbito da estimação em pequenos domínios, a utilização de estruturas de covariância heterogéneas, tem a desvantagem de exigir a estimação de um maior número de componentes de variância, podendo piorar desta forma a precisão das estimativas dos parâmetros de interesse.

Tabela 7.2.6: Estatísticas *I* de Moran e *c* de Geary, para cada período de tempo

t	**Estatística *I* de Moran**					**Estatística *c* de Geary**				
	I	**Normal**		**Monte Carlo**		***c***	**Normal**		**Monte Carlo**	
		Z	***p***	***Z***	***p***		***Z***	***p***	***Z***	***p***
1	0,299	2,93	0,0017	2,93	0,0017	0,557	-3,09	0,0010	-3,09	0,0010
2	0,318	3,10	0,0010	3,11	0,0009	0,537	-3,23	0,0006	-3,23	0,0006
3	0,251	2,52	0,0059	2,58	0,0049	0,496	-3,51	0,0002	-3,51	0,0002
4	0,245	2,46	0,0069	2,47	0,0068	0,578	-2,94	0,0016	-2,94	0,0016
5	0,338	3,27	0,0005	3,32	0,0004	0,455	-3,79	0,0001	-3,79	0,0001
6	0,309	3,02	0,0013	3,05	0,0011	0,493	-3,53	0,0002	-3,54	0,0002
7	0,291	2,86	0,0021	2,90	0,0019	0,475	-3,65	0,0001	-3,65	0,0001

Por último, os resultados do teste Burridge indicam a existência de associação espacial significativa incorporada na estrutura de erro, em todos os períodos de tempo, com excepção do período 1. Esses resultados encontram-se na tabela 7.2.7.

Tabela 7.2.7: Resultados do teste de Burridge, para cada período de tempo

t	T_B	***p***
1	-0,99	0,2454
2	2,38	0,0236
3	4,72	0,0000
4	-5,23	0,0000
5	2,51	0,0172
6	3,19	0,0024
7	5,27	0,0000

Desta forma, parece também clara a necessidade de incorporar no modelo que assiste a estimação, efeitos aleatórios específicos de domínio com uma estrutura de covariância espacial.

A aparente assimetria dos resíduos sugere que a hipótese da normalidade dos resíduos não se verifica exactamente (*vide* histograma e gráfico QQ no apêndice 5). De forma a verificar essa hipótese, foram efectuados os conhecidos testes de normalidade[75], cujos resultados são apresentados na tabela 7.2.8.

[75] Foram efectuados os testes de SW e de KS, porque *n*=196. Note-se que a estatística-teste W original (Shapiro e Wilk, 1965) é válida apenas para dimensões amostrais compreendidas entre 3 e 50 observações, mas Royston (1982) apresentou uma extensão desta estatística-teste para dimensões amostrais até 2.000 observações, através de uma aproximação à distribuição Normal. Esta extensão é utilizada pelo SAS. Por sua vez, é preferível aplicar o teste de KS a amostras grandes (Stephens, 1974).

Tabela 7.2.8: Resultados dos testes de normalidade

Teste	Estatística	*p*
Shapiro-Wilk (SW)	W=0,9097	0,0000
Kolmogorov-Smirnov (KS)	D=0,0772	0,0100

Ambos os resultados apresentados na tabela 7.2.8 conduzem à rejeição da hipótese da normalidade dos resíduos. Contudo, continua a admitir-se a hipótese da normalidade para todos os modelos especificados. De facto, pode esperar-se que desvios em relação à normalidade tenham apenas um pequeno impacto nas estimativas pontuais dos parâmetros de interesse por duas razões. Por um lado, porque as expressões do BLUP podem ser derivadas mesmo numa situação de não normalidade, e por outro lado, porque a estimação das componentes de variância não exige o pressuposto da normalidade pelo facto de ser efectuada através do método dos momentos. Na realidade, os desvios em relação à normalidade apenas podem ter um impacto mais significativo na estimação da terceira componente do EQMP dos EBLUP, uma vez que esta é a única componente que exige a hipótese da normalidade na sua estimação.

7.2.3 Desenho do estudo por simulação *design-based*

No âmbito da aplicação da metodologia, foi gerada uma pseudo-população finita de empresas a partir de uma amostra real conhecida. Posteriormente, foi efectuada uma simulação de Monte Carlo, na qual foi extraída em cada réplica (ou simulação) uma amostra aleatória independente da pseudo-população.

A realização de estudos empíricos, do tipo *design-based* ou *model-based*, baseados em simulação de Monte Carlo é cada vez mais frequente, uma vez que esta ferramenta permite avaliar as propriedades estatísticas de estimadores, que de outra forma dificilmente seriam avaliadas, devido à impossibilidade frequente de derivação de expressões explícitas da sua variância e do seu valor esperado. Isto deve-se ao facto de em estudos empíricos com dados simulados ser possível comparar as estimativas com os verdadeiros valores dos parâmetros, permitindo desta forma perceber-se como os estimadores se poderão comportar em aplicações reais, nas quais os verdadeiros valores dos parâmetros são desconhecidos.

No contexto da estimação em pequenos domínios, alguns trabalhos que comparam estimadores por métodos de Monte Carlo cujas amostras foram extraídas de pseudo-populações finitas geradas a partir de censos devem-se a Falorsi *et al.* (1994) e a Ghosh *et al.* (1996), e de pseudo-populações geradas a partir da replicação de amostras reais devem-se a Singh *et al.* (1994, 2005), Falorsi *et al.* (1999), Coelho (2000), Lehtonen *et al.* (2003), Chandra e Chambers (2006*a*, 2006*b*, 2006*c*), Fabrizi *et al.* (2007), Pratesi e Salvati (2008), entre outros.

7.2.3.1 Geração de uma pseudo-população

Para realizar este estudo por simulação foi utilizada uma pseudo-população de A=4.659 empresas. Esta pseudo-população finita, U^*, foi gerada a partir da replicação das empresas de uma amostra real, s, do IPTH. Uma vez que neste inquérito as dimensões amostrais de unidades primárias (empresas) não são constantes ao longo do tempo (*vide* tabela 7.2.1), então decidiu formar-se uma nova amostra com dimensão igual à dimensão média das amostras recolhidas nas sete vagas em que foi aplicado o IPTH (458 empresas). A constituição desta nova amostra de a=458 empresas foi efectuada através de uma amostragem aleatória simples de todas as empresas que fizeram parte de pelo menos uma das sete amostras.

A pseudo-população, U^*, foi então gerada através da replicação de cada uma das empresas, g, de uma amostra de a=458 empresas, proporcionalmente ao inverso das suas probabilidades de inclusão de primeira ordem[76], na situação de uma sondagem por conglomerados previamente estratificados. Todas as réplicas de uma dada empresa partilham as mesmas probabilidades de inclusão e as mesmas unidades secundárias, ou seja, as mesmas transacções de habitações.

[76] O inverso da probabilidade de inclusão de primeira ordem também se pode designar por coeficiente de extrapolação. O arredondamento do coeficiente de extrapolação foi efectuado de forma aleatória, mas tendo em consideração a sua parte decimal. Considere-se $\left(1/\pi_g\right)^*$ uma variável aleatória que representa o valor arredondado. As probabilidades do arredondamento ter sido efectuado por defeito e por excesso são dadas, respectivamente, por: $P\left[\left(\frac{1}{\pi_g}\right)^* = \text{int}\left(\frac{1}{\pi_g}\right)\right] = 1 - \frac{1}{\pi_g} + \text{int}\left(\frac{1}{\pi_g}\right)$ e $P\left[\left(\frac{1}{\pi_g}\right)^* = \text{int}\left(\frac{1}{\pi_g}\right) + 1\right] = \frac{1}{\pi_g} - \text{int}\left(\frac{1}{\pi_g}\right)$.

A pseudo-população foi posteriormente arquivada numa base de sondagem com as seguintes variáveis: *código de NUTSIII*, *código de estrato* e *código da empresa de mediação imobiliária*. Simultaneamente, foram criadas sete bases de dados constituídas pelas transacções de habitações efectuadas em cada trimestre pelas empresas pertencentes à pseudo-população.

7.2.3.2 Descrição das simulações

A partir da pseudo-população finita (considerada fixa) de empresas, foram extraídas L=1.000 amostras aleatórias independentes, utilizando um plano de sondagem semelhante ao que foi utilizado no IPTH – sondagem aleatória por conglomerados previamente estratificados, sem reposição. Contudo, admitiu-se que cada uma dessas amostras aleatórias tinha uma dimensão de a=229 empresas (com metade da dimensão da amostra real) com o objectivo de se testar a reacção das propriedades dos estimadores propostos, quando estes são utilizados em problemas de estimação em verdadeiros pequenos domínios, ou seja, quando a maior parte dos domínios de interesse apresenta dimensões amostrais muito pequenas ou mesmo nulas.

Para além disso, admitiu-se também que o IPTH foi implementado segundo um painel "puro" e que a população foi estratificada pelo cruzamento das variáveis *NUTSIII* e *concelho* (apenas para as áreas Metropolitanas de Lisboa e do Porto). O facto de se ter assumido a implementação do IPTH segundo um painel "puro", significa que no primeiro período de observação foi seleccionada uma amostra aleatória de a=229 empresas, e que foram observadas todas as transacções efectuadas por essa amostra (fixa) de empresas nas sete vagas do inquérito, ou seja, ao longo dos sete trimestres.

Note-se que, apesar do plano de sondagem utilizado neste estudo por simulação diferir ligeiramente do plano de sondagem original, pelo facto de não considerar a rotação da amostra, ele continua a apresentar a sua característica principal, ou seja, é um plano de sondagem informativo. Para além disso, foi considerada uma redução substancial das dimensões amostrais de forma a tornar o problema de estimação em pequenos domínios mais realista. As dimensões pseudo-populacional e amostral de empresas em cada estrato, bem como as respectivas taxas de sondagem encontram-se no apêndice 6.

Admitiu-se que o IPTH tinha sido implementado segundo um painel, porque se acreditou não existirem vantagens em considerar que o IPTH tivesse sido implementado segundo um inquérito longitudinal com rotação, uma vez que o objectivo consiste em fazer inferência sobre as unidades secundárias (transacções) através da utilização de modelos de nível área que utilizam informação espacial/seccional e/ou cronológica. Este tipo de modelos permite fazer a estimação do preço médio de transacção da habitação para qualquer domínio, independentemente de se terem observado ou não transacções nesse domínio, e independentemente das empresas que realizaram transacções nesse domínio terem pertencido à amostra em vagas anteriores.

Para cada amostra (réplica) foram calculadas estimativas directas do preço médio de transacção da habitação em cada NUTSIII de Portugal continental através do estimador directo, e estimativas indirectas através do estimador sintético pelo quociente, do estimador sintético pela regressão e de alguns estimadores combinados revistos no quarto capítulo, e dos estimadores propostos nos capítulos quinto e sexto.

7.2.4 Estimação

7.2.4.1 Introdução

Com base nos dados disponíveis, pretende produzir-se estimativas, com precisão adequada, do preço médio de transacção da habitação por metro quadrado de área útil para os sete trimestres, ao nível de desagregação NUTSIII. Para tal, será avaliada a qualidade de diversos estimadores (directo, sintéticos e combinados) com o fim de se identificar aquele que permite produzir as supracitadas estimativas com os melhores níveis de enviesamento e de precisão. Será dado, naturalmente, maior destaque aos estimadores combinados propostos.

O parâmetro de interesse é a média da variável *preço de transacção da habitação*[77] (por metro quadrado de área útil) ao nível de NUTSIII. A variável auxiliar utilizada neste

[77] O valor da variável *preço de transacção da habitação* (por metro quadrado de área útil) foi calculado pelo quociente entre o valor da variável *valor de transacção* e o valor da variável *área útil*.

estudo[78] é o *preço de avaliação bancária da habitação*[79] (também por metro quadrado). A impossibilidade de ligação ao nível individual entre os dados provenientes das amostras do IPTH e a informação auxiliar relativa à avaliação bancária, limitou fortemente a natureza dos métodos de estimação que poderiam ser usados neste contexto. Desta forma, foram utilizados apenas modelos de nível área de estimação em pequenos domínios que relacionam a média da variável de interesse em cada NUTSIII (domínio de interesse) com a média da variável auxiliar nesse domínio específico.

A utilização dessa variável auxiliar justifica-se pelo facto do preço médio de transacção da habitação ao nível de NUTSIII estar directamente correlacionado com o preço médio de avaliação bancária da habitação, ao mesmo nível de agregação. Tendo-se verificado que essa correlação é ligeiramente superior quando existe um desfasamento temporal entre esses dois preços médios, decidiu utilizar-se a informação auxiliar desfasada de um trimestre em relação ao momento de inferência. A tabela 7.2.9 apresenta os valores dos coeficientes de correlação linear entre o preço médio de transacção da habitação e o preço médio de avaliação bancária da habitação ao nível de NUTSIII, sem desfasamento e com desfasamento temporal de um trimestre, para cada um dos sete trimestres.

Tabela 7.2.9: Coeficientes de correlação entre o preço médio de transacção da habitação e o preço médio de avaliação bancária da habitação, ao nível de NUTSIII

Trimestre	Sem desfasamento temporal	Com desfasamento temporal de um trimestre
1	0,819	0,851
2	0,844	0,849
3	0,841	0,846
4	0,848	0,852
5	0,857	0,869
6	0,814	0,820
7	0,845	0,876

Os resultados anteriores não são surpreendentes, uma vez que é do conhecimento geral que as avaliações bancárias das habitações em Portugal continental são efectuadas pelas instituições bancárias nos meses que antecedem a transacção da habitação.

[78] O *preço de avaliação bancária da habitação* poderá, com toda a certeza, não ser a única variável explicativa do preço de transacção de uma habitação. Neste estudo foi utilizada apenas a variável *preço de avaliação bancária da habitação*, dada a inexistência de outra informação auxiliar disponível.

[79] O valor da variável *preço de avaliação bancária da habitação* foi calculado pelo quociente entre o valor da variável *valor de avaliação* e o valor da variável *área*.

Para se proceder à estimação do preço médio de transacção da habitação, ao nível de NUTSIII, através de estimadores *design-based*, foi necessário considerar-se a estimação em domínios numa sondagem por conglomerados previamente estratificados.

Nas secções seguintes são apresentados todos os estimadores da média da variável de interesse utilizados neste estudo por simulação *design-based*, bem como os estimadores da variância ou do EQM associados a esses estimadores. Em primeiro lugar, são apresentados os estimadores tradicionais: estimador directo pós-estratificado, estimador sintético pelo quociente, estimador sintético pela regressão e estimador combinado com pesos dependentes dos dados. Em segundo lugar, são apresentados os estimadores EBLUP: estimador seccional de Fay-Herriot, estimador espacial de Nicola Salvati, estimador temporal de Rao-Yu e os dois estimadores propostos (estimador assistido por um modelo espaciotemporal sem e com restrições).

É altura de salientar que uma classe de estimadores indirectos que, *a priori*, poderia parecer potencialmente interessante para responder aos objectivos de estimação do preço médio de transacção da habitação ao nível de NUTSIII, garantindo simultaneamente a consistência interna, é a classe dos estimadores directos modificados. Estes estimadores são normalmente denominados por estimadores directos modificados, apesar de constituírem uma das classes da família dos estimadores indirectos, porque garantem o não enviesamento aproximado do ponto de vista do plano de sondagem. A classe de estimadores directos modificados poderia parecer adequada à partida por duas razões: em primeiro lugar, porque os estimadores desta classe utilizam informação relativa à variável de interesse de fora do domínio de estudo, melhorando desta forma a precisão relativamente ao estimador directo; e, em segundo lugar, porque estes estimadores verificam a propriedade da consistência interna.

Contudo, nenhum dos estimadores pertencentes à referida classe pode ser utilizado no âmbito deste estudo porque: (*i*) são desconhecidas as dimensões populacionais dos domínios de interesse, N_i - número total de transacções de habitações efectuadas na *i*-ésima NUTSIII, o que inviabiliza o uso de um estimador directo modificado pós-estratificado; e (*ii*) não é possível fazer-se a ligação, ao nível individual, entre as observações amostrais da variável de interesse e as observações da variável auxiliar, impossibilitando desta forma o uso de estimadores directos modificados pelo quociente e pela regressão.

Na exposição seguinte, utiliza-se a notação *ti* para representar o *i*-ésimo domínio no momento *t*. Os domínios correspondem às 28 NUTSIII de Portugal continental, *i*=1, ..., 28, enquanto os momentos de tempo correspondem aos sete trimestres em que foram realizadas as vagas do IPTH, *t*=1, ..., 7, ou seja, aos quatro trimestres do ano de 2002 e aos três primeiros trimestres do ano de 2003. Em todo o caso, é de salientar que todos os estimadores que não tiram partido da informação temporal são utilizados de forma independente para calcular estimativas para cada período de tempo. Na exposição seguinte, considere-se também que $\mu_{x,t-1,i}$ representa o preço médio de avaliação bancária da habitação por metro quadrado na *i*-ésima NUTSIII no trimestre (*t*-1), e que $\boldsymbol{\mu}_{x,t-1,i} = \left(1, \mu_{x,t-1,i}\right)'$ e $\mathbf{X}_{t-1} = col_{1\le i\le m}\left(\boldsymbol{\mu}'_{x,t-1,i}\right)$ (*i*=1, …, 28; *t*=1, …, 7).

7.2.4.2 Estimadores tradicionais da média

A. Estimador directo pós-estratificado

No caso de uma sondagem por conglomerados previamente estratificados, o estimador directo pós-estratificado da média no *i*-ésimo domínio no momento *t*, é dado por:

$$\hat{\mu}_{ti}^{dir} = \frac{\hat{\tau}_{ti,HT}}{\hat{N}_{ti}} = \frac{\sum_{h=1}^{H}\sum_{g\in s_{Gh}}\sum_{j\in U_{hg}\cap ti}\frac{y_{thgj}}{\pi_{thgj}}}{\sum_{h=1}^{H}\sum_{g\in s_{Gh}}\sum_{j\in U_{hg}\cap ti}\frac{1}{\pi_{thgj}}} = \frac{\sum_{h=1}^{H}\sum_{g\in s_{Gh}}\sum_{j\in U_{hg}\cap ti}\frac{A_h}{a_h}y_{thgj}}{\sum_{h=1}^{H}\sum_{g\in s_{Gh}}\sum_{j\in U_{hg}\cap ti}\frac{A_h}{a_h}} = \frac{\sum_{h=1}^{H}\frac{A_h}{a_h}\sum_{g\in s_{Gh}}\tau_{tgi}}{\sum_{h=1}^{H}\frac{A_h}{a_h}\sum_{g\in s_{Gh}}N_{tgi}}, \qquad (7.2.4)$$

onde a_h e A_h representam, respectivamente, o número de empresas na amostra e na população do estrato *h*; y_{hgj} é o preço da *j*-ésima transacção de habitação por metro quadrado efectuada pela *g*-ésima empresa pertencente ao estrato *h*; τ_{tgi} é o somatório dos preços de todas as transacções efectuadas pela *g*-ésima empresa pertencente ao estrato *h*, na *i*-ésima NUTSIII no trimestre *t*; e N_{tgi} representa o número total de transacções de habitações efectuadas pela *g*-ésima empresa pertencente ao estrato *h*, na *i*-ésima NUTSIII no trimestre *t* (*g*=1, …, 229; *h*=1, …, 51; *i*=1, …, 28; *t*=1, …, 7). Este estimador directo não utiliza qualquer tipo de informação auxiliar.

Uma vez que o estimador directo (7.2.4) é um estimador aproximadamente centrado, então a medição da sua incerteza é efectuada através do cálculo de uma estimativa da

variância desse estimador. Dada a complexidade do plano de sondagem subjacente, não existe um estimador exacto da sua variância, sendo necessário usar uma aproximação. Decidiu, então, utilizar-se o método delta, *i.e.*, baseado em desenvolvimentos em séries de Taylor, para obter um estimador aproximado da variância do estimador directo. Um estimador da variância *design-based* do estimador directo (7.2.4) é dado por (Woodruff, 1971):

$$\hat{V}_d\left(\hat{\mu}_{ti}^{dir}\right)=\sum_{h=1}^{H}\hat{V}_h\left(\hat{\mu}_{ti}^{dir}\right), \tag{7.2.5}$$

onde $\hat{V}_h\left(\hat{\mu}_{ti}^{dir}\right)=\frac{a_{th}\left(1-f_{th}\right)}{a_{th}-1}\sum_{g=1}^{a_h}\left(r_{thg\bullet}-\bar{r}_{th\bullet\bullet}\right)^2$ quando $a_{th}>1$, na qual $r_{th\bullet\bullet}=\frac{1}{a_{th}}\sum_{g=1}^{a_h}r_{thg\bullet}$, $r_{thg\bullet}=\frac{1}{v_{t\bullet\bullet\bullet}}\sum_{j=1}^{n_{hg}}v_{thgj}\left(z_{thgj}-\hat{\mu}_{ti}^{dir}\right)$ e $v_{t\bullet\bullet\bullet}=\sum_{h=1}^{H}\sum_{g=1}^{a_h}\sum_{j=1}^{n_{hg}}v_{thgj}$. Tem-se ainda que $z_{thgj}=y_{thgj}I_{ti}(h,g,j)$ e $v_{thgj}=\pi_{thgj}^{-1}I_{ti}(h,g,j)$, onde $I_{ti}(h,g,j)$ é uma variável indicatriz, tal que $I_{ti}(h,g,j)=1$ se a observação (*h*,*g*,*j*) pertence ao domínio *i* no período *t* e $I_{ti}(h,g,j)=0$ em caso contrário. Quando $a_{th}=1$, então $\hat{V}_h\left(\hat{\mu}_{ti}^{dir}\right)$ é omisso se $a_{th'}=1$ para *h'*=1, …, *H*, e $\hat{V}_h\left(\hat{\mu}_{ti}^{dir}\right)=0$ se $a_{th'}>1$ para algum $1 < h' < H$.

B. Estimador sintético pelo quociente

No âmbito do mesmo plano de sondagem, um estimador sintético pelo quociente da média no *i*-ésimo domínio no momento *t*, deduzido a partir do estimador proposto por Singh e Tessier (1976), é dado por[80]:

[80] No âmbito da estimação baseada num estimador sintético pelo quociente, decidiu utilizar-se o estimador apresentado em vez do conhecido estimador sintético pelo quociente separado (utilizado no contexto de sondagens estratificadas ou pós-estratificadas), pelas seguintes razões: (*i*) não existe informação que permita estratificar a variável auxiliar por *valor de volume de negócios* (477 estratos); (*ii*) quando a estratificação é definida pelas variáveis *NUTSIII* e *concelhos das áreas metropolitanas*, então a maioria dos domínios coincide com os estratos (51 estratos); e, (*iii*) principalmente, porque o estimador utilizado verifica a consistência interna ao nível de NUTSII, ao contrário do estimador sintético pelo quociente separado.

$$\hat{\mu}_{ti}^{\sin Q} = \mu_{x,t-1,i} \frac{\hat{\mu}_{te,dir}}{\mu_{x,t-1,e}} = \frac{\mu_{x,t-1,i}}{\mu_{x,t-1,e}} \frac{\hat{\tau}_{te,HT}}{\hat{N}_{te}} = \frac{\mu_{x,t-1,i}}{\mu_{x,t-1,e}} \frac{\sum_{h=1}^{H} \frac{A_h}{a_h} \sum_{g \in s_{Gh}} \tau_{tge}}{\sum_{h=1}^{H} \frac{A_h}{a_h} \sum_{g \in s_{Gh}} N_{tge}}, \quad (7.2.6)$$

onde *e* representa uma área que contém um conjunto de domínios, sendo neste caso definida como uma NUTSII; $\mu_{x,t-1,e}$ é o preço médio de avaliação bancária da habitação na *e*-ésima NUTSII no trimestre (*t*-1); τ_{tge} é o somatório dos preços de todas as transacções efectuadas pela *g*-ésima empresa pertencente ao estrato *h*, na *e*-ésima NUTSII no trimestre *t*; e N_{tge} representa o número total de transacções de habitações efectuadas pela *g*-ésima empresa pertencente ao estrato *h*, na *e*-ésima NUTSII no trimestre *t* (*g*=1, …, 229; *h*=1, …, 51; *e*=1, …, 5; *i*=1, …, 28; *t*=1, …, 7).

É de salientar que este estimador verifica a consistência interna, ou seja, as estimativas directas do preço médio de transacção da habitação produzidas ao nível de NUTSII coincidem com o resultado do agrupamento das estimativas sintéticas desse preço médio produzidas ao nível de NUTSIII.

O estimador sintético pelo quociente (7.2.6) é enviesado, sendo o seu enviesamento dado por $B\left(\hat{\mu}_{ti}^{\sin Q}\right) \approx \mu_{x,t-1,i}\left(\frac{\mu_t}{\mu_{x,t-1}} - \frac{\mu_{ti}}{\mu_{x,t-1,i}}\right)$. Contudo, não é possível estimar este enviesamento a não ser que a população seja conhecida[81]. Um estimador da variância *design-based* do estimador sintético pelo quociente (7.2.6) é dado por:

$$\hat{V}_d\left(\hat{\mu}_{ti}^{\sin Q}\right) = \left(\frac{\mu_{x,t-1,i}}{\mu_{x,t-1,e}}\right)^2 \hat{V}_d\left(\hat{\mu}_{te}^{dir}\right), \quad (7.2.7)$$

onde $\hat{V}_d\left(\hat{\mu}_{te}^{dir}\right)$ é obtido da mesma forma que (7.2.5).

[81] No entanto, no âmbito deste estudo empírico, o enviesamento do estimador sintético pelo quociente será avaliado do ponto de vista da amostragem repetida.

C. Estimador sintético pela regressão

Um estimador sintético pela regressão é dado pela parte sintética do estimador combinado assistido pelo modelo de Fay-Herriot (4.2.3), com componente de variância estimada pelo método dos momentos de Prasad-Rao (4.2.13):

$$\hat{\mu}_{ti}^{\sin R} = \boldsymbol{\mu}'_{x,t-1,i}\hat{\boldsymbol{\beta}}_t = \hat{\beta}_{t1} + \mu_{x,t-1,i}\hat{\beta}_{t2}\,, \tag{7.2.8}$$

onde $\hat{\boldsymbol{\beta}}_t = \left(\hat{\beta}_{t1}, \hat{\beta}_{t2}\right)' = \left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i}\boldsymbol{\mu}'_{x,t-1,i}\left(\hat{\sigma}_{u,t}^2 + \hat{\sigma}_{\varepsilon,ti}^2\right)^{-1}\right]^{-1}\left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i}\hat{\mu}_{ti}^{dir}\left(\hat{\sigma}_{u,t}^2 + \hat{\sigma}_{\varepsilon,ti}^2\right)^{-1}\right]$ e $\hat{\mu}_{ti}^{dir}$ é uma estimativa directa do preço médio de transacção da habitação na *i*-ésima NUTSIII no trimestre *t*, calculada através de (7.2.4) (*i*=1, …, 28; *t*=1, …, 7). Este estimador é normalmente denominado na literatura por estimador BLUP sintético.

À semelhança do estimador (7.2.6), também o estimador sintético pela regressão (7.2.8) é enviesado, sendo o seu enviesamento dado por $B\left(\hat{\mu}_{ti}^{\sin R}\right) \approx \boldsymbol{\mu}'_{x,t-1,i}\boldsymbol{\beta}_t - \mu_{ti}$. Também neste caso só seria possível estimar o enviesamento se a população fosse conhecida[82]. Um estimador da variância *model-based* do estimador sintético pela regressão (7.2.8) é dado por:

$$\hat{V}\left(\hat{\mu}_{ti}^{\sin R}\right) = \boldsymbol{\mu}'_{x,t-1,i}\hat{V}\left(\hat{\boldsymbol{\beta}}_t\right)\boldsymbol{\mu}'_{x,t-1,i}\,, \tag{7.2.9}$$

onde $\hat{V}\left(\hat{\boldsymbol{\beta}}_t\right) = \left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i}\boldsymbol{\mu}'_{x,t-1,i}\left(\hat{\sigma}_{u,t}^2 + \hat{\sigma}_{\varepsilon,ti}^2\right)^{-1}\right]^{-1}$.

D. Estimador combinado com pesos dependentes dos dados

Uma forma natural de equilibrar o enviesamento potencial do estimador sintético com a instabilidade do estimador directo, procurando desta forma evitar que a qualidade do estimador do parâmetro de interesse fique totalmente dependente da veracidade do modelo postulado, consiste em utilizar um estimador combinado definido como uma média ponderada desses dois estimadores. Um estimador combinado com pesos

[82] No âmbito deste estudo empírico, o enviesamento do estimador sintético pela regressão também será avaliado do ponto de vista da amostragem repetida.

dependentes dos dados[83], para a média no *i*-ésimo domínio no momento *t*, pode ser escrito como:

$$\hat{\mu}_{ti}^{com} = \hat{\gamma}_{ti}\hat{\mu}_{ti}^{dir} + (1-\hat{\gamma}_{ti})\hat{\mu}_{ti}^{\sin R}, \tag{7.2.10}$$

onde $\hat{\gamma}_{ti} = 1 - \dfrac{\hat{V}_d(\hat{\mu}_{ti}^{dir})}{(\hat{\mu}_{ti}^{\sin R} - \hat{\mu}_{ti}^{dir})^2}$, verificando $0 \le \hat{\gamma}_{ti} \le 1$, e $\hat{\mu}_{ti}^{dir}$ e $\hat{\mu}_{ti}^{\sin R}$ são estimativas do preço médio de transacção da habitação, calculadas através de (7.2.4) e (7.2.8), respectivamente (*i*=1, …, 28; *t*=1, …, 7). Por último, um estimador *design-based* da variância do estimador combinado (7.2.10) é dado por:

$$\hat{V}(\hat{\mu}_{ti}^{com}) = \hat{\gamma}_{ti}^2\hat{V}_d(\hat{\mu}_{ti}^{dir}) + (1-\hat{\gamma}_{ti})^2\hat{V}(\hat{\mu}_{ti}^{\sin R}). \tag{7.2.11}$$

onde $\hat{V}_d(\hat{\mu}_{ti}^{dir})$ e $\hat{V}(\hat{\mu}_{ti}^{\sin R})$ são obtidas a partir de (7.2.5) e (7.2.9), respectivamente.

7.2.4.3 Estimadores EBLUP da média

A. Estimador seccional de Fay-Herriot

O primeiro estimador combinado, obtido pela metodologia EBLUP[84], é assistido pelo modelo de Fay-Herriot (4.2.3) com componente de variância estimada pelo método dos momentos de Prasad-Rao (4.2.13). Este estimador é dado por:

$$\hat{\mu}_{ti}^{FH} = \hat{\gamma}_{ti}\hat{\mu}_{ti}^{dir} + (1-\hat{\gamma}_{ti})\mathbf{\mu}'_{x,t-1,i}\hat{\beta}_t, \tag{7.2.12}$$

[83] No âmbito da estimação baseada em estimadores combinados tradicionais, decidiu utilizar-se o estimador apresentado em vez de um estimador combinado com pesos dependentes da dimensão da amostra, propostos por Drew *et al*. (1982) e Särndal e Hidiroglou (1989), porque os pesos usados neste último estimador exigem o conhecimento das dimensões populacionais nos domínios de interesse, as quais são desconhecidas neste estudo.

[84] Tal como foi referido na secção 2.6.3, existe um vasto conjunto de estimadores combinados com pesos dependentes dos dados obtidos através da metodologia EBLUP. Os principais estimadores desta família foram apresentados no capítulo quarto. Para facilitar a distinção dos estimadores combinados com pesos dependentes dos dados utilizados no âmbito deste estudo empírico, decidiu denominar-se o conjunto dos combinados com pesos dependentes dos dados obtidos através da metodologia EBLUP simplesmente por estimadores EBLUP.

onde $\hat{\boldsymbol{\beta}}_t = \left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i} \boldsymbol{\mu}'_{x,t-1,i} \left(\hat{\sigma}^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)^{-1}\right]^{-1} \left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i} \hat{\mu}^{dir}_{ti} \left(\hat{\sigma}^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)^{-1}\right]$; $\hat{\mu}^{dir}_{ti}$ é uma estimativa directa do preço médio de transacção da habitação na *i*-ésima NUTSIII no trimestre *t*, calculada através de (7.2.4); e $\hat{\gamma}_{ti} = \hat{\sigma}^2_{u,t} / \left(\hat{\sigma}^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)$ é um ponderador (*i*=1, …, 28; *t*=1, …, 7). Um estimador analítico do EQMP do estimador (7.2.12) é dado por:

$$eqmp\left(\hat{\mu}^{FH}_{ti}\right) = g_{1ti}\left(\hat{\sigma}^2_u\right) + g_{2ti}\left(\hat{\sigma}^2_u\right) + 2g_{3ti}\left(\hat{\sigma}^2_u\right), \tag{7.2.13}$$

onde $g_{1ti}\left(\hat{\sigma}^2_u\right) = \hat{\gamma}_{ti}\hat{\sigma}^2_{\varepsilon,ti}$; $g_{2ti}\left(\hat{\sigma}^2_u\right) = \left(1-\hat{\gamma}_{ti}\right)^2 \boldsymbol{\mu}'_{x,t-1,i}\left[\sum_{i=1}^{m} \boldsymbol{\mu}_{x,t-1,i}\boldsymbol{\mu}'_{x,t-1,i}\left(\hat{\sigma}^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)^{-1}\right]^{-1} \boldsymbol{\mu}_{x,t-1,i}$;

$g_{3ti}\left(\sigma^2_u\right) = \left[\sigma^4_{\varepsilon,ti}\left(\sigma^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)^{-3}\right] \overline{V}\left(\hat{\sigma}^2_{u,t}\right)$; e $\overline{V}\left(\hat{\sigma}^2_{u,t}\right) \approx 2m^{-2}\sum_{i=1}^{m}\left(\hat{\sigma}^2_{u,t} + \hat{\sigma}^2_{\varepsilon,ti}\right)^2$.

B. Estimador espacial de Nicola Salvati

O segundo estimador combinado, que tira partido da informação espacial, é assistido pelo modelo de Salvati (4.5.4), com componente de variância estimada por uma extensão do método III de Henderson, (5.9.2), e com coeficiente de associação espacial conhecido[85], ϕ=0,29. Este estimador é dado por:

$$\hat{\mu}^{NS}_{ti} = \boldsymbol{\mu}'_{x,t-1,i}\hat{\boldsymbol{\beta}}_t + \mathbf{m}'_{ti}\hat{\boldsymbol{\Lambda}}_t\left(\hat{\boldsymbol{\mu}}^{dir}_t - \mathbf{X}_{t-1}\hat{\boldsymbol{\beta}}_t\right), \tag{7.2.14}$$

onde $\hat{\boldsymbol{\beta}}_t = \left(\mathbf{X}'_{t-1}\hat{\mathbf{V}}^{-1}_t\mathbf{X}_{t-1}\right)^{-1}\mathbf{X}'_{t-1}\hat{\mathbf{V}}^{-1}_t\hat{\boldsymbol{\mu}}^{dir}_t$, $\hat{\boldsymbol{\Lambda}}_t = \hat{\sigma}^2_{u,t}\hat{\mathbf{B}}^{-1}_t\hat{\mathbf{V}}^{-1}_t$, $\mathbf{X}_{t-1} = col_{1\le i\le m}\left(\boldsymbol{\mu}'_{x,t-1,i}\right)$ e $\hat{\boldsymbol{\mu}}^{dir}_t = col_{1\le i\le m}\left(\hat{\mu}^{dir}_{ti}\right)$ (*i*=1, …, 28; *t*=1, …, 7). Um estimador analítico do EQMP do estimador (7.2.14) é dado por:

$$eqmp\left(\hat{\mu}^{NS}_{ti}\right) = g_{1ti}(\hat{\sigma}^2_u) + g_{2ti}(\hat{\sigma}^2_u) + 2g_{3ti}(\hat{\sigma}^2_u), \tag{7.2.15}$$

onde $g_{1ti}\left(\hat{\sigma}^2_u\right) = \hat{\sigma}^2_{u,t}\boldsymbol{\varsigma}_{tii} - \hat{\sigma}^2_{u,t}\boldsymbol{\varsigma}'_{ti}\hat{\mathbf{V}}^{-1}_t\hat{\sigma}^2_{u,t}\boldsymbol{\varsigma}_{ti}$,

$g_{2ti}\left(\hat{\sigma}^2_u\right) = \left[\boldsymbol{\mu}_{x,t-1,i} - \hat{\sigma}^2_{u,t}\mathbf{X}'_{t-1}\hat{\mathbf{V}}^{-1}_t\boldsymbol{\varsigma}_{ti}\right]'\left(\mathbf{X}'_{t-1}\hat{\mathbf{V}}^{-1}_t\mathbf{X}_{t-1}\right)^{-1}\left[\boldsymbol{\mu}_{x,t-1,i} - \hat{\sigma}^2_{u,t}\mathbf{X}'_{t-1}\hat{\mathbf{V}}^{-1}_t\boldsymbol{\varsigma}_{ti}\right]$ e

[85] Admitiu-se que o coeficiente de associação espacial é conhecido, sendo igual à média das sete estatísticas globais *I* de Moran (tabela 7.2.6), ϕ=0,293.

$g_{3ti}\left(\hat{\sigma}_u^2\right)=tr\left[\mathbf{L}_{ti}\left(\hat{\sigma}_{u,t}^2\right)\mathbf{V}_t\left(\hat{\sigma}_{u,t}^2\right)\mathbf{L}'_{ti}\left(\hat{\sigma}_{u,t}^2\right)\overline{\mathbf{V}}\left(\hat{\sigma}_{u,t}^2\right)\right]$. Para além disso $\mathbf{L}_{ti}\left(\hat{\sigma}_{u,t}^2\right)=\left(\boldsymbol{\varsigma}'_{ti}-\hat{\sigma}_{u,t}^2\boldsymbol{\varsigma}'_{ti}\hat{\mathbf{V}}_t^{-1}\mathbf{B}_t^{-1}\right)\hat{\mathbf{V}}_t^{-1}$ e $\overline{\mathbf{V}}\left(\tilde{\sigma}_{u,t}^2\right)=2k_1^2tr\left(\mathbf{C}_1\mathbf{V}_t\mathbf{C}_1\mathbf{V}_t\right)$ com $k_1=\left[m-r\left(\mathbf{H}^{(1)}\right)\right]^{-1}$, $\mathbf{V}_t=\mathbf{R}_t+\sigma_{u,t}^2\mathbf{B}_t^{-1}$ e $\mathbf{C}_1=\mathbf{C}^{(1)'}\left[\mathbf{I}_m-\mathbf{C}^{(1)}\mathbf{X}_{t-1}\left(\mathbf{X}'_{t-1}\mathbf{C}^{(1)'}\mathbf{C}^{(1)}\mathbf{X}_{t-1}\right)^{-}\mathbf{X}'_{t-1}\mathbf{C}^{(1)'}\right]\mathbf{C}^{(1)}$ com $\mathbf{C}^{(1)}=\left(\mathbf{I}_m-\phi\mathbf{W}\right)$.

C. Estimador temporal de Rao-Yu

Este estimador combinado, que utiliza informação seccional e cronológica, é assistido pelo modelo de Rao-Yu (4.3.4), com componentes de variância estimadas por uma extensão do método III de Henderson, (4.3.9) e (4.3.10), e com coeficiente de autocorrelação temporal conhecido[86], ρ=0,37. O estimador de Rao-Yu é dado por:

$$\hat{\mu}_{it}^{RY}=\boldsymbol{\mu}'_{x,i,t-1}\hat{\boldsymbol{\beta}}+\left[\hat{\sigma}_v^2\mathbf{1}_T+\hat{\sigma}^2\boldsymbol{\gamma}_t\right]'\hat{\mathbf{V}}_i^{-1}\left[\hat{\boldsymbol{\mu}}_i^{dir}-\mathbf{X}_i\hat{\boldsymbol{\beta}}\right], \tag{7.2.16}$$

onde $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\mu}}^{dir}$, $\hat{\boldsymbol{\mu}}^{dir}=col_{1\le i\le m}\left(\hat{\boldsymbol{\mu}}_i^{dir}\right)$, $\hat{\boldsymbol{\mu}}_i^{dir}=col_{1\le t\le T}\left(\hat{\mu}_{it}^{dir}\right)$, $\mathbf{X}=col_{1\le i\le m}\left(\mathbf{X}_i\right)$ e $\mathbf{X}_i=col_{1\le t\le T}\left(\boldsymbol{\mu}'_{x,i,t-1}\right)$ (i=1, …, 28; t=1, …, 7). Um estimador analítico do EQMP do estimador (7.2.16) é dado por:

$$eqmp\left(\hat{\mu}_{it}^{RY}\right)=g_{1it}\left(\hat{\boldsymbol{\psi}}\right)+g_{2it}\left(\hat{\boldsymbol{\psi}}\right)+2g_{3it}\left(\hat{\boldsymbol{\psi}}\right), \tag{7.2.17}$$

onde $g_{1it}\left(\hat{\boldsymbol{\psi}}\right)$ é dado por (4.3.12), $g_{2it}\left(\hat{\boldsymbol{\psi}}\right)$ é dado por (4.3.13) e $g_{3it}\left(\hat{\boldsymbol{\psi}}\right)$ é dado por (4.3.14), com as devidas adaptações para o problema em estudo.

D. Estimador espaciotemporal proposto sem restrições

Este estimador combinado, que tira partido da informação espacial/seccional e cronológica, é assistido pelo modelo espaciotemporal (5.2.5) proposto no quinto capítulo. As suas componentes de variância são estimadas por uma extensão do método

[86] Admitiu-se que o coeficiente de autocorrelação temporal é conhecido, sendo igual ao coeficiente de autocorrelação de primeira ordem, calculado com base nos resíduos do método dos mínimos quadrados ordinários do modelo de regressão linear, ρ=0,372.

III de Henderson, (5.5.4) e (5.5.7), e os coeficiente de associação espacial e de autocorrelação temporal são conhecidos, ϕ=0,29 e ρ=0,37. Este estimador é dado por[87]:

$$\hat{\mu}_{it}^{LP} = \boldsymbol{\mu}'_{x,i,t-1}\hat{\boldsymbol{\beta}} + \left(\hat{\sigma}_u^2 \boldsymbol{\varsigma}_i' \otimes \mathbf{1}_T' + \hat{\sigma}^2 \boldsymbol{\zeta}_{it}'\right)\hat{\mathbf{V}}^{-1}\left[\hat{\boldsymbol{\mu}}^{dir} - \mathbf{X}\hat{\boldsymbol{\beta}}\right], \tag{7.2.18}$$

onde $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\mu}}^{dir}$, $\hat{\boldsymbol{\mu}}^{dir} = col_{1\le i\le m}\left(\hat{\boldsymbol{\mu}}_i^{dir}\right)$, $\hat{\boldsymbol{\mu}}_i^{dir} = col_{1\le t\le T}\left(\hat{\mu}_{it}^{dir}\right)$, $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$ e $\mathbf{X}_i = col_{1\le t\le T}\left(\boldsymbol{\mu}'_{x,i,t-1}\right)$ (i=1, …, 28; t=1, …, 7). Um estimador analítico do EQMP deste EBLUP é dado por:

$$eqmp\left(\hat{\mu}_{it}^{LP}\right) = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) + 2g_{3it}(\hat{\boldsymbol{\psi}}), \tag{7.2.19}$$

onde $g_{1it}(\hat{\boldsymbol{\psi}})$ é dado por (5.6.2), $g_{2it}(\hat{\boldsymbol{\psi}})$ é dado por (5.6.3) e $g_{3it}(\hat{\boldsymbol{\psi}})$ é dado por (5.6.4), com as devidas adaptações para o problema em estudo.

E. Estimador espaciotemporal proposto com restrições

O estimador espaciotemporal com restrições, que tira partido da informação espacial/seccional e cronológica e que garante a consistência interna das estimativas, é assistido pelo modelo espaciotemporal (6.4.1) sujeito às restrições (6.4.2). Este estimador, que também utiliza estimativas das componentes de variância obtidas através do método dos momentos, (5.5.4) e (5.5.7), é dado por:

$$\begin{aligned}\hat{\mu}_{it}^{LPR} = {} & \boldsymbol{\mu}'_{x,i,t-1}\widehat{\boldsymbol{\beta}} + \left(\hat{\sigma}_u^2 \boldsymbol{\varsigma}'_i \otimes \mathbf{1}'_t + \hat{\sigma}^2 \boldsymbol{\zeta}'_{it}\right)\widehat{\mathbf{V}}^{-1}\left(\widehat{\boldsymbol{\mu}}^{dir} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ & + \left(\boldsymbol{\mu}'_{x,i,t-1}\mathbf{C}_{11}\boldsymbol{X}^{R\,\prime} + \boldsymbol{\mu}'_{x,i,t-1}\mathbf{C}'_{21}\mathbf{Q}'_v + \boldsymbol{m}'_{it}\mathbf{C}_{21}\mathbf{X}^{R\,\prime} + \right. \\ & \left. \boldsymbol{m}'_{it}\mathbf{C}_{22}\mathbf{Q}_v{}'\right)(\mathbf{Q}\mathbf{C}\mathbf{Q}')^{-1}\left(\boldsymbol{q} - \mathbf{Q}\tilde{\boldsymbol{\xi}}\right),\end{aligned} \tag{7.2.20}$$

onde $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\mu}}^{dir}$, $\hat{\boldsymbol{\mu}}^{dir} = col_{1\le i\le m}\left(\hat{\boldsymbol{\mu}}_i^{dir}\right)$, $\hat{\boldsymbol{\mu}}_i^{dir} = col_{1\le t\le T}\left(\hat{\mu}_{it}^{dir}\right)$, $\mathrm{X} = col_{1\le i\le m}(\mathbf{X}_i)$ e $\mathbf{X}_i = col_{1\le t\le T}\left(\boldsymbol{\mu}'_{x,i,t-1}\right)$. No que se refere a este estimador, decidiu garantir-se a consistência interna das estimativas ao nível de Portugal continental e ao nível de NUTSII, pelo que se tem $\boldsymbol{q} = col_{1\le t\le 7}\left(\hat{\mu}_t^{dir}\right)$ e $\boldsymbol{q} = col_{1\le e\le 5;1\le t\le 7}\left(\hat{\mu}_{et}^{dir}\right)$, respectivamente. A garantia da consistência interna a estes dois níveis origina dois

[87] À semelhança do que foi efectuado para os outros estimadores, decidiu utilizar-se o índice superior "LP" para denominar o estimador espaciotemporal, por ser devido a Luís Pereira.

estimadores, representados, respectivamente, por $\hat{\mu}_{it}^{LPR1}$ e $\hat{\mu}_{it}^{LPR2}$. A estimação do EQMP destes EBLUP tem que ser efectuado por um dos métodos de reamostragem apresentados nas secções 6.3.6. e 6.3.7.

É de salientar que todos os modelos de estimação em domínios envolvem estimativas das componentes de variância produzidas através do método dos momentos, o qual não exige a verificação da hipótese na normalidade dos erros do modelo. Por último, é ainda de notar que todos os modelos de estimação em domínios que assistem os estimadores *model-based* apresentados nesta subsecção envolvem variâncias amostrais, $\sigma_{\varepsilon,it}^2$, que se admitem conhecidas. Contudo, elas são desconhecidas nas aplicações práticas, tendo sido substituídas neste estudo por estimativas calculadas através de um método delta. Desta forma, as estimativas das variâncias amostrais calculadas através de (7.2.5) são consideradas como uma *proxy* de $\sigma_{\varepsilon,it}^2$.

7.2.5 Medidas de qualidade dos estimadores

Na avaliação das propriedades dos estimadores, foi adoptada a abordagem habitualmente seguida nos estudos por simulação deste tipo, referidos na secção 7.2.3. Assim, a qualidade dos estimadores dos parâmetros de interesse foi avaliada através de um conjunto de medidas de enviesamento, de precisão e de eficiência *design-based*, bem como através de taxas de cobertura de IC. Na exposição seguinte, onde são apresentadas as referidas medidas utilizadas, L representa o número de réplicas; μ_{it} representa o verdadeiro valor do parâmetro de interesse (média da variável de interesse) no i-ésimo domínio no período t; $\hat{\mu}_{l,it}$ representa uma estimativa do parâmetro de interesse no i-ésimo domínio no período t, obtida na l-ésima réplica e $eqmp\left(\hat{\mu}_{l,it}\right)$ representa uma sua estimativa do EQMP (l=1, …, L). Note-se que μ_{it} permanece inalterado nas L réplicas sob a abordagem *design-based*, porque a pseudo-população finita considerada é fixa para todas as L réplicas.

As medidas de qualidade *design-based*, utilizadas para avaliar a qualidade dos estimadores do parâmetro de interesse foram calculadas ao nível dos domínios individuais e ao nível de grupos de domínios.

7.2.5.1 Domínios individuais

As medidas *design-based* utilizadas para avaliar a qualidade dos estimadores do parâmetro de interesse, representados genericamente por $\hat{\mu}_{it}^{f}$, $f \in \{dir, sinQ, sinR, com, FH, NS, RY, LP, LPR1, LPR2\}$, são baseadas no enviesamento, $B(\hat{\mu}_{it}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\mu}_{l,it} - \mu_{it})$, no EQM, $EQM(\hat{\mu}_{it}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\mu}_{l,it} - \mu_{it})^2$, na variância, $V(\hat{\mu}_{it}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\mu}_{l,it} - \bar{\hat{\mu}}_{it})^2$ onde $\bar{\hat{\mu}}_{it} = \frac{1}{L}\sum_{l=1}^{L}\hat{\mu}_{l,it}$, e no erro absoluto, $EA(\hat{\mu}_{it}) = \frac{1}{L}\sum_{l=1}^{L}|\hat{\mu}_{l,it} - \mu_{it}|$. Ao nível dos domínios individuais, ou seja, ao nível de cada domínio *i* no período *t* (*i*=1, …, 28; *t*=1, …, 7), são utilizadas as seguintes medidas:

(*i*) Enviesamento relativo absoluto (BRA): $BRA(\hat{\mu}_{it}) = \left|\frac{B(\hat{\mu}_{it})}{\mu_{it}}\right|$;

(*ii*) Rácio de enviesamento absoluto (RBA): $RBA(\hat{\mu}_{it}) = \left|\frac{B(\hat{\mu}_{it})}{\sqrt{V(\hat{\mu}_{it})}}\right|$;

(*iii*) Erro relativo absoluto (ERA): $ERA(\hat{\mu}_{it}) = \frac{EA(\hat{\mu}_{it})}{\mu_{it}}$;

(*iv*) Erro padrão relativo (EPR): $EPR(\hat{\mu}_{it}) = \frac{\sqrt{EQM(\hat{\mu}_{it})}}{\mu_{it}}$;

(*v*) Coeficiente de variação (CV): $CV(\hat{\mu}_{it}) = \frac{\sqrt{V(\hat{\mu}_{it})}}{\mu_{it}}$;

(*vi*) Eficiência relativa (EFR): $EFR(\hat{\mu}_{it}) = \left[\frac{EQM(\hat{\mu}_{it}^{dir})}{EQM(\hat{\mu}_{it}^{f})}\right]^{\frac{1}{2}}$;

(*vii*) Taxa de cobertura do IC *design-based* (TCDB): $TCDB(\hat{\mu}_{it}) = 100\frac{R(\hat{\mu}_{it})}{L}$, onde $R(\hat{\mu}_{it})$ corresponde ao número de simulações para as quais o IC $\hat{\mu}_{l,it} \pm z_{\alpha/2}\sqrt{V(\hat{\mu}_{it})}$ contém o verdadeiro valor do parâmetro, μ_{it};

(*viii*) Taxa de cobertura do IC *model-based* (TCMB): $TCMB(\hat{\mu}_{it}) = 100\dfrac{R(\hat{\mu}_{it})}{L}$, onde $R(\hat{\mu}_{it})$ corresponde ao número de simulações para as quais o IC $\hat{\mu}_{l,it} \pm z_{\alpha/2}\sqrt{eqmp(\hat{\mu}_{l,it})}$ contém o verdadeiro valor do parâmetro, μ_{it}.

Note-se que os IC *design-based* são construídos a partir das variâncias *design-based*, as quais são obtidas através da simulação de Monte Carlo, mas que geralmente são desconhecidas nas aplicações práticas. Por sua vez, os IC *model-based* são construídos a partir de estimativas do EQMP dos estimadores EBLUP, calculadas através de uma metodologia delta ou por reamostragem[88]. Contudo, ambas as taxas de cobertura calculadas no âmbito deste estudo empírico são obtidas sob amostragem repetida. Considerando-se que a abordagem proposta para estimação do parâmetro de interesse é do tipo *model-assisted*[89], então parece ser extremamente importante a comparação de IC alternativos (*design-based* e *model-based*) sob amostragem repetida, como instrumento de avaliação da qualidade das medidas de precisão *model-based* num contexto de uma população real e de um método de amostragem realista.

É também importante salientar que, quando a estimação do parâmetro de interesse é efectuada com base em estimadores do tipo *model-assisted*[90], a avaliação da precisão das estimativas pode ser efectuada (e é normalmente efectuada) com base em estimativas do EQMP *model-based* desses estimadores, *i.e.*, com base em estimativas do EQMP baseadas nas propriedades do modelo. Como neste trabalho foram propostos estimadores EBLUP para os quais foi possível deduzir, segundo diferentes metodologias, aproximações para o estimador do EQMP *model-based* desses

[88] No âmbito deste estudo empírico foram utilizados estimadores do EQMP dos EBLUP baseados na aproximação analítica, por serem aqueles que apresentam globalmente melhores propriedades, apesar de não serem uniformemente os melhores estimadores, tal como será apresentado no subcapítulo 7.3. Os estimadores do EQMP dos EBLUP baseados em métodos de reamostragem constituem uma excelente alternativa aos estimadores delta, sobretudo em modelos que envolvem a estimação de um elevado número de parâmetros ou que envolvem restrições, nos quais se torna impossível manipular algumas expressões algébricas.

[89] No sentido em que não se supõe que a população alvo seja exactamente gerada através de um modelo de superpopulação, mas apenas que possa ser aproximadamente descrita por tal modelo.

[90] Apesar dos estimadores propostos no quinto capítulo não serem considerados estimadores *model-based* puros, a qualidade da estimação baseada nesses estimadores pode ser avaliada à luz desses critérios.

estimadores propostos, então só poderá ser efectuada uma boa avaliação da precisão das estimativas se for utilizado o estimador do EQMP que apresentar melhores propriedades. Neste sentido, foi efectuado, no âmbito deste trabalho, um estudo empírico por simulação *model-based*, apresentado no subcapítulo 7.3, tendo como objectivo a avaliação do desempenho de estimadores alternativos do EQMP dos estimadores combinados do parâmetro de interesse. Pode, portanto, considerar-se que esse estudo empírico responde a um sub-problema do problema principal deste trabalho, que é um problema de sondagens, e que consiste em estimar o preço médio de transacção da habitação com a melhor precisão possível.

7.2.5.2 Grupos de domínios

Tendo em conta o grande número de domínios de interesse (28 NUTSIII) e o número de períodos considerados neste estudo (7 trimestres), e com o objectivo de facilitar a análise relativa ao comportamento médio dos diversos estimadores em análise, decidiu formar-se grupos de domínios. Os 28 domínios foram divididos em seis grupos mutuamente exclusivos em função da sua dimensão média (unidades secundárias – transacções). Cada grupo de domínios, denotado por G_b, b=1, ..., 6, contém m_b domínios. Na tabela 7.2.10 são apresentados os grupos de domínios, no que se refere à sua dimensão média amostral e ao número de domínios incluídos em cada grupo. A constituição de cada grupo de domínios encontra-se no apêndice 7.

Tabela 7.2.10: Dimensão média amostral e número de domínios por grupos de domínios de interesse

Grupo (G_b)	Dimensão média amostral	Número de domínios (m_b)
1	1 transacção	2
2	2 a 5 transacções	4
3	6 a 9 transacções	3
4	10 a 19 transacções	8
5	20 a 29 transacções	5
6	Pelo menos 30 transacções	6

A divisão dos 28 domínios em seis grupos foi efectuada em função da dimensão média da amostra em cada domínio, pois a avaliação dos méritos relativos dos estimadores propostos não deve ser feita sem se ter em conta as dimensões amostrais dos domínios. Por exemplo, as conclusões retiradas para o grupo 6, que inclui as NUTSIII onde, em média, foram observadas pelo menos 30 transacções ao longo dos sete períodos, não

serão iguais às conclusões retiradas para o grupo 2, considerado um grupo de "pequenos domínios". A formação do grupo 1, com uma especificidade e natureza diferente dos restantes, teve como objectivo evitar que a inclusão no grupo 2 dos domínios que o integram viesse alterar as conclusões sobre os méritos relativos dos diversos estimadores. Com efeito, não foi descurado o facto dos estimadores poderem apresentar variância *design-based* nula nos domínios do grupo 1, pelo facto de só apresentarem uma transacção em todas as simulações. Devido a esta situação, decidiu substituir-se uma variância nula, num particular domínio *i* no período *t*, por um pequeno valor, neste caso igual a 10^{-4}, para ser possível apresentar os resultados do rácio de enviesamento.

Para a análise dos resultados decidiu, então, calcular-se a média de todas as medidas de precisão e de enviesamento expostas acima para cada grupo de domínios, para todos os períodos do tempo. As medidas resultantes apresentam-se em seguida:

(*i*) Enviesamento relativo absoluto médio (BRAM):

$$BRAM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} BRA(\hat{\mu}_{it});$$

(*ii*) Rácio de enviesamento absoluto médio (RBAM):

$$RBAM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} RBA(\hat{\mu}_{it});$$

(*iii*) Erro relativo absoluto médio (ERAM): $ERAM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} ERA(\hat{\mu}_{it});$

(*iv*) Erro padrão relativo médio (EPRM): $EPRM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} EPR(\hat{\mu}_{it});$

(*v*) Coeficiente de variação médio (CVM): $CVM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} CV(\hat{\mu}_{it});$

(*vi*) Eficiência relativa média (EFRM): $EFRM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} EFR(\hat{\mu}_{it});$

(*vii*) Taxa de cobertura do IC *design-based* média (TCDBM):

$$TCDBM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} TCDB(\hat{\mu}_{it});$$

(*viii*) Taxa de cobertura do IC *model-based* média (TCMBM):

$$TCMBM(\hat{\mu}_{it}) = \frac{1}{m_b T}\sum_{i=1}^{m_b}\sum_{t=1}^{T} TCMB(\hat{\mu}_{it}).$$

Por último, é ainda de salientar que, por conveniência de análise de resultados, em alguns casos considerou-se a média das medidas de qualidade dos estimadores ao nível de cada domínio, tendo-se calculado a média sobre os sete períodos do tempo. Em todo o caso, continuam a denominar-se por medidas "médias".

7.2.6 Diagnóstico aos modelos de estimação em domínios

No âmbito deste estudo empírico são utilizados os seguintes modelos de estimação em domínios: modelo de Fay-Herriot, modelo de Salvati, modelo de Rao-Yu e modelo espaciotemporal proposto. Todos estes modelos são casos particulares do modelo linear misto geral (3.2.1), diferindo sobretudo na estrutura de covariâncias dos efeitos aleatórios, uma vez que todos os modelos têm termo independente e a mesma variável explicativa. Os modelos de Fay-Herriot e de Salvati são modelos contemporâneos, sendo que o primeiro tem uma estrutura de covariâncias homogénea e diagonal, enquanto no segundo considera-se a existência de associação espacial entre os efeitos aleatórios de domínio. Por sua vez, os modelos de Rao-Yu e espaciotemporal são modelos longitudinais, sendo que o primeiro tem uma estrutura de covariâncias homogénea diagonal por blocos, com autocorrelação temporal em cada bloco, enquanto no segundo considera-se a existência de associação espacial entre os efeitos aleatórios de domínio e de autocorrelação temporal entre os efeitos aleatórios de domínio-tempo.

Apesar do objectivo principal deste trabalho estar circunscrito ao contexto das sondagens, ou seja, consistir na produção de estimativas dos parâmetros de interesse em pequenos domínios com a melhor precisão possível, não se pode menosprezar o trabalho de diagnóstico dos modelos que assistem a estimação em domínios, o qual constitui uma tarefa nobre da econometria. A este respeito, Ghosh e Rao (1994) sublinham que a estimação *model-based* ou *model-assisted* deve ser acompanhada de

um diagnóstico cuidadoso dos pressupostos dos modelos, de forma a averiguar se os modelos se ajustam bem aos dados. Em particular, é necessário testar a significância estatística dos efeitos fixos, das componentes de variância e a normalidade dos erros do modelo[91], tal como é sugerido por Diggle (1988) e estendido por Wolfinger (1993), e efectuado, por exemplo, em Battese *et al*. (1988), Coelho (2000), Militino *et al*. (2007*a*, 2007*b*) e Ugarte *et al*. (2009). Antes de se apresentarem os resultados, importa especificar previamente os ensaios de hipótese e indicar as estatísticas-teste utilizadas, o que é feito em cada caso.

7.2.6.1 Teste à significância estatística dos efeitos fixos

O ensaio de hipóteses à significância dos efeitos fixos tem as seguintes hipóteses subjacentes: $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$, k=0, 1. A estatística-teste de Wald é utilizada para realizar este ensaio, sendo dada por (Littell *et al.*, 2006):

$$t = \frac{\mathbf{l}_k \hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{l}_k \hat{\mathbf{C}} \mathbf{l}_k'}}, \tag{7.2.21}$$

onde $\mathbf{l}_k$ é um vector 1×2 com o valor um na k-ésima posição e zero na outra, $\hat{\boldsymbol{\beta}}$ é um vector 2×1 com as estimativas dos efeitos fixos e $\hat{\mathbf{C}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}$ é a matriz de covariâncias de $\left[\left(\hat{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\right) \ \left(\hat{\boldsymbol{\upsilon}}' - \boldsymbol{\upsilon}'\right)\right]'$. Sob a hipótese nula, e assumindo a normalidade dos erros do modelo, a estatística-teste (7.2.21) tem distribuição aproximadamente t de Student (McLean e Sanders, 1988), sendo que os graus de liberdade podem ser estimados, por exemplo, pelo método de Satterwaitte (Giesbrecht e Burns, 1985; McLean e Sanders, 1988; Fai e Cornelius, 1996) ou pelo método de Kenward e Roger (1997). Neste caso optou-se por utilizar o método de Kenward-Roger, pelo facto de envolver resultados normalmente utilizados no contexto da estimação em pequenos domínios, designadamente a estimação da matriz de covariâncias dos efeitos fixos e dos

[91] Note-se que apesar dos estimadores das componentes de variância não exigirem a normalidade dos efeitos aleatórios específicos de domínio e dos erros da sondagem, e dos estimadores EBLUP serem robustos à não normalidade desses erros do modelo linear misto, a estimação do EQMP dos EBLUP exige que se verifique essa normalidade.

efeitos aleatórios, $\hat{\mathbf{C}}$, através do método proposto por Prasad e Rao (1990) e por Harville e Jeske (1992).

Passa-se agora ao diagnóstico da significância estatística dos efeitos fixos dos modelos que assistem a estimação em domínios, os quais foram estimados pelo método da MVR. Neste diagnóstico, as componentes de variância foram estimadas externamente aos modelos através do método dos momentos. Os resultados das estimativas dos parâmetros dos modelos, bem como algumas medidas de diagnóstico são apresentados nas tabelas 7.2.11 a 7.2.14.

Tabela 7.2.11: Estimativas dos hiper-parâmetros e medidas de diagnóstico ao modelo seccional de Fay-Herriot, para cada trimestre

t	Compo-nentes de variância	Termo independente				Variável auxiliar				Critérios de informação[92]	
	$\widehat{\sigma}^2$	$\widehat{\beta}_0$	t_{obs}	*gl*	*p*	$\widehat{\beta}_1$	t_{obs}	*gl*	*p*	*AIC*	*AICC*
1	13.566	-115,02	-1,87	26	0,0728	0,9234	6,71	26	0,0000	340,1	340,1
2	11.295	-178,43	-2,21	26	0,0361	0.9822	7,58	26	0,0000	335,3	335,3
3	24.060	-95,39	-1,39	26	0,1775	0,9131	5,06	26	0,0000	343,9	343,9
4	26.832	-238,81	-2,09	26	0,0465	1,0292	5,46	26	0,0000	347,5	347,5
5	16.460	-312,08	-1,89	26	0,0697	1,0788	6,70	26	0,0000	338,5	338,5
6	14.944	-195,92	-2,06	26	0,0495	0,9577	7,20	26	0,0000	342,2	342,2
7	16.582	-176,81	-2,10	26	0,0456	0,9230	7,38	26	0,0000	342,0	342,0

Os resultados da tabela 7.2.11, referentes ao modelo de Fay-Herrriot, evidenciam que a variável auxiliar tem poder explicativo significativo sobre a variável dependente do modelo, em todos os trimestres. Por sua vez, o termo independente só não é significativo nos trimestres 1, 3 e 5.

[92] Existem diversos critérios de informação, sendo que todos eles apresentam formas diversas que podem introduzir alguma imprecisão na linguagem e confusão ao decisor. A forma aqui utilizada está apresentada na forma "quanto menor melhor". O critério de informação de Akaike (AIC) é calculado da seguinte forma: $AIC = -2\log[l(\Theta)] + 2q$, onde $l(\Theta)$ representa a verosimilhança máxima no espaço q-dimensional de parâmetros, Θ. O critério *AICC* é uma versão corrigida pela dimensão amostral, *n*, do *AIC*, proposta por Hurvich e Tsai (1989), dada por: $AICC = -2\log[l(\Theta)] + 2qn/(n-q-1)$, porque o *AIC* pode ser subenviesado em amostras de pequenas dimensões. Burnham e Anderson (2002) sugerem que a utilização do *AICC* em vez do *AIC* quando $n/q < 40$ para o caso do modelo com maior q, o que ocorre nos casos dos modelos de Fay-Herriot e de Salvati. Note-se que no caso particular dos modelos de estimação em domínios, a dimensão amostral, *n*, corresponde ao número de pequenos domínios, *m*.

Tabela 7.2.12: Estimativas dos hiper-parâmetros e medidas de diagnóstico ao modelo espacial de Salvati (ϕ=0,29), para cada trimestre

t	Compo-nentes de variância	Termo independente				Variável auxiliar				Critérios de informação	
	$\widehat{\sigma}_u^2$	$\widehat{\beta}_0$	t_{obs}	gl	p	$\widehat{\beta}_1$	t_{obs}	gl	p	*AIC*	*AICC*
1	10.080	-14,50	-1,40	26	0,1733	0,8087	4,85	26	0,0001	352,3	352,3
2	1.552	-60,94	-1,84	26	0,0772	0,8549	9,64	26	0,0001	334,5	334,5
3	17.159	68,84	1,80	26	0,0835	0,7882	3,90	26	0,0006	349,6	349,6
4	22.581	-185,10	-2,06	26	0,0495	0,9422	4,09	26	0,0004	356,9	356,9
5	12.760	-218,94	-2,17	26	0,0393	0,9763	5,29	26	0,0000	346,3	346,3
6	7.927	-49,52	-1,69	26	0,1030	0,8376	6,12	26	0,0000	345,9	345,9
7	11.800	-216,24	-2,11	26	0,0446	0,9232	6,16	26	0,0000	349,6	349,6

Os resultados relativos ao modelo espacial de Salvati, expostos na tabela 7.2.12, demonstram igualmente que a variável auxiliar tem poder explicativo significativo sobre a variável dependente do modelo, em todos os trimestres. Neste caso, o termo independente só é significativo nos trimestres 4, 5 e 7.

Tabela 7.2.13: Estimativas dos hiper-parâmetros e medidas de diagnóstico ao modelo seccional e cronológico de Rao-Yu (ρ=0,37)

Componentes de variância		Termo independente				Variável auxiliar				Critérios de informação	
$\widehat{\sigma}_v^2$	$\widehat{\sigma}^2$	$\widehat{\beta}_0$	t_{obs}	gl	p	$\widehat{\beta}_1$	t_{obs}	gl	p	*AIC*	*AICC*
10.976	766	267,76	3,73	194	0,0002	0,5243	7,73	194	0,0000	2.271,8	2.271,8

Tabela 7.2.14: Estimativas dos hiper-parâmetros e medidas de diagnóstico ao modelo espaciotemporal (ϕ=0,29 ; ρ=0,37)

Componentes de variância		Termo independente				Variável auxiliar				Critérios de informação	
$\widehat{\sigma}_u^2$	$\widehat{\sigma}^2$	$\widehat{\beta}_0$	t_{obs}	gl	p	$\widehat{\beta}_1$	t_{obs}	gl	p	*AIC*	*AICC*
100	766	-42,18	-2,03	194	0,0437	0,8293	25,42	194	0,0000	2.484,2	2.484,2

Por último, os resultados apresentados nas tabelas 7.2.13 e 7.2.14, referentes aos modelos longitudinais de Rao-Yu e espaciotemporal, permitem concluir que os efeitos fixos são significativos em ambos os modelos.

7.2.6.2 Teste à significância estatística das componentes de variância

No âmbito desta tese, propôs-se que as componentes de variância fossem estimadas através do método dos momentos de forma a não ser necessário a verificação da normalidade dos erros do modelo. Para além disso, no processo de estimação dos parâmetros de interesse em domínios, admitiu-se que os coeficientes de autocorrelação temporal e de associação espacial são conhecidos. Em todo o caso, continua a ser fundamental que todas essas componentes de variância sejam estatisticamente significativas, tendo em vista a obtenção de estimativas *model-based* válidas dos parâmetros de interesse. Desta forma, para ser possível testar a significância estatística das referidas componentes de variância, tem que se efectuar a sua estimação através de um dos métodos de verosimilhança. Admita-se, portanto, que neste estudo de diagnóstico aos modelos que assistem a estimação, as componentes de variância são estimadas pelo método da MVR.

O ensaio de hipóteses à significância estatística de um parâmetro de variância, genericamente representado por σ^2, tem as seguintes hipóteses subjacentes: $H_0 : \sigma^2 = 0$ contra $H_1 : \sigma^2 > 0$ [93]. Para a realização deste tipo de ensaio de hipóteses recorre-se à estatística-teste de Wald (Verbeke e Molenberghs, 2000):

$$t = \frac{\hat{\sigma}^2}{\sqrt{\hat{V}(\hat{\sigma}^2)}}, \tag{7.2.22}$$

a qual segue aproximadamente uma distribuição normal padrão. Esta estatítica-teste é válida para amostras grandes, mas pode não ser fiável para amostras pequenas ou para parâmetros que tenham uma distribuição amostral assimétrica (Littell *et al.*, 2006). Segundo estes autores, uma melhor alternativa consiste em utilizar o teste de razão de verosimilhanças (*RV*). Este teste compara dois modelos, sendo que um deles (modelo reduzido com parâmetros pertencentes ao subespaço q_R-dimensional, Θ_R, do espaço de parâmetros q-dimensional, Θ) está encaixado no outro (modelo completo com

[93] Se a componente de variância for um coeficiente de autocorrelação temporal ou de associação espacial, representado genericamente por φ, então as hipóteses subjacentes são: $H_0: \varphi = 0$ contra $H_1: \varphi \neq 0$.

parâmetros pertencentes ao espaço Θ), sendo a sua estatística-teste dada por (Verbeke e Molenberghs, 2000):

$$RV = -2\log\left[\frac{l_{MVR}\left(\hat{\Theta}_R\right)}{l_{MVR}\left(\hat{\Theta}\right)}\right], \tag{7.2.23}$$

onde $l_{MVR}(\Theta_R)$ e $l_{MVR}(\Theta)$ representam as verosimilhanças dos modelos reduzido e completo, respectivamente. Sob a hipótese a testar de nulidade das componentes de variância adicionais do modelo completo, esta estatística-teste segue aproximadamente uma distribuição do qui-quadrado (χ^2), com um número de graus de liberdade igual à diferença entre o número de parâmetros do espaço Θ e o número de parâmetros do subespaço Θ_R. Segundo Verbeke e Molenberghs (2000), o teste de razão de verosimilhanças também pode ser utilizado como teste marginal à necessidade de incorporar mais efeitos aleatórios num modelo. Neste caso, quando se compara um modelo com q_1 efeitos aleatórios com outro modelo com q_2 efeitos aleatórios, a distribuição assimptótica da estatística-teste é uma mistura de uma $\chi^2_{(q_1)}$ com uma $\chi^2_{(q_2)}$, com pesos iguais, normalmente representada por $\chi^2_{(q_1:q_2)}$.

Antes de se avançar para a apresentação dos resultados deste diagnóstico, é importante referir que o procedimento MIXED do programa SAS não permite a estimação pelos métodos de verosimilhança de estruturas de covariância espaciais do tipo SAR. Desta forma, decidiu testar-se a existência de associação espacial na estrutura de erro dos modelos utilizando uma estrutura espacial linear[94]. Esta decisão foi suportada por duas razões: (*i*) o objectivo desta análise consiste em testar se existe evidência de covariância espacial, independentemente do seu tipo; e (*ii*) a estrutura espacial linear é a estrutura disponível no SAS que mais se aproxima da estrutura espacial do tipo SAR.

Passa-se agora ao diagnóstico da significância estatística das componentes de variância dos modelos que assistem a estimação em domínios, os quais foram integralmente estimados pelo método da MVR. Os programas em SAS que suportam esta estimação encontram-se no apêndice 8. Os resultados das estimativas das componentes de

[94] A estrutura espacial linear define que a covariância entre dois efeitos aleatórios específicos de domínio é dada por: $cov(u_i; u_{i'}) = \sigma_u^2(1 - \phi w_{ii'}) \times I(\phi w_{ii'} < 1)$, onde $I(\phi w_{ii'} < 1)$ é uma variável indicatriz que é igual a um quando $w_{ii'} < \phi$ e é igual a zero em caso contrário (Littell *et al.*, 2006).

variância dos modelos, bem como algumas medidas de diagnóstico, são apresentados nas tabelas 7.2.15 a 7.2.18.

Tabela 7.2.15: Estimativa da componente de variância do modelo seccional de Fay-Herriot e respectivas medidas de diagnóstico, para cada trimestre

t	$\widehat{\sigma}^2$	t_{obs}	p	$-2\log[l_{MVR}(\Theta)]$	*AIC*	*AICC*
1	12.057	3,09	0,0010	339,9	341,9	342,1
2	9.420	2,96	0,0015	335,1	337,1	337,2
3	24.060	*	*	343,9	345,9	346,1
4	26.831	*	*	347,5	349,5	349,7
5	8.373	2,86	0,0021	335,1	337,1	337,3
6	13.249	3,13	0,0009	342,1	344,1	344,3
7	11.259	3,01	0,0013	340,8	342,8	343,0

Nota: * O processo de estimação convergiu, mas a matriz hessiana não é definida positiva.

Pelos resultados disponíveis na tabela 7.2.15, pode verificar-se que a componente de variância do modelo de Fay-Herriot é estatisticamente significativa. Desta forma, confirma-se a existência de evidência empírica no sentido de introdução no modelo de efeitos aleatórios específicos de domínio.

Tabela 7.2.16: Estimativas das componentes de variância do modelo espacial de Salvati e respectivas medidas de diagnóstico, para cada trimestre

t	$\widehat{\sigma}_u^2$	t_{obs}	p	$\widehat{\phi}$	t_{obs}	p	$-2\log[l_{MVR}(\Theta)$	*AIC*	*AICC*
1	16.305	3,00	0,0013	0,269	2,55	0,0108	338,4	342,4	343,0
2	6.886	*	*	0,771	*	*	336,0	340,0	340,5
3	16.553	3,18	0,0007	0,153	1,16	0,2460	334,9	338,9	339,4
4	21.019	4,13	0,0000	0,146	0,89	0,3735	339,2	343,2	343,7
5	14.699	2,19	0,0143	0,189	1,72	0,0854	335,9	339,9	340,4
6	8.314	*	*	0,686	*	*	344,4	348,4	348,9
7	17.349	2,71	0,0034	0,229	2,51	0,0121	341,8	345,8	346,3

Nota: * O processo de estimação convergiu, mas a matriz hessiana não é definida positiva.

Os resultados disponíveis relativos ao diagnóstico da significância estatística das componentes de variância do modelo espacial, apresentados na tabela 7.2.16, permitem verificar que o parâmetro de variância é estatisticamente significativo. Relativamente ao coeficiente de associação espacial, verifica-se que ele só é estatisticamente significativo nos trimestres 1 e 7. Contudo, é de notar que no trimestre 5 esse coeficiente ainda é estatisticamente significativo ao nível de significância de 10%. Pode, portanto, concluir-se que existe alguma evidência empírica para a introdução de associação espacial entre os efeitos aleatórios específicos de domínio.

Tabela 7.2.17: Estimativas das componentes de variância do modelo seccional e cronológico de Rao-Yu e respectivas medidas de diagnóstico

Componentes de variância	t_{obs}	p	Verosimilhança e critérios de informação
$\hat{\sigma}_v^2 = 16.237$	2,83	0,0023	$-2\log[l_{MVR}(\Theta)]$= 2.264,0
$\hat{\sigma}^2 = 1608$	3,41	0,0000	*AIC*= 2.270,0
$\hat{\rho} = 0{,}359$	1,68	0,0930	*AICC*= 2.270,1

Os resultados apresentados na tabela 7.2.17 indicam que os dois parâmetros de variância do modelo de Rao-Yu são estatisticamente significativos. Contudo, esses resultados também evidenciam que o coeficiente de autocorrelação temporal só é estatisticamente significativo ao nível de significância de 10%.

Vai agora realizar-se um teste marginal (de razão de verosimilhanças) à necessidade de incorporar cada um dos dois tipos de efeitos aleatórios do modelo de Rao-Yu. Quando se estima um caso particular do modelo de Rao-Yu sem efeitos aleatórios de domínio, obtém-se $-2\log[l_{MVR}(\Theta)]$ =2.271,5. Tem-se, então, que *RV*=2.271,5–2.264,0=7,5. Tendo em conta o valor crítico obtido pela mistura, $\chi^2_{(1:2)}$=0,5×3,84+0,5×5,99=4,92, verifica-se que existe evidência estatística para a inclusão de efeitos aleatórios específicos de domínio. A estimação de um caso particular do modelo de Rao-Yu sem efeitos aleatórios de domínio-tempo produz $-2\log[l_{MVR}(\Theta)]$ =2.345,9. Neste caso também se verifica que existe evidência estatística para a inclusão de efeitos aleatórios específicos de domínio-tempo, uma vez que *RV*=2.345,9–2.264,0=81,9 > 4,92[95].

Desta forma, parece existir evidência empírica suficiente para se concluir da existência de autocorrelação temporal nos efeitos aleatórios específicos de domínio-tempo, bem como de efeitos aleatórios específicos de domínio.

[95] O valor crítico mantém-se, apesar dos efeitos aleatórios específicos de domínio-tempo envolverem a estimação de duas componentes de variância, porque o número de graus de liberdade depende apenas do número de efeitos aleatórios em cada modelo.

Tabela 7.2.18: Estimativas das componentes de variância do modelo espaciotemporal e respectivas medidas de diagnóstico

Componentes de variância	t_{obs}	p	**Verosimilhança e critérios de informação**
$\hat{\sigma}_u^2 = 9.203$	2,16	0,0154	$-2\log[l_{MVR}(\Theta)]$= 2.270,2
$\hat{\phi} = 0{,}146$	1,67	0,0949	*AIC*= 2.278,2
$\hat{\sigma}^2 = 16.280$	2,93	0,0017	*AICC*= 2.278,5
$\hat{\rho} = 0{,}962$	52,04	0,0000	-------

Pela análise dos resultados apresentados na tabela 7.2.18, pode verificar-se que todas as componentes de variância são estatisticamente significativas. A única dúvida recai sobre a significância estatística do coeficiente de associação espacial (valor-*p*=0,0949). No entanto, ele é significativo para um nível de significância de 10%.

À semelhança do que foi realizado para o modelo de Rao-Yu, vai agora realizar-se um teste marginal à necessidade de incorporar cada um dos dois tipos de efeitos aleatórios do modelo espaciotemporal. Quando se estima um caso particular do modelo espaciotemporal sem efeitos aleatórios de domínio, obtém-se $-2\log[l_{MVR}(\Theta)]$ =2.273,3. Tem-se, então, que *RV*=2.273,3–2.270,2=3,1 < 4,92. É, no entanto, de salientar que se se considerar um nível de significância de 10%, existe evidência estatística para a inclusão de efeitos aleatórios específicos de domínio com associação espacial entre eles. A estimação de um caso particular do modelo espaciotemporal sem efeitos aleatórios de domínio-tempo produz $-2\log[l_{MVR}(\Theta)]$ =3.768,8. Neste caso, verifica-se que existe evidência estatística clara para a inclusão de efeitos aleatórios específicos de domínio-tempo com estrutura AR(1), uma vez que *RV*=3.768,8–2.270,2=1.498,6 ≫ 4,92.

Perante estes resultados, parece existir evidência empírica suficiente para se concluir da existência de autocorrelação temporal nos efeitos aleatórios específicos de domínio-tempo, bem como da existência de associação espacial entre os efeitos aleatórios específicos de domínio[96].

Por último, note-se que, do ponto de vista puramente econométrico, os critérios de informação utilizados favorecem claramente os modelos com estruturas de covariância

[96] O que interessa aqui reter é que existe evidência estatística da existência de associação espacial entre os efeitos aleatórios específicos de domínio, independentemente do seu tipo. Daqui em diante, será utilizada a estrutura de covariância espacial do tipo SAR, subjacente ao modelo de estimação em domínios proposto no capítulo quinto.

dos efeitos aleatórios mais simples[97], ou seja, favorecem o modelo de Fay-Herriot relativamente ao modelo de Salvati e favorecem o modelo de Rao-Yu relativamente ao modelo espaciotemporal. Esta observação é válida independentemente das componentes de variância dos modelos serem estimadas pelo método dos momentos ou pelo método da MVR. Importa, contudo, referir mais uma vez que o objectivo deste trabalho consiste em determinar os estimadores dos parâmetros de interesse em pequenos domínios com as melhores propriedades, apesar do melhor estimador poder não ser assistido pelo modelo de estimação em domínios que apresenta o melhor "quadro econométrico".

7.2.6.3 Teste à normalidade dos erros

O diagnóstico à normalidade dos efeitos aleatórios e dos erros da sondagem dos modelos é baseado nesses erros transformados pelas matrizes de transformação $\hat{\mathbf{G}}^{-1/2}$ e $\hat{\mathbf{R}}^{-1/2}$, da seguinte forma:

$$\hat{\boldsymbol{\upsilon}}^{*} = \hat{\mathbf{G}}^{-1/2}\hat{\boldsymbol{\upsilon}} \overset{\circ}{\sim} N(\mathbf{0};\mathbf{I}), \tag{7.2.24}$$

$$\hat{\boldsymbol{\varepsilon}}^{*} = \hat{\mathbf{R}}^{-1/2}\hat{\boldsymbol{\varepsilon}} \overset{\circ}{\sim} N(\mathbf{0};\mathbf{I}), \tag{7.2.25}$$

uma vez que Fuller e Battese (1973) mostraram que os erros transformados são aproximadamente não correlacionados e com variâncias aproximadamente iguais a um. As matrizes de transformação podem ser obtidas pelo método proposto por Fuller e Battese (1973)[98]. Naturalmente que as hipóteses nulas da normalidade dos efeitos

[97] Os critérios de informação, apesar de baseados na verosimilhança restrita, permitem comparar os modelos de estimação em domínios de Fay-Herriot com o de Salvati, bem como o modelo de Rao-Yu com o espaciotemporal, porque ambos os modelos de cada par têm os mesmos efeitos fixos.

[98] Considere-se um vector aleatório $\boldsymbol{\xi}$ tal que $E(\boldsymbol{\xi}) = \mathbf{0}$ e $V(\boldsymbol{\xi}) = \boldsymbol{\Sigma}$. Segundo o Lema 1 de Fuller e Battese (1973), se $\boldsymbol{\Sigma}$ é uma matriz $n \times n$ simétrica e definida positiva, com r valores próprios distintos λ_i, $i=1, \ldots, r$, respectivamente de multiplicidade m_i, $i=1, \ldots, r$, onde $\sum_{i=1}^{r} m_i = n$, então a transformação de $\boldsymbol{\xi}$ é efectuada através de uma matriz $\boldsymbol{\Sigma}^{-1/2}$ tal que $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \mathbf{I}_n$. Prova-se que esta relação é satisfeita por $\boldsymbol{\Sigma}^{-1/2} = \sum_{i=1}^{r} \lambda_i^{-1/2}\mathbf{A}_i$, onde $\mathbf{A}_i$ é uma matriz $n \times n$ mutuamente ortogonal, simétrica e idempotente, que satisfaz a condição $\boldsymbol{\Sigma}\mathbf{A}_i = \lambda_i \mathbf{A}_i$, e que é dada por $\mathbf{A}_i = \mathbf{B}_i\mathbf{B}_i'$, sendo $\mathbf{B}_i$ uma matriz

aleatórios transformados, $\hat{\boldsymbol{\upsilon}}^*$, bem como dos erros da sondagem transformados, $\hat{\boldsymbol{\varepsilon}}^*$, podem ser testadas através dos conhecidos testes de SW ou de KS.

Os resultados do teste à normalidade dos erros dos modelos, efectuado através do teste de SW, são apresentados nas tabelas 7.2.19 e 7.2.20. Os histogramas e os gráficos Q-Q[99] dos erros transformados encontram-se nos apêndices 9 a 12. Nestes apêndices são também apresentadas as medidas de assimetria e curtose.

Tabela 7.2.19: Resultados do teste SW à normalidade dos erros dos modelos de Fay-Herriot e de Salvati, para cada trimestre

t	**Modelo de Fay-Herriot**				**Modelo de Salvati**			
	Efeitos aleatórios		**Erros da sondagem**		**Efeitos aleatórios**		**Erros da sondagem**	
	W	***p***	***W***	***p***	***W***	***p***	***W***	***p***
1	0,9392	0,1056	0,9793	0,8328	0,9649	0,4517	0,9654	0,4647
2	0,9756	0,7367	0,9637	0,4241	0,9770	0,7746	0,9842	0,9357
3	0,9765	0,7607	0,5651	0,0001	0,9756	0,7365	0,7354	0,0001
4	0,9770	0,7740	0,5479	0,0001	0,9711	0,6117	0,7013	0,0001
5	0,9740	0,6909	0,7838	0,0001	0,9662	0,4822	0,9067	0,0165
6	0,9542	0,2517	0,9385	0,1014	0,9674	0,5126	0,9770	0,7731
7	0,9202	0,0351	0,7102	0,0001	0,9505	0,2039	0,9281	0,0551

Tabela 7.2.20: Resultados do teste SW à normalidade dos erros dos modelos de Rao-Yu e espaciotemporal

Tipo de erro	**Modelo de Rao-Yu**		**Modelo espaciotemporal**	
	W	***p***	***W***	***p***
Efeitos aleatórios de domínio	0,9686	0,5446	0,9288	0,0576
Efeitos aleatórios de domínio-tempo	0,9356	0,0001	0,9420	0,0001
Erros da sondagem	0,9329	0,0001	0,9897	0,0527

Na tabela 7.2.19 pode observar-se que a normalidade dos efeitos aleatórios do modelo de Fay-Herriot só não se verifica no trimestre 7, enquanto a normalidade dos erros da sondagem não se verifica nos trimestres 3, 4, 5 e 7. Por sua vez, existe evidência estatística para não rejeitar a hipótese da normalidade dos efeitos aleatórios do modelo de Salvati em todos os trimestres, bem como para não rejeitar a hipótese da normalidade dos erros da sondagem nos trimestres 1, 2, 6 e 7.

$n \times m_i$ dos m_i vectores próprios ortonormados associados a λ_i, i=1, …, r. Note-se que a determinação de $\boldsymbol{\Sigma}^{-1/2}$ é muito fácil quando $\boldsymbol{\Sigma}$ é uma matriz diagonal.

[99] Os gráficos Q-Q também são normalmente conhecidos por gráficos quantil-quantil.

A partir da tabela 7.2.20 conclui-se que se verifica a normalidade dos efeitos aleatórios de domínio no modelo de Rao-Yu, conquanto tal não se verifica relativamente aos efeitos aleatórios de domínio-tempo e aos erros da sondagem. Por sua vez, no âmbito do modelo espaciotemporal só não se verifica a normalidade dos efeitos aleatórios de domínio-tempo.

Note-se, no entanto, que os desvios em relação à normalidade evidenciados não parecem constituir uma grave violação aos pressupostos dos modelos de estimação em pequenos domínios em estudo por dois motivos. Por um lado, porque os desvios em relação à normalidade devem-se sobretudo ao achatamento e não à assimetria das distribuições (conforme se pode observar nos histogramas e nos gráficos Q-Q dos erros dos modelos, e se pode confirmar pelas medidas de assimetria e achatamento[100], apresentados nos apêndices 9 a 12), sendo que as referidas distribuições permanecem razoavelmente simétricas, mas consideravelmente menos achatadas do que a distribuição normal (leptocúrticas). Por outro lado, porque no contexto dos modelos utilizados, os pressupostos da normalidade dos erros só são verdadeiramente necessários na medição da incerteza dos EBLUP, ou seja, na estimação da vulgarmente designada terceira componente do EQMP dos EBLUP. A este respeito relembre-se que Lahiri e Rao (1995) mostraram que o estimador do EQMP do EBLUP proposto por Prasad e Rao (1990) é robusto à violação da hipótese da normalidade dos efeitos aleatórios, quando aplicado ao modelo de Fay-Herriot e utilizando estimativas ANOVA das componentes de variância obtidas pelo método dos momentos de Prasad-Rao, desde que se verifique a simetria dos erros do modelo. Tendo em conta este resultado, e considerando-se que as distribuições dos erros dos modelos utilizados são aproximadamente simétricas, então os afastamentos observados em relação à normalidade não parecem ser suficientes para invalidar as estimativas do EQMP *model-based* dos estimadores dos parâmetros de interesse.

[100] Note-se que numa distribuição Normal se tem g_1=0 e g_2=0.

7.2.7 Resultados do estudo por simulação *design-based*

7.2.7.1 Introdução

Em cada uma das L=1.000 réplicas deste estudo por simulação, foram calculadas estimativas da média da variável de interesse em cada NUTSIII para cada trimestre, utilizando os estimadores apresentados nas subsecções 7.2.4.2 e 7.2.4.3. Em cada réplica foram também calculadas as estimativas dos EQMP *model-based* dos estimadores EBLUP. Nesta secção pretende avaliar-se as propriedades dos estimadores do parâmetro de interesse. Relembre-se que esta avaliação é efectuada sob uma perspectiva de amostragem repetida, sendo as estimativas obtidas em cada amostra extraída da pseudo-população comparadas com os respectivos parâmetros de interesse dessa pseudo-população. Uma análise sumária da distribuição dos parâmetros de interesse na pseudo-população é apresentada na subsecção 7.2.7.2.

Com o objectivo de facilitar a análise das propriedades dos estimadores, decidiu formar-se dois grupos compatíveis de estimadores. No grupo A, formado por oito estimadores, encontram-se os quatro estimadores tradicionais e os quatro estimadores EBLUP sem restrições. O grupo B é formado pelo melhor estimador do grupo A e pelos três estimadores que verificam a consistência interna na publicação das estimativas: o estimador sintético pelo quociente e os dois estimadores EBLUP com restrições. A avaliação das propriedades dos estimadores dos grupos A e B é efectuada, respectivamente, nas subsecções 7.2.7.3 e 7.2.7.4.

Em primeiro lugar, foram calculadas as medidas de qualidade dos estimadores do grupo A ao nível individual, para as primeiras 100, 500 e para as 1.000 réplicas. Os resultados das medidas principais (enviesamento, EQM, variância e erro absoluto) são apresentados, respectivamente, nos apêndices 13, 14 e 15. Globalmente, pode verificar-se que os resultados obtidos para as 100, 500 e 1.000 réplicas são convergentes, pelo que se considera desnecessário efectuar um maior número de réplicas neste estudo. Para além disso, os erros relativos da simulação de Monte Carlo[101] com 1.000 réplicas, apresentados na tabela 7.2.21, indicam que estes são muito pequenos. Em termos

[101] Os erros relativos da simulação de Monte Carlo são calculados da seguinte forma: $Erro\ relativo\ de\ simulação = \sqrt{L^{-1}V(\hat{\mu}_{it})}/\bar{\hat{\mu}}_{it}$.

globais, o maior erro relativo médio da simulação ocorre na estimação dos parâmetros de interesse através do estimador directo (0,36%), enquanto o menor erro relativo médio verifica-se para o estimador sintético pela regressão (0,09%).

Tabela 7.2.21: Erros relativos (%) da simulação de Monte Carlo, em cada trimestre

t	$\widehat{\mu}_i^{dir}$	$\widehat{\mu}_i^{sinQ}$	$\widehat{\mu}_i^{sinR}$	$\widehat{\mu}_i^{com}$	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$
1	0,37	0,17	0,08	0,33	0,18	0,17	0,18	0,14	0,33	0,19
2	0,37	0,15	0,10	0,32	0,17	0,16	0,18	0,13	0,33	0,18
3	0,39	0,23	0,10	0,35	0,20	0,20	0,18	0,14	0,32	0,24
4	0,37	0,27	0,09	0,32	0,18	0,18	0,17	0,13	0,31	0,24
5	0,35	0,25	0,09	0,31	0,16	0,16	0,16	0,13	0,30	0,22
6	0,33	0,14	0,08	0,30	0,15	0,15	0,16	0,13	0,31	0,17
7	0,31	0,22	0,08	0,27	0,14	0,14	0,15	0,13	0,29	0,19

Desta forma, todas as análises de resultados apresentadas seguidamente são baseadas nas 1.000 réplicas. Os resultados das restantes medidas de qualidade dos estimadores do grupo A (enviesamento relativo absoluto, rácio de enviesamento absoluto, erro relativo absoluto, erro padrão relativo, coeficiente de variação, eficiência relativa, taxa de cobertura do IC *design-based* e taxa de cobertura do IC *model-based*), do estudo por simulação baseado em 1.000 réplicas, são também apresentados no apêndice 15.

A realização deste estudo empírico exigiu a programação em linguagem SAS de programas com as seguintes funções: (*i*) geração de uma pseudo-população e cálculo do verdadeiro valor dos parâmetros de interesse (apêndice 16); (*ii*) cálculo das estimativas dos parâmetros de interesse através dos estimadores tradicionais (apêndice 17); (*iii*) cálculo das estimativas dos parâmetros de interesse através dos estimadores EBLUP (apêndice 18); e (*iv*) cálculo das medidas de avaliação da qualidade dos estimadores (apêndice 19). O tempo total de processamento computacional deste estudo por simulação foi de cerca de 42 horas.

7.2.7.2 Análise exploratória da distribuição do preço médio de transacção da habitação na pseudo-população

Uma vez que este estudo empírico tem como objectivo comparar as propriedades *design-based* de um conjunto de estimadores, através da comparação das estimativas dos parâmetros de interesse produzidas em cada réplica com os verdadeiros valores desses parâmetros, em cada um dos domínios, então começa por apresentar-se uma

breve análise exploratória da distribuição do preço médio de transacção da habitação na pseudo-população (*vide* apêndice 20). Esta distribuição caracteriza-se por apresentar um valor médio global de 1.080€ e um CV de 32,6%. A referida distribuição é aproximadamente simétrica (g_1=0,91) e é leptocúrtica (g_2= 1,57). Uma característica interessante da distribuição do preço médio de transacção da habitação é que existem disparidades significativas entre as diferentes NUTSIII (domínios de interesse) em cada trimestre, bem como ao longo dos sete trimestres. Numa análise trimestral, verifica-se que os preços médios de transacção da habitação variam entre os 1.041€ no primeiro trimestre e os 1.129€ no sexto trimestre, apresentando CV pouco superiores a 30% em todos os trimestres (tabela 7.2.22).

Tabela 7.2.22: Verdadeiros valores dos parâmetros média e CV (%) do preço de transacção da habitação na pseudo-população, em cada trimestre

Trimestre	Média	CV
1	1.041	31,5%
2	1.063	32,3%
3	1.039	33,8%
4	1.092	34,1%
5	1.093	32,2%
6	1.129	32,2%
7	1.126	32,6%

7.2.7.3 Avaliação das propriedades dos estimadores do grupo A

A. Análise das medidas de enviesamento dos estimadores do grupo A

Numa primeira análise dos resultados das medidas de enviesamento ao nível das 28 NUTSIII, pode notar-se que todos os estimadores apresentam algum enviesamento. Esta primeira observação está ilustrada nos gráficos 7.2.1 e 7.2.2, nos quais são apresentados, respectivamente, o enviesamento médio e o enviesamento relativo médio dos estimadores do grupo A, ao nível de NUTSIII.

Gráfico 7.2.1: Enviesamento médio dos estimadores do grupo A, ao nível de NUTSIII

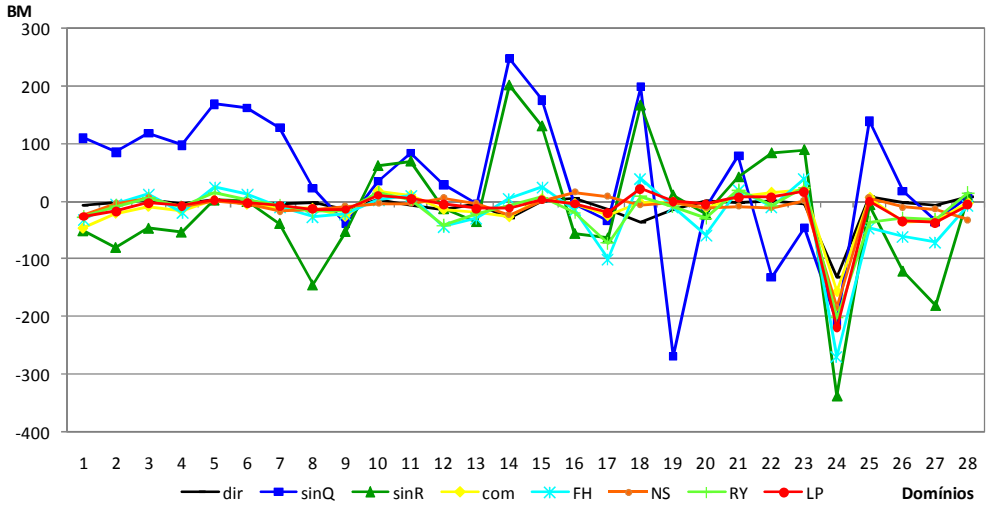
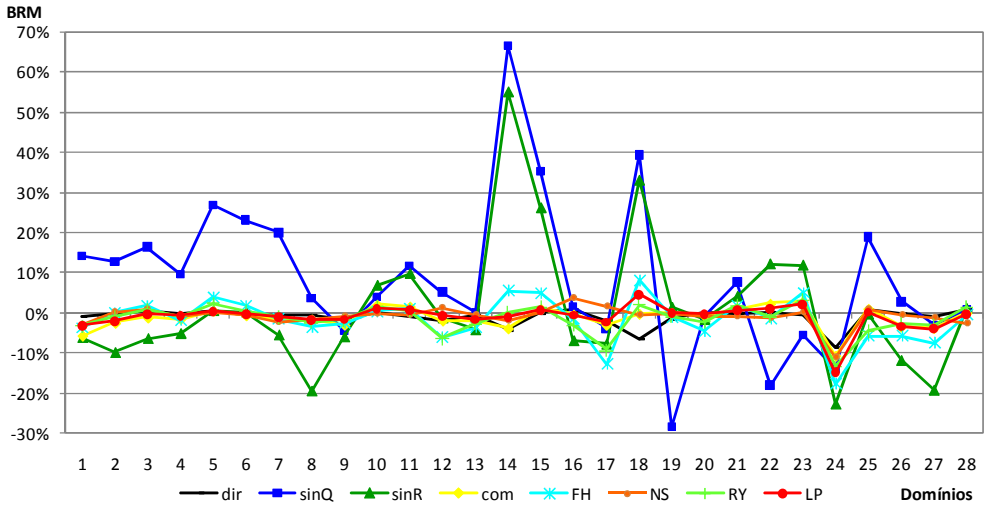


Gráfico 7.2.2: Enviesamento relativo médio dos estimadores do grupo A, ao nível de NUTSIII



A partir da observação dos gráficos 7.2.1 e 7.2.2, podem identificar-se três comportamentos distintos. Em primeiro lugar, os dois estimadores sintéticos são os que apresentam o pior comportamento em termos de enviesamento na generalidade das 28 NUTSIII, sendo particularmente mau o comportamento do estimador sintético pelo quociente. Em segundo lugar, todos os estimadores EBLUP apresentam um comportamento muito semelhante em termos de enviesamento em todas as NUTSIII. É, também, de notar que os níveis de enviesamento relativo médio dos estimadores EBLUP em nada se comparam com esses níveis de enviesamento dos estimadores sintéticos. Em terceiro lugar, o estimador directo apresenta níveis de enviesamento médio pouco expressivos, confirmando empiricamente que este estimador é aproximadamente centrado. Os enviesamentos observados neste estimador, embora de

muito pequeno valor, resultam, por um lado, do erro de simulação, e por outro lado, do facto de se imputarem valores nulos às estimativas directas quando estas não existem, ou seja, quando as dimensões amostrais dos domínios de interesse são nulas.

Os gráficos 7.2.1 e 7.2.2 ilustram, ainda, que todos os estimadores, com excepção dos sintéticos, tendem a subestimar o verdadeiro valor dos parâmetros de interesse. Verifica-se globalmente que existe subestimação do verdadeiro valor desses parâmetros em cerca de dois terços dos domínios individuais, sendo o enviesamento relativo médio global igual a -1,0% para o estimador directo, -1,3% para o estimador combinado, e igual a -1,6%, -0,7%, -1,5% e -1,0%, respectivamente, para os estimadores EBLUP de FH, NS, RY e de LP.

Numa análise ao nível de grupos de domínios, cujos resultados são apresentados na tabela 7.2.23, verifica-se que o enviesamento relativo absoluto médio e o rácio de enviesamento absoluto médio tendem a diminuir com o aumento das dimensões amostrais nos domínios. Este comportamento é globalmente semelhante para todos os estimadores. Contudo, é de destacar um comportamento atípico no grupo 1.

Tabela 7.2.23: Medidas de enviesamento médio dos estimadores do grupo A, por grupo de NUTSIII

Grupo	$\hat{\mu}_i^{dir}$	$\hat{\mu}_i^{sinQ}$	$\hat{\mu}_i^{sinR}$	$\hat{\mu}_i^{com}$	$\hat{\mu}_i^{FH}$	$\hat{\mu}_i^{NS}$	$\hat{\mu}_{it}^{RY}$	$\hat{\mu}_{it}^{LP}$
BRAM (%)								
1	2,32	52,59	42,78	2,71	10,10	5,52	3,39	2,31
2	4,80	18,53	19,70	6,16	12,90	8,76	8,09	6,82
3	1,45	16,31	11,21	3,16	6,10	4,29	3,57	2,50
4	1,25	13,74	9,29	2,70	5,42	4,76	3,04	1,55
5	1,13	12,76	9,39	2,32	3,97	2,75	1,77	1,20
6	0,64	8,60	4,63	1,63	3,29	2,38	1,86	0,89
RBAM								
1	0,13	7,90	10,53	0,17	1,59	1,81	1,87	4,02
2	0,24	2,67	8,12	0,38	2,09	0,88	0,94	0,80
3	0,10	1,94	5,77	0,30	1,60	1,03	0,80	0,68
4	0,11	2,70	3,80	0,28	1,22	0,86	0,51	0,31
5	0,11	2,86	3,87	0,26	1,20	0,69	0,36	0,23
6	0,08	2,43	1,36	0,24	1,10	0,70	0,47	0,22

Nota: BRAM–enviesamento relativo absoluto médio; RBAM–rácio de enviesamento absoluto médio.

É altura de salientar que os resultados obtidos no grupo de domínios 1, quer em termos de enviesamento, quer posteriormente em termos de precisão, não se devem exclusivamente à sua dimensão média amostral unitária, mas sim à pequena (ou nula)

variabilidade existente nos domínios com amostras superiores à unidade[102]. Embora não seja apresentado nesta tese, foi efectuado paralelamente outro estudo empírico *design-based* utilizando o mesmo plano de sondagem, mas no qual foi seleccionada uma amostra longitudinal de 458 empresas (o dobro da dimensão amostral utilizada no estudo empírico que está aqui a ser apresentado) em cada uma das 1.000 réplicas. A partir dos resultados obtidos nesse outro estudo empírico, foi possível observar um comportamento semelhante, em termos de enviesamento e de precisão, para todos os estimadores (inclusive nos domínios correspondentes ao grupo 1), ao observado no estudo empírico que está a ser apresentado neste texto (*vide* apêndice 21)[103]. Este facto parece confirmar a tal especificidade das NUTSIII do grupo 1, pois mesmo com o dobro da dimensão amostral, continuam a não ter um comportamento uniforme com os outros grupos de maiores dimensões amostrais.

Voltando à análise dos resultados deste estudo empírico, note-se que o estimador directo é, naturalmente, o estimador que apresenta melhor desempenho em termos de todas as medidas de enviesamento: o enviesamento relativo absoluto médio situa-se entre os 0,64% no grupo 6 e os 4,80% no grupo 2, enquanto o rácio de enviesamento absoluto médio está compreendido entre os 0,08 e os 0,24, respectivamente.

Também se pode observar na tabela 7.2.23 que os estimadores sintéticos são os que apresentam pior comportamento em termos de enviesamento, tal como já foi observado pelos gráficos 7.2.1 e 7.2.2. Neste subgrupo de estimadores sintéticos, o estimador pelo quociente é o mais enviesado. Este estimador não só apresenta o pior enviesamento relativo médio em 57% das 28 NUTSIII, como também apresenta rácios de enviesamento absoluto médio superiores a 2 em mais de metade das NUTSIII. Por sua vez, o estimador sintético pela regressão é o que apresenta o pior desempenho ao nível do rácio de enviesamento absoluto médio (tal como se pode observar na tabela 7.2.23), pelo facto de apresentar variâncias médias muito pequenas.

No que se refere ao estimador combinado com pesos dependentes dos dados, tal como seria de esperar, apresenta níveis de enviesamento relativo absoluto médio e de rácio de

[102] Através de uma análise exploratória aos dados do grupo 1, foi possível observar que existem domínios com todas as observações iguais.

[103] Por facilidade de apresentação, decidiu apresentar-se apenas as medidas principais (enviesamento, EQM, variância e erro absoluto), ao nível de NUTSIII, referentes aos estimadores tradicionais.

enviesamento absoluto médio compreendidos entre os respectivos níveis do estimador directo e do estimador sintético pela regressão, embora mais próximos dos valores do estimador directo. Estes resultados tornam claro o ganho potencial, em termos de enviesamento, que ocorre quando se utiliza um estimador combinado deste tipo, tal como referido no subcapítulo 2.6.

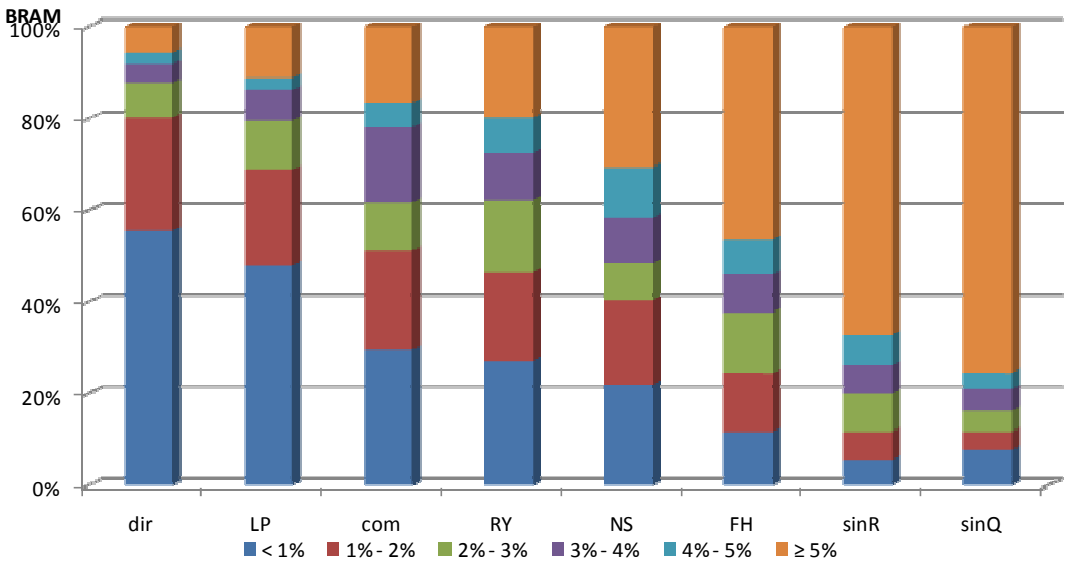
Através de uma análise comparativa das medidas de enviesamento dos estimadores EBLUP, é possível observar que os estimadores assistidos por modelos de estimação em domínios que utilizam mais informação (temporal e/ou espacial) tendem a apresentar melhor comportamento em termos de enviesamento. Estes ganhos são particularmente evidentes no estimador EBLUP espaciotemporal, o qual apresenta valores do enviesamento relativo absoluto médio pouco superiores aos do estimador directo, que é, como se sabe, teoricamente um estimador aproximadamente centrado. No grupo 1, o estimador EBLUP espaciotemporal apresenta mesmo um enviesamento relativo absoluto médio menor do que o do estimador directo. Por outro lado, o estimador EBLUP de FH é o que apresenta o pior desempenho em termos de enviesamento, naquele subgrupo de estimadores. Note-se, ainda, que o estimador EBLUP de RY apresenta um desempenho globalmente melhor do que o estimador EBLUP de NS. Este facto evidencia que a introdução de informação temporal na estrutura de erro do modelo de estimação em domínios tem um maior impacto na redução do enviesamento do estimador, do que a introdução de informação espacial. No entanto, a utilização simultânea de informação seccional/espacial e temporal na estrutura de erro do modelo é, como se referiu acima, a melhor solução, tendo em vista a obtenção de estimativas com o menor enviesamento possível.

Por último, note-se que os estimadores EBLUP apresentam globalmente, em termos de enviesamento, um desempenho muito melhor do que os estimadores sintéticos, mas ligeiramente pior do que o estimador directo e do que o estimador combinado com pesos dependentes dos dados. Contudo, note-se que existe uma excepção: o estimador EBLUP espaciotemporal tende a apresentar um comportamento melhor do que o estimador combinado com pesos dependentes dos dados. É, também, de salientar que apesar dos estimadores EBLUP serem enviesados, o seu enviesamento relativo absoluto médio é inferior a 10%, na esmagadora maioria dos casos. Em particular, o estimador EBLUP espaciotemporal apresenta um enviesamento relativo absoluto médio que se

situa entre os 0,89% no grupo 6 e os 6,82% no grupo 2, o que o torna num excelente estimador alternativo ao estimador directo, em termos de enviesamento.

No sentido de se evidenciar este facto com maior clareza, decidiu calcular-se a percentagem de domínios individuais por classes[104] de enviesamento relativo absoluto médio, para os oito estimadores em avaliação. O gráfico 7.2.3 ilustra estes resultados.

Gráfico 7.2.3: Percentagem de domínios individuais por classes de enviesamento relativo absoluto médio, para os estimadores do grupo A



No gráfico 7.2.3 pode observar-se, por exemplo, (*i*) que o estimador directo apresenta enviesamentos relativos absolutos médios não superiores a 1% em 56% dos domínios individuais, enquanto essa percentagem é de 48% para o estimador EBLUP espaciotemporal, ou (*ii*) que o estimador directo apresenta enviesamentos relativos absolutos médios não superiores a 5% em 94% dos domínios individuais, enquanto essa percentagem é de 89% para o estimador EBLUP espaciotemporal. Pelo contrário, ambos os estimadores sintéticos apresentam enviesamentos relativos absolutos médios superiores ou iguais a 5% em cerca de 80% dos domínios individuais.

Portanto, globalmente é possível concluir que o estimador directo é o melhor estimador, tendo em conta exclusivamente o enviesamento, porque é aproximadamente centrado (ou ligeiramente subenviesado), do ponto de vista *design-based*. Contudo, este

[104] As seis classes consideradas foram as seguintes: [0%; 1%[, [1%; 2%[, [2%; 3%[, [3%; 4%[, [4%; 5%[e [5%; +∞[.

estimador é apenas pouco melhor do que o estimador EBLUP espaciotemporal, o qual também se pode considerar aproximadamente centrado (ou ligeiramente subenviesado) do ponto de vista *design-based*, na maioria dos grupos de domínios.

Perante um quadro de estimadores tendencialmente enviesados (se bem que alguns deles sejam apenas ligeiramente enviesados), pretende identificar-se, numa óptica *design-based*, não só os estimadores que apresentam enviesamentos mais moderados, mas sobretudo os que apresentam também melhor comportamento em termos de precisão. Vai encetar-se, então, uma análise global das medidas de precisão apresentadas na secção 7.2.5, para todos os estimadores do grupo A.

B. Análise das medidas de precisão dos estimadores do grupo A

Na análise das medidas de precisão, começa por apresentar-se, nos gráficos 7.2.4 a 7.2.6, o erro absoluto médio, o EQM médio e a variância média dos estimadores do grupo A, ao nível das NUTSIII.

Gráfico 7.2.4: Erro absoluto médio dos estimadores do grupo A, ao nível de NUTSIII

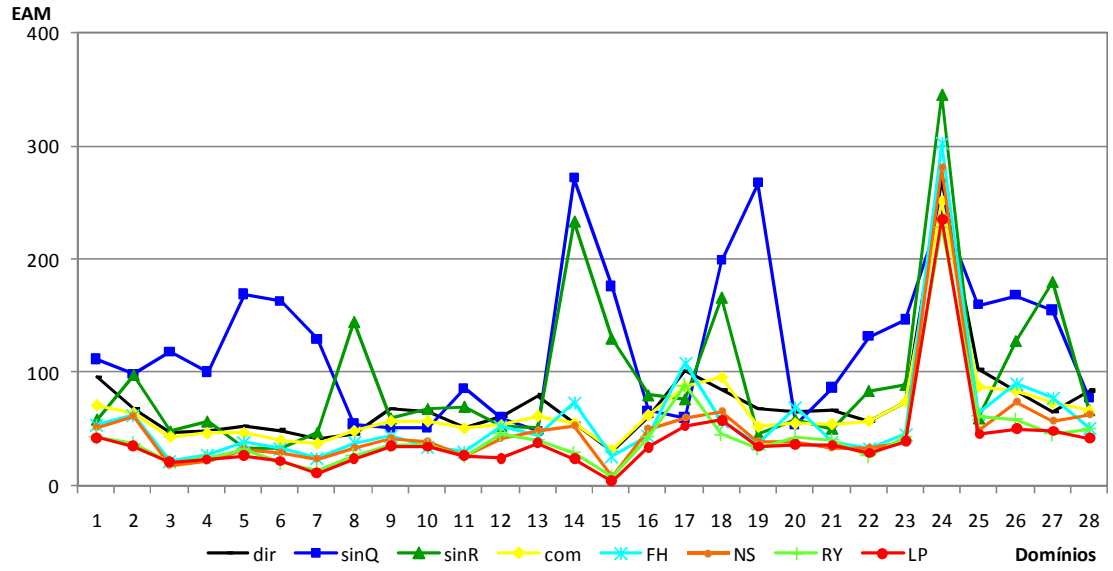


Gráfico 7.2.5: EQM médio dos estimadores do grupo A, ao nível de NUTSIII

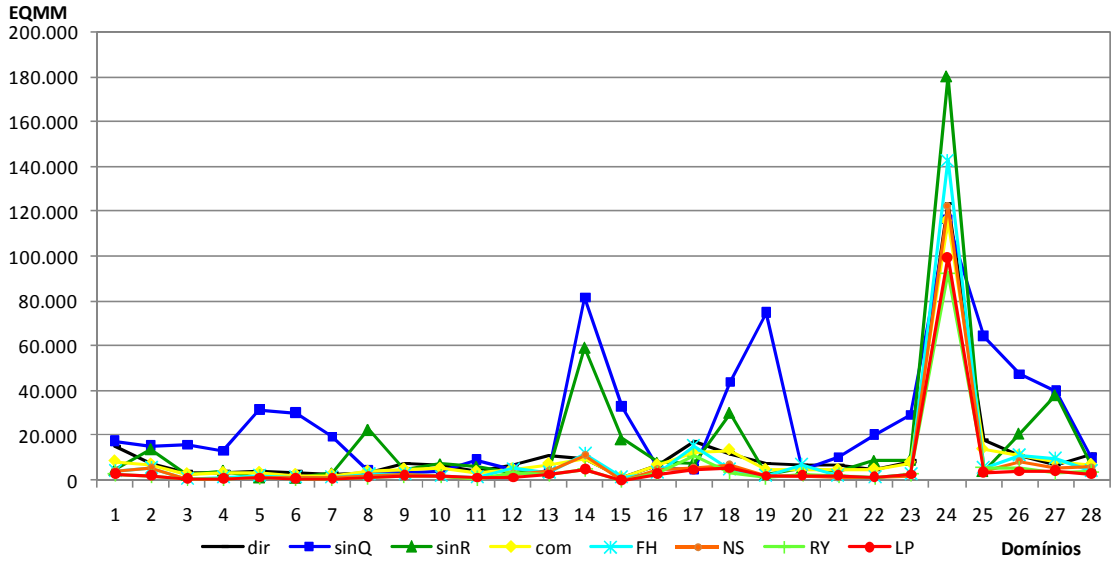
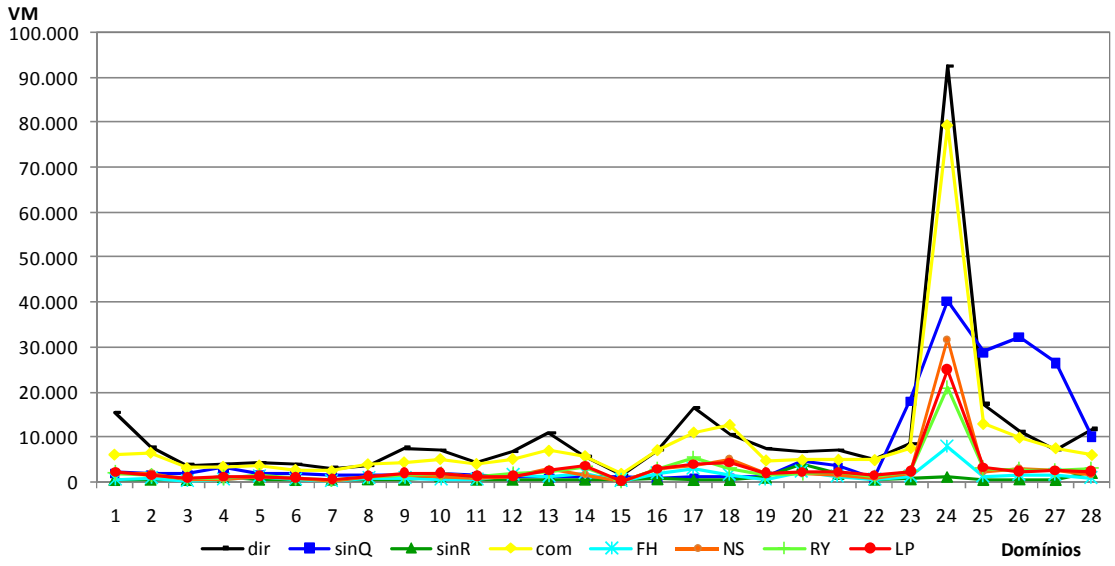


Gráfico 7.2.6: Variância média dos estimadores do grupo A, ao nível de NUTSIII



No gráfico 7.2.5 observa-se que os estimadores sintéticos apresentam um comportamento desastroso em termos de precisão, dado que apresentam os valores mais elevados no EQM médio para a generalidade das NUTSIII. Em particular, o estimador sintético pelo quociente destaca-se pela negativa em termos de EQM médio, na maioria das NUTSIII. O mau comportamento de ambos os estimadores sintéticos deve-se sobretudo ao seu elevado enviesamento, uma vez que a variância desses estimadores é das mais pequenas, tal como se pode observar no gráfico 7.2.6. Pelo gráfico 7.2.4 confirma-se a péssima prestação dos estimadores sintéticos em termos de precisão, pois

apresentam também os valores mais elevados em termos de erro absoluto médio para a maioria das NUTSIII.

Pelo contrário, os estimadores EBLUP são os que apresentam melhor desempenho em termos de precisão, uma vez que apresentam os menores valores no EQM médio. Este desempenho dos estimadores EBLUP deve-se não só às suas reduzidas variâncias, mas sobretudo ao pequeno enviesamento destes estimadores. Neste subgrupo de estimadores, destaca-se o estimador EBLUP espaciotemporal, sendo o estimador que apresenta menores valores no erro absoluto médio e no EQM médio em 93% das NUTSIII.

Da análise do gráfico 7.2.6, sobressai ainda o facto do comportamento, em termos de variância, do estimador combinado com pesos dependentes dos dados estar naturalmente compreendido entre o comportamento do estimador directo e do estimador sintético pela regressão. Contudo, o comportamento deste estimador combinado está novamente mais próximo do estimador directo, à semelhança do que acontece ao nível de enviesamento. Isto deve-se ao facto de se terem dado pesos globalmente maiores ao estimador directo do que ao estimador sintético pela regressão, resultantes das estimativas da variância do estimador directo serem pequenas, quando comparadas com o quadrado da diferença entre as estimativas directas e sintéticas do parâmetro de interesse (conforme referido na secção 2.6.3).

Pela análise dos gráficos 7.2.4 e 7.2.5, é possível observar que os estimadores directo e combinado apresentam valores médios de erro absoluto e de EQM muito próximos. Para além disso, verifica-se ainda que o estimador combinado consegue apresentar valores mais baixos nessas medidas do que o estimador sintético pela regressão, num grande número de NUTSIII.

Passa-se agora à análise das medidas de precisão relativa dos estimadores do grupo A. Em primeiro lugar, são apresentados nos gráficos 7.2.7 e 7.2.8, respectivamente, o erro relativo absoluto médio e o erro padrão relativo médio, ao nível de NUTSIII. Numa análise das medidas de precisão relativa, sobressai igualmente o mau comportamento dos estimadores sintéticos e o bom desempenho dos estimadores combinados e, em particular, dos estimadores EBLUP. De facto, todos os estimadores combinados conseguem apresentar ganhos de precisão relativamente ao estimador sintético pela

regressão. Por último, é particularmente agradável observar que o estimador EBLUP espaciotemporal é o que apresenta melhores níveis de precisão na esmagadora maioria das NUTSIII.

Gráfico 7.2.7: Erro relativo absoluto médio dos estimadores do grupo A, ao nível de NUTSIII

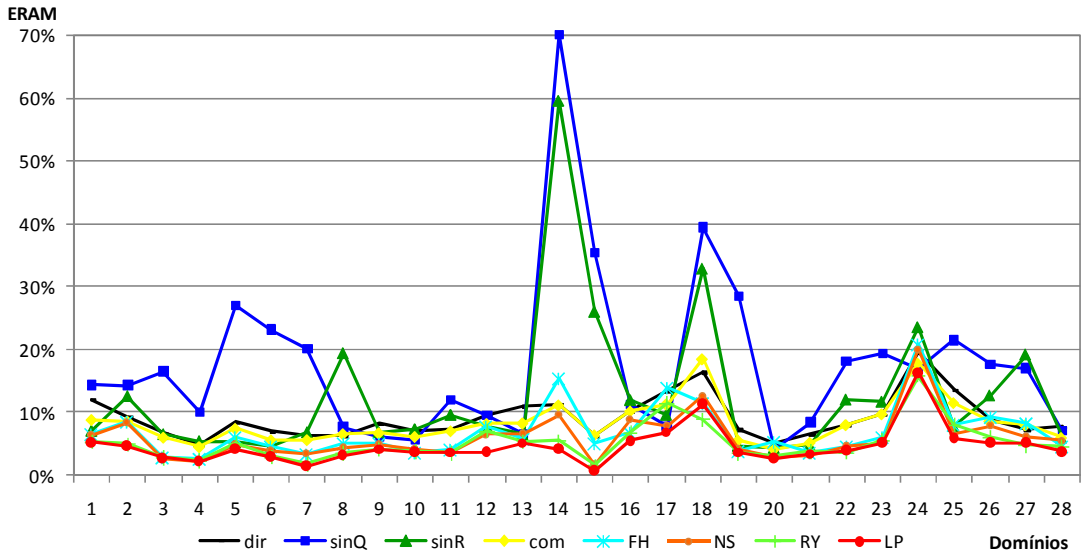
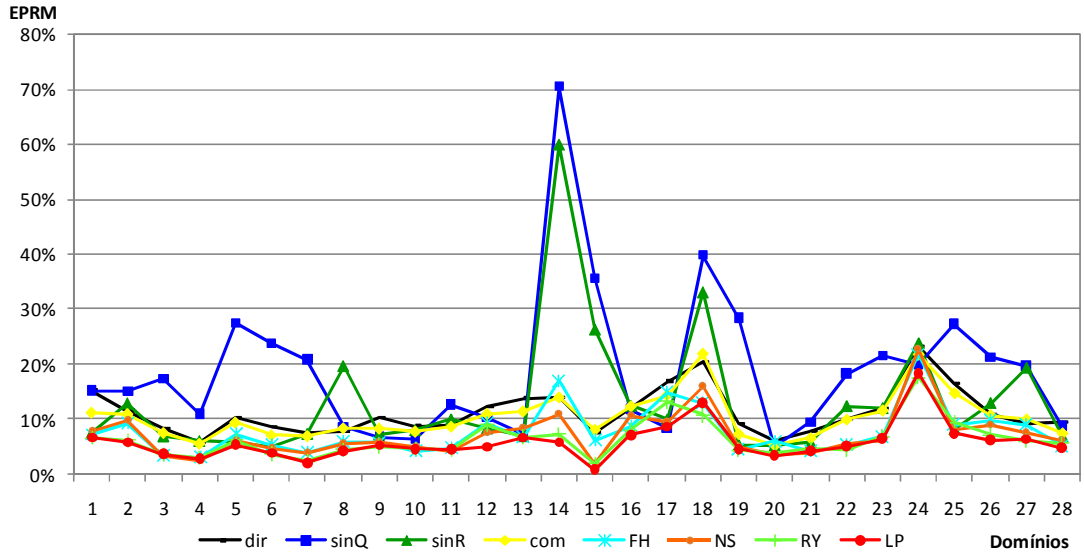


Gráfico 7.2.8: Erro padrão relativo médio dos estimadores do grupo A, ao nível de NUTSIII



Note-se ainda que os resultados apresentados no gráfico 7.2.7, relativos à medida erro relativo absoluto médio, estão em total concordância com os resultados apresentados no gráfico 7.2.8, relativos à medida erro padrão relativo médio. É apenas de salientar que os valores obtidos na segunda medida de qualidade são um pouco mais elevados, devido

ao maior peso que essa medida atribui aos valores extremos. Desta forma, os comentários efectuados aos resultados obtidos nestas duas medidas são idênticos.

Numa análise ao nível de grupos de domínios, cujos resultados são apresentados na tabela 7.2.24, verifica-se que a precisão dos estimadores tende a melhorar com o aumento das dimensões amostrais nos domínios, com excepção do grupo 1. À semelhança do que ocorria ao nível do enviesamento, esse comportamento verifica-se genericamente para todos os estimadores. Da mesma forma, o grupo 1 continua a apresentar um comportamento atípico, tal como já foi referido anteriormente.

Tabela 7.2.24: Medidas de precisão média dos estimadores do grupo A, por grupo de NUTSIII

Grupo	$\widehat{\mu}_i^{dir}$	$\widehat{\mu}_i^{sinQ}$	$\widehat{\mu}_i^{sinR}$	$\widehat{\mu}_i^{com}$	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
				ERAM (%)				
1	8,72	52,59	42,78	8,86	10,26	5,68	3,54	2,49
2	14,46	20,39	19,75	14,13	13,83	12,17	10,54	9,92
3	9,01	19,47	11,39	8,47	6,60	5,43	4,93	4,25
4	9,19	14,16	9,51	8,06	6,43	6,34	5,17	4,35
5	7,92	14,32	9,46	7,60	4,61	4,21	4,00	3,96
6	6,36	10,52	5,54	5,34	3,96	3,73	3,56	3,26
				EPRM (%)				
1	10,87	53,02	43,03	11,05	11,59	6,47	4,70	3,31
2	17,97	22,45	19,99	17,18	14,90	14,44	12,30	11,66
3	11,19	22,69	11,71	10,58	7,37	6,55	6,02	5,31
4	11,57	15,04	9,90	10,25	7,50	7,63	6,47	5,61
5	9,99	15,44	9,80	9,24	5,26	5,10	4,99	5,00
6	7,96	11,60	6,33	6,69	4,58	4,47	4,44	4,08
				CVM (%)				
1	10,07	6,33	4,05	10,21	5,38	3,06	3,06	2,06
2	16,91	9,59	2,42	15,69	6,25	10,17	8,50	8,52
3	11,04	13,29	2,29	9,98	3,56	4,42	4,56	4,52
4	11,45	4,97	2,54	9,69	4,60	5,48	5,43	5,20
5	9,90	6,53	2,41	8,82	3,19	4,00	4,52	4,78
6	7,92	5,24	3,63	6,40	2,88	3,39	3,92	3,91

Nota: ERAM–erro relativo absoluto médio; EPRM–erro padrão relativo médio; CVM–coeficiente de variação médio.

A partir da tabela 7.2.24 verifica-se que os dois estimadores sintéticos são os que apresentam o pior desempenho ao nível do erro relativo absoluto médio. A única excepção é protagonizada pelo estimador sintético pela regressão no grupo 6 (ERAM=5,54%), no qual o estimador directo é ligeiramente pior (ERAM=6,36%). É, no entanto, de salientar a redução acentuada que os estimadores sintéticos apresentam no erro relativo absoluto médio com o aumento das dimensões amostrais nos domínios. Na tabela 7.2.24 pode também observar-se um comportamento parecido, em termos de

erro relativo absoluto médio, entre os estimadores directo e combinado, com vantagem para o segundo estimador. Note-se também que o estimador combinado com pesos dependentes dos dados apresenta menor erro relativo absoluto médio do que o estimador sintético pela regressão, sobretudo nos grupos de domínios de menores dimensões amostrais. É, ainda, de salientar que os estimadores EBLUP são os que apresentam os melhores resultados globais nessa medida de qualidade, mais uma vez com clara vantagem para o estimador EBLUP espaciotemporal (o ERAM varia entre o mínimo de 2,49% no grupo 1 e o máximo de 9,92% no grupo 2).

Através de uma análise comparativa das medidas erro relativo absoluto médio e erro padrão relativo médio dos estimadores EBLUP, é possível observar que os estimadores assistidos por modelos de estimação em domínios que utilizam mais informação (temporal e/ou espacial) tendem a apresentar melhor comportamento em termos de precisão, do que aqueles que utilizam menos informação. Estes ganhos são particularmente evidentes para o estimador que utiliza simultaneamente informação temporal e espacial, o qual apresenta os menores valores nessas medidas em todos os grupos de domínios, ou seja, independentemente das dimensões amostrais nos domínios de interesse. Este resultado verifica-se globalmente para todas as NUTSIII, tal como pode ser observado nos gráficos 7.2.7 e 7.2.8.

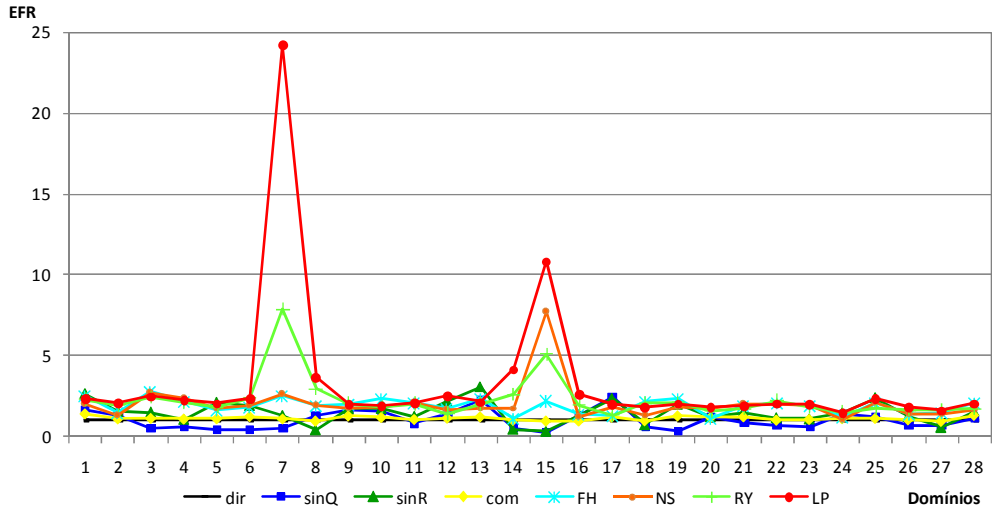
Relativamente à medida CV médio, verifica-se, pela tabela 7.2.24, que o estimador directo é o que apresenta o pior comportamento. A este respeito, relembre-se que o estimador directo é caracterizado por ser aproximadamente centrado (tal como se verificou), mas poder apresentar variâncias elevadas quando se trabalha com domínios de pequenas dimensões amostrais, como é o caso de grande parte das NUTSIII. Por outro lado, verifica-se que o estimador sintético pela regressão é o vencedor, uma vez que apresenta os menores valores no CV médio na maioria dos grupos de domínios. Este resultado relativo ao estimador sintético pela regressão já era expectável, pois este tipo de estimadores é conhecido por apresentar enviesamentos consideráveis, mas muito pequenas variâncias. Em todo o caso, verifica-se que este estimador sintético tem a concorrência dos estimadores EBLUP, que conseguem nalguns casos apresentar níveis de CV médio muito próximos (no grupo 6), ou mesmo menores (no grupo 1), do que os do estimador sintético pela regressão.

O facto dos estimadores EBLUP apresentarem níveis de CV médios competitivos com os observados para o estimador que apresenta menor variância média, aliado ao facto daqueles estimadores (especialmente o estimador EBLUP espaciotemporal) competirem com o estimador directo em termos de enviesamento, torna-os bastante precisos. De facto, esta é uma característica notável dos estimadores EBLUP. Com efeito, no grupo dos dois estimadores menos enviesados (estimador directo e estimador EBLUP espaciotemporal), o estimador que apresenta menor CV médio é o estimador EBLUP espaciotemporal, o que parece evidenciar que este é o melhor estimador para estimar o preço médio de transacção da habitação ao nível das NUTSIII de Portugal continental.

C. Análise da medida de eficiência relativa dos estimadores do grupo A

A eficiência relativa média dos estimadores do grupo A face ao estimador directo, por NUTSIII, é apresentada no gráfico 7.2.9.

Gráfico 7.2.9: Eficiência relativa média dos estimadores do grupo A face ao estimador directo, ao nível de NUTSIII



Em primeiro lugar, é de destacar que os valores de eficiência relativa média ilustrados no gráfico 7.2.9 evidenciam que o estimador EBLUP espaciotemporal é o estimador globalmente mais eficiente, embora não seja o mais eficiente em cerca de metade das NUTSIII. É, contudo, de salientar que apesar do estimador EBLUP espaciotemporal não ser o mais eficiente nesses casos, ele encontra-se muito próximo do estimador mais eficiente (que é na maior parte desses casos o estimador EBLUP de FH). Através do gráfico 7.2.9, também se pode observar que o estimador directo é pontualmente mais

eficiente do que os estimadores sintético pelo quociente, sintético pela regressão e combinado com pesos dependentes dos dados, em determinadas NUTSIII.

A eficiência relativa média dos estimadores do grupo A face ao estimador directo, por grupo de NUTSIII, é apresentada na tabela 7.2.25.

Tabela 7.2.25:Medida de eficiência relativa média dos estimadores do grupo A, por grupo de NUTSIII

Grupo	$\widehat{\mu}_i^{dir}$	$\widehat{\mu}_i^{sinQ}$	$\widehat{\mu}_i^{sinR}$	$\widehat{\mu}_i^{com}$	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
1	1,00	0,37	0,38	0,98	1,68	4,73	3,87	7,46
2	1,00	1,25	1,45	1,08	1,50	1,40	1,65	1,76
3	1,00	0,76	1,42	1,05	2,03	2,05	3,79	9,36
4	1,00	1,23	1,91	1,13	1,86	1,70	2,11	2,45
5	1,00	0,83	1,35	1,10	2,17	2,11	2,13	2,10
6	1,00	0,94	1,52	1,20	2,00	1,92	1,86	1,98

Os valores de eficiência relativa média apresentados nesta tabela evidenciam claramente que o estimador globalmente mais eficiente é o estimador EBLUP espaciotemporal. As únicas excepções encontram-se nos grupos 5 e 6, nos quais o estimador EBLUP de FH (EFRM=2,17 e EFRM=2,00), consegue ser ligeiramente mais eficiente do que o estimador EBLUP espaciotemporal (EFRM=2,10 e EFRM=1,98). Apesar de não se ter verificado empiricamente que o estimador EBLUP espaciotemporal tem EQM médio uniformemente mínimo para todos os domínios individuais, são, contudo, de salientar as suas boas qualidades por várias razões:

(*i*) Apresenta ganhos de precisão, quando comparado com todos os outros estimadores, sendo esses ganhos particularmente significativos relativamente ao estimador directo;

(*ii*) Os ganhos de precisão relativamente ao estimador directo devem-se sobretudo a ganhos "líquidos" na variância, uma vez que se verificou empiricamente que o enviesamento do estimador EBLUP espaciotemporal é apenas ligeiramente superior ao do estimador directo;

(*iii*) Os ganhos de precisão relativamente ao estimador directo são expressivos mesmo em domínios com maiores dimensões amostrais (EFRM=1,98 no grupo 6, com uma dimensão média amostral superior ou igual a 30 observações).

Em termos de eficiência relativa, é ainda possível observar na tabela 7.2.25 que o estimador directo só consegue ser significativamente mais eficiente do que o estimador sintético pelo quociente. Este ganho de eficiência verifica-se em quatro dos seis grupos de domínios. Por sua vez, é também perceptível que a combinação linear com pesos dependentes dos dados entre o estimador directo e o estimador sintético pela regressão, gera ganhos de eficiência moderados (entre os 5% e os 20%) na generalidade dos grupos de domínios.

D. Análise das taxas de cobertura dos IC dos estimadores do grupo A

Passa-se agora para a análise dos resultados do estudo empírico relativos às taxas de cobertura dos IC *design-based* e *model-based* para a média, em cada domínio individual. Uma vez que a abordagem proposta para estimação do parâmetro de interesse em cada domínio individual é do tipo *model-assisted*, no sentido em que não se supõe que a população alvo seja exactamente gerada por um modelo de superpopulação, mas apenas que possa ser aproximadamente descrita por tal modelo, então torna-se necessário conhecer o interesse e a qualidade dos vários indicadores de precisão que podem ser produzidos para os estimadores EBLUP propostos. Um sinal da importância desses indicadores é dado pela comparação das taxas de cobertura de IC alternativos, sob amostragem repetida.

A tabela 7.2.26 apresenta as taxas médias de cobertura dos IC *design-based* por grupo de NUTSIII, calculadas sob amostragem repetida, e para um grau de confiança de 95%, enquanto a tabela 7.2.27 exibe a percentagem de IC *design-based* por classe de taxa de cobertura. A análise destas duas tabelas permite identificar três comportamentos distintos.

Tabela 7.2.26:Taxas médias de cobertura dos IC *design-based* (em %) dos estimadores do grupo A, por grupo de NUTSIII

Grupo	$\widehat{\mu}_i^{dir}$	$\widehat{\mu}_i^{sinQ}$	$\widehat{\mu}_i^{sinR}$	$\widehat{\mu}_i^{com}$	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
1	94,3	0,8	0,3	94,0	58,8	56,0	57,9	25,4
2	94,7	49,3	12,8	95,2	47,6	79,2	80,4	83,4
3	94,9	57,5	21,1	94,1	59,6	77,4	83,0	86,7
4	95,0	39,5	32,3	94,0	71,4	83,4	90,5	92,5
5	94,8	39,4	19,0	95,0	72,8	86,6	92,4	94,0
6	95,0	62,9	67,9	94,5	75,1	84,6	91,4	94,0

Tabela 7.2.27: Percentagem de IC *design-based*, por classe de taxa de cobertura

Classe de Cobertura (%)	$\widehat{\mu}_i^{dir}$	$\widehat{\mu}_i^{sinQ}$	$\widehat{\mu}_i^{sinR}$	$\widehat{\mu}_i^{com}$	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
[0 ; 40 [	0,0	46,4	64,3	0,0	19,4	4,6	3,1	5,6
[40 ; 60 [	0,0	11,2	5,1	0,0	16,3	7,1	1,5	1,0
[60 ; 80 [	0,0	14,3	11,7	0,0	20,9	20,9	12,8	7,1
[80 ; 90 [	1,0	9,7	8,2	4,1	16,3	22,4	19,4	13,3
[90 ; 100]	99,0	18,4	10,7	95,9	27,0	44,9	63,3	73,0

Em primeiro lugar, verifica-se, pela tabela 7.2.26, que os dois estimadores sintéticos apresentam taxas médias de cobertura que nem atingem os 70% no grupo 6, o que os torna nos piores estimadores, de acordo com este indicador. O mau desempenho destes estimadores pode ser confirmado também pela tabela 7.2.27, na qual se observa que o estimador sintético pela regressão proporciona taxas médias de cobertura inferiores a 80% em cerca de 81% dos domínios individuais, enquanto o estimador sintético pela regressão proporciona taxas médias de cobertura inferiores a 80% em cerca de 72% dos domínios individuais. Estes resultados não são de estranhar, dados os elevados rácios de enviesamento absolutos médios obtidos para esses dois estimadores, o que ilustra bem as fragilidades da estimação sintética, e em particular da estimação sintética pela regressão, num contexto de amostragem repetida.

Em segundo lugar, as tabelas 7.2.26 e 7.2.27 evidenciam que os estimadores directo e combinado com pesos dependentes dos dados apresentam taxas médias de cobertura dos IC que se situam numa pequena vizinhança de 95%, e que a esmagadora maioria (99,0% para o estimador directo e 95,9% para o estimador combinado) dos IC *design-based* estão contidos na classe de cobertura [90%; 100%]. O excelente desempenho destes estimadores deve-se aos baixos rácios de enviesamento absoluto médios (conforme tabela 7.2.23). Apesar do comportamento desses dois estimadores ser semelhante, é de notar que as taxas médias de cobertura associadas ao estimador directo tendem a aumentar com a dimensão amostral, o que não se verifica com o referido estimador combinado.

Em terceiro lugar, pode identificar-se o subgrupo dos estimadores EBLUP. Estes estimadores tendem a apresentar taxas médias de cobertura tanto mais elevadas quanto maiores forem as dimensões amostrais e mais informação for incluída no modelo que assiste a estimação. Destaca-se, em particular, a superioridade do estimador EBLUP espaciotemporal face aos restantes estimadores EBLUP, o qual apresenta uma taxa

média de cobertura global de aproximadamente 86%, e cerca de 73% dos IC *design-based* estão contidos na classe de cobertura [90%; 100%].

Em termos globais, é de salientar que a maioria dos estimadores se comporta de forma semelhante, apresentando taxas médias de cobertura que tendem a aumentar com a dimensão amostral. A este respeito relembre-se que se observou, na análise das medidas de enviesamento, que os rácios de enviesamento absoluto médios tendiam a diminuir com o aumento das dimensões amostrais nos domínios. De facto, verifica-se agora que a rácios de enviesamento absoluto médios superiores a 0,6 correspondem taxas médias de cobertura inferiores a 90%.

Por último, é de notar que todos os estimadores combinados apresentam taxas médias de cobertura dos IC *design-based* muito superiores às do estimador sintético pela regressão. Destaca-se a grande importância deste resultado, o qual só por si revela uma grande vantagem em utilizar um estimador combinado, qualquer que ele seja, em detrimento do estimador sintético pela regressão.

A tabela 7.2.28 apresenta as taxas médias de cobertura dos IC *model-based* produzidos pelos estimadores EBLUP, por grupo de domínios, calculadas sob amostragem repetida. Por sua vez, a tabela 7.2.29 expõe a percentagem de IC *model-based* por classe de taxa de cobertura.

Tabela 7.2.28: Taxas médias de cobertura dos IC *model-based* (em %) dos estimadores EBLUP, por grupo de NUTSIII

Grupo	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
1	86,9	93,1	97,5	97,5
2	42,2	56,7	78,3	81,0
3	72,5	83,0	90,9	92,0
4	79,5	82,2	87,9	88,6
5	91,3	92,6	94,6	94,4
6	84,7	87,7	97,0	97,2

Tabela 7.2.29: Percentagem de IC *model-based*, por classe de taxa de cobertura

Classe de Cobertura (%)	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$
[0 ; 40 [	15,8	6,1	1,5	1,5
[40 ; 60 [	7,7	6,1	2,6	2,0
[60 ; 80 [	14,8	17,9	9,2	8,2
[80 ; 90 [	10,7	18,4	14,8	10,7
[90 ; 100]	51,0	51,5	71,9	77,6

A partir da análise da tabela 7.2.28, observa-se que as taxas médias de cobertura não apresentam uma tendência linear definida com o aumento da dimensão amostral dos domínios, ao contrário do que se observou nas taxas médias de cobertura dos IC *design-based*. Verifica-se também, que as taxas médias de cobertura dos IC *model-based* apresentam um afastamento moderado do grau de confiança definido, sobretudo nos grupos 2 e 4, sendo esse afastamento mais acentuado para o caso dos estimadores EBLUP de FH e de NS, ou seja, para o caso dos estimadores que não usam informação temporal. A tabela 7.2.28 ilustra ainda que o estimador EBLUP espaciotemporal volta a ser o melhor, uma vez que é o que apresenta taxas médias de cobertura do IC *model-based* uniformemente mais elevadas e a maior percentagem de domínios individuais na melhor classe de cobertura (77,6%, de acordo com a tabela 7.2.29).

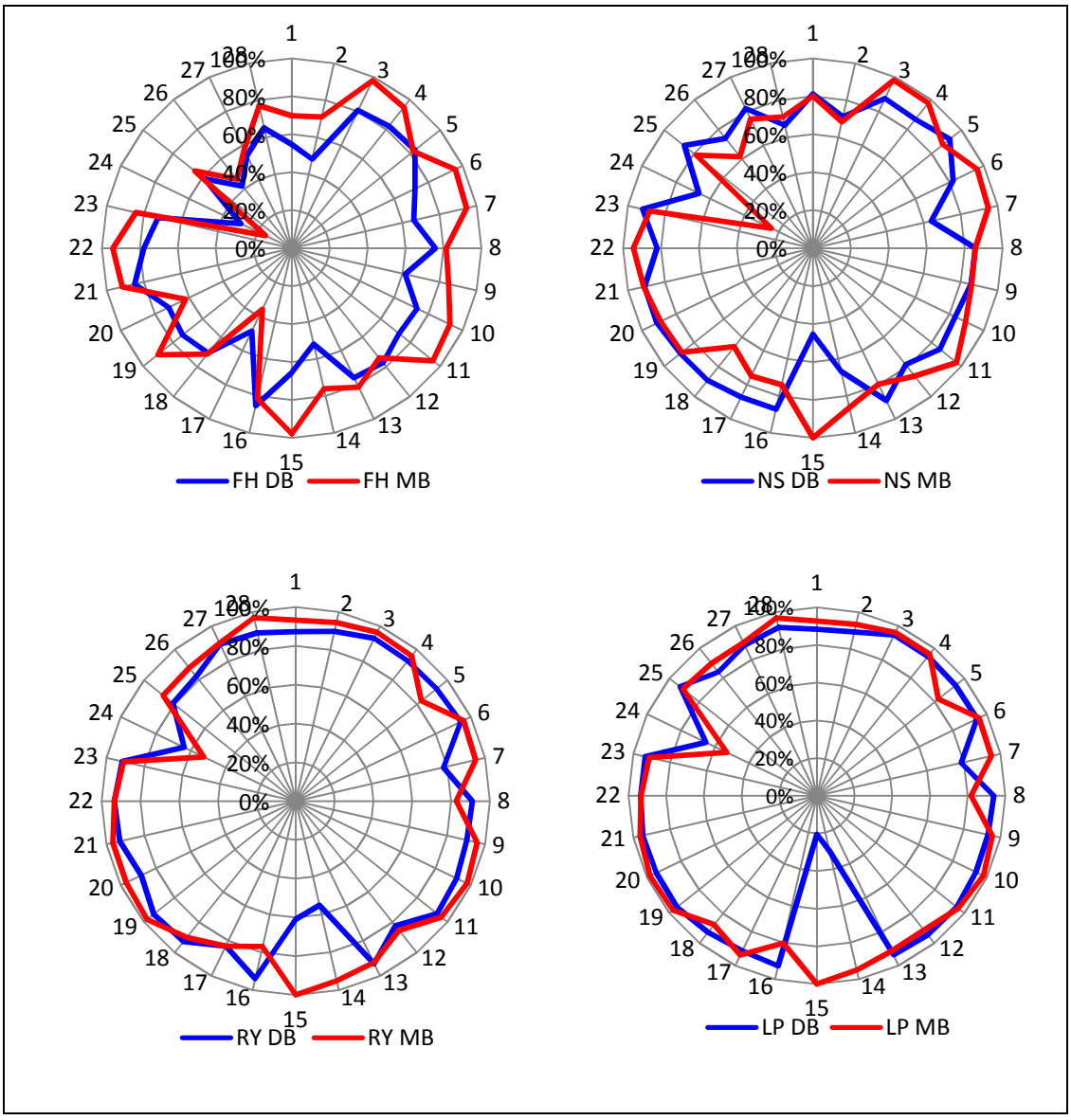
Importa agora comparar as taxas de cobertura dos IC *model-based* com as taxas de cobertura dos IC *design-based*. Tal pode ser efectuado pela observação conjunta das tabelas 7.2.26 e 7.2.28 ou das tabelas 7.2.27 e 7.2.29. Para facilitar esta comparação foram produzidos gráficos em radar com as taxas de cobertura *design-based* (em azul) e *model-based* (em vermelho), para cada estimador EBLUP ao nível de NUTSIII (figura 7.2.3).

A partir da figura 7.2.3 é possível observar que as taxas médias de cobertura dos IC *model-based* tendem a ser superiores às taxas médias de cobertura dos IC *design-based*. Para além disso, observa-se que a diferença entre as duas taxas médias de cobertura diminui com o aumento da utilização de informação espacial e/ou temporal nos modelos que assistem a estimação. Em particular, as diferenças entre as referidas taxas médias de cobertura são muito pouco expressivas no caso do estimador EBLUP espaciotemporal.

Este resultado é muito animador, pois parece confirmar a boa qualidade das estimativas *model-based* do EQMP do EBLUP espaciotemporal, uma vez que estas estimativas plagiam, em geral, as estimativas da variância sob amostragem repetida, pelo facto do enviesamento do estimador EBLUP espaciotemporal ser ligeiro. Verifica-se assim, que as boas taxas de cobertura dos IC *model-based* se devem à boa qualidade do estimador *model-based* do EQMP e não às elevadas amplitudes dos IC *model-based* [105].

[105] Note-se que é de esperar que IC baseados no EQMP *model-based* apresentem uma probabilidade de cobertura superior ao nível de confiança definido. De facto, quando se está perante estimadores

Figura 7.2.3:Taxas médias de cobertura dos IC *design-based* e *model-based* dos estimadores EBLUP, ao nível de NUTSIII



Nota: FH DB – IC *design-based* do estimador de Fay-Herriot; FH MB - IC *model-based* do estimador de Fay-Herriot; NS DB – IC *design-based* do estimador de Nicola Salvati; NS MB - IC *model-based* do estimador de Nicola Salvati; RY DB – IC *design-based* do estimador de Rao-Yu; RY MB - IC *model-based* do estimador de Rao-Yu; LP DB – IC *design-based* do estimador Luís Pereira; LP MB - IC *model-based* do estimador de Luís Pereira.

enviesados, o preço a pagar pelo aumento da taxa de cobertura é o aumento da amplitude dos intervalos devido a elevados EQMP, quando comparada com a amplitude que poderia ser obtida através de um IC baseado na variância.

7.2.7.4 Avaliação das propriedades dos estimadores do grupo B

Vai agora apresentar-se a análise das medidas de qualidade dos estimadores do grupo B. Uma vez que se concluiu que o estimador EBLUP espaciotemporal é o melhor estimador do grupo A, então o grupo B é formado por esse estimador, pelo estimador sintético pelo quociente e pelos estimadores EBLUP espaciotemporais com restrições. Relembre-se que o estimador sintético pelo quociente e um dos estimadores EBLUP espaciotemporais com restrições (LPR2) garantem a consistência interna na publicação das estimativas ao nível de NUTSII, enquanto o outro estimador EBLUP espaciotemporal com restrições (LPR1) garante essa consistência interna ao nível de Portugal continental, conforme foi definido na subsecção 7.2.4.3. Os resultados das medidas de qualidade, ao nível individual, dos estimadores EBLUP espaciotemporais com restrições estão compilados no apêndice 22.

A. Análise das medidas de enviesamento dos estimadores do grupo B

O enviesamento médio e o enviesamento relativo médio dos estimadores do grupo B estão ilustrados, respectivamente, nos gráficos 7.2.10 e 7.2.11.

Gráfico 7.2.10: Enviesamento médio dos estimadores do grupo B, ao nível de NUTSIII

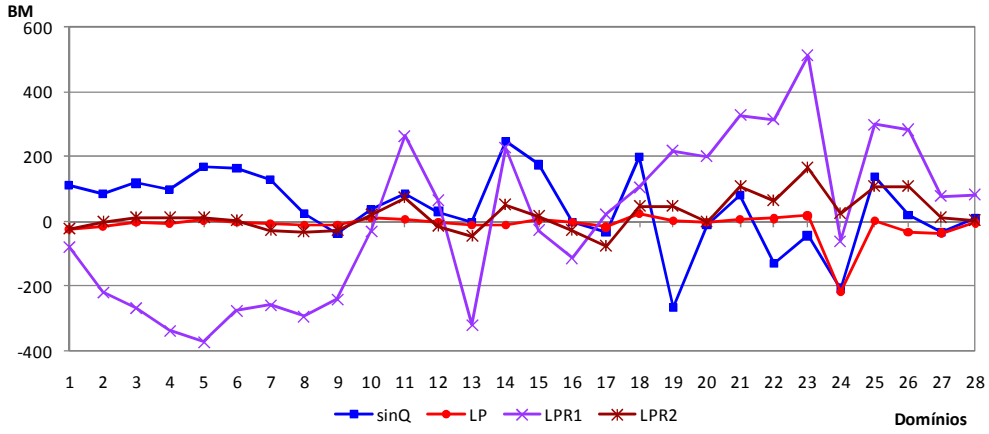
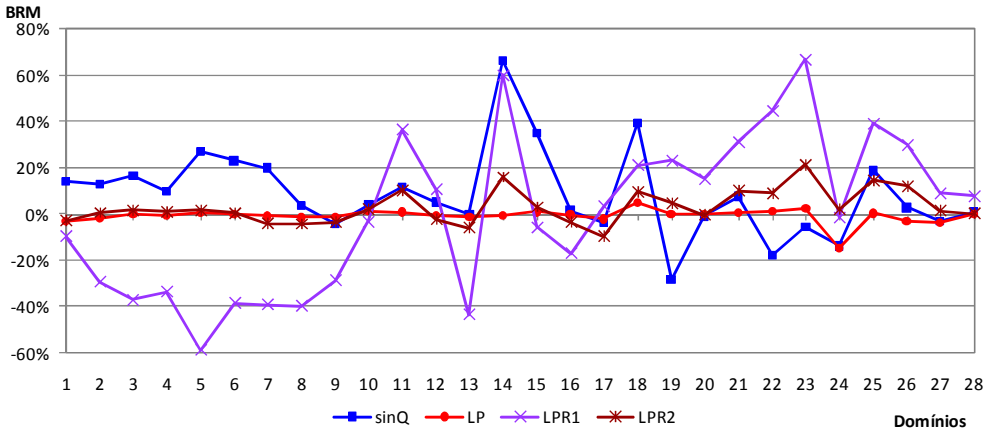


Gráfico 7.2.11: Enviesamento relativo médio dos estimadores do grupo B, ao nível de NUTSIII



Nestes gráficos pode observar-se que a garantia da consistência interna na publicação das estimativas leva ao aumento do enviesamento dos estimadores, embora o aumento do enviesamento do estimador EBLUP espaciotemporal que garante a consistência interna ao nível de NUTSII (LPR2) seja marginal para a maioria das NUTSIII. Através da observação dos gráficos 7.2.10 e 7.2.11, também sobressai que a garantia da consistência ao nível de Portugal continental (através do estimador LPR1) é a que conduz aos maiores enviesamentos médios para a generalidade das 28 NUTSIII. É, também, de notar que não existe um padrão claro de subestimação ou de sobreestimação quando se exige a garantia da consistência interna. Contudo, consegue identificar-se que o estimador sintético pelo quociente tende a sobreestimar o verdadeiro valor dos parâmetros nas NUTSIII da região *Norte*, ao contrário do estimador LPR1 que tende a subestimá-lo; e que os estimadores LPR1 e LPR2 tendem a sobreestimar esse verdadeiro valor nas NUTSIII da região da *Grande Lisboa*, *Alentejo* e *Algarve* (NUTSIII 19 a 28).

Passa-se agora para uma análise ao nível de grupos de domínios, cujas medidas de enviesamento médio dos estimadores do grupo B são apresentadas na tabela 7.2.30. Pela análise desta tabela, verifica-se, de imediato, que a garantia da consistência interna na publicação das estimativas conduz ao aumento do enviesamento, tal como também já tinha sido observado através dos gráficos 7.2.10 e 7.2.11.

Tabela 7.2.30:Medidas de enviesamento médio dos estimadores do grupo B, por grupo de NUTSIII

Grupo	$\widehat{\mu}_{it}^{sinQ}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$	$\widehat{\mu}_{it}^{sinQ}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$
	BRAM (%)				RBAM			
1	52,59	2,31	33,22	13,84	7,90	4,02	2,74	2,05
2	18,53	6,82	18,08	8,84	2,67	0,80	2,24	0,84
3	16,31	2,50	29,45	8,32	1,94	0,68	3,33	0,90
4	13,74	1,55	30,91	5,13	2,70	0,31	3,22	0,99
5	12,76	1,20	42,63	9,35	2,86	0,23	3,96	1,72
6	8,60	0,89	19,12	3,49	2,43	0,22	2,92	0,89

Nota: BRAM–enviesamento relativo absoluto médio; RBAM–rácio de enviesamento absoluto médio.

Relativamente ao estimador sintético pelo quociente, é mais uma vez de salientar o seu mau desempenho em termos de enviesamento, pois os seus níveis de enviesamento relativo absoluto médio são significativamente superiores aos do estimador EBLUP espaciotemporal que também garante a consistência interna ao nível de NUTSII (LPR2). Da mesma forma, os rácios de enviesamento absoluto médio do estimador sintético são muito elevados, o que conduz a péssimas taxas médias de cobertura dos intervalos de confiança *design-based* (conforme tabela 7.2.31).

O aumento do enviesamento é muito acentuado quando se exige a garantia da consistência interna ao nível de Portugal continental, uma vez que o enviesamento relativo absoluto médio mínimo passa a ser igual a 18% (grupo 2) e os rácios de enviesamento absoluto médios são superiores a 2 em todos os grupos de domínios. Contudo, pode verificar-se, pela tabela 7.2.30, apenas uma moderada deterioração no enviesamento relativo absoluto e no rácio de enviesamento absoluto médios quando se exige a garantia da consistência interna ao nível de NUTSII, através do estimador EBLUP. Note-se, em particular, que apenas num grupo de domínios o enviesamento relativo absoluto médio é superior a 10% (grupo 1), e em quatro dos seis grupos de domínios os rácios de enviesamento absoluto médios são inferiores à unidade. Isto conduz a taxas médias de cobertura dos IC *design-based* na ordem dos 80% nesses grupos (tabela 7.2.31). É, ainda, de salientar que os enviesamentos relativos absolutos médios tendem a decrescer quando se exige a garantia da consistência interna ao nível de NUTSII, mas tal não se verifica quando se exige a garantia da consistência interna ao nível de Portugal continental.

B. Análise das medidas de precisão dos estimadores do grupo B

As medidas de precisão absoluta dos estimadores do grupo B, designadamente o erro absoluto médio, o EQM médio e a variância média, ao nível de NUTSIII, estão ilustradas nos gráficos 7.2.12 a 7.2.14.

Gráfico 7.2.12: Erro absoluto médio dos estimadores do grupo B, ao nível de NUTSIII

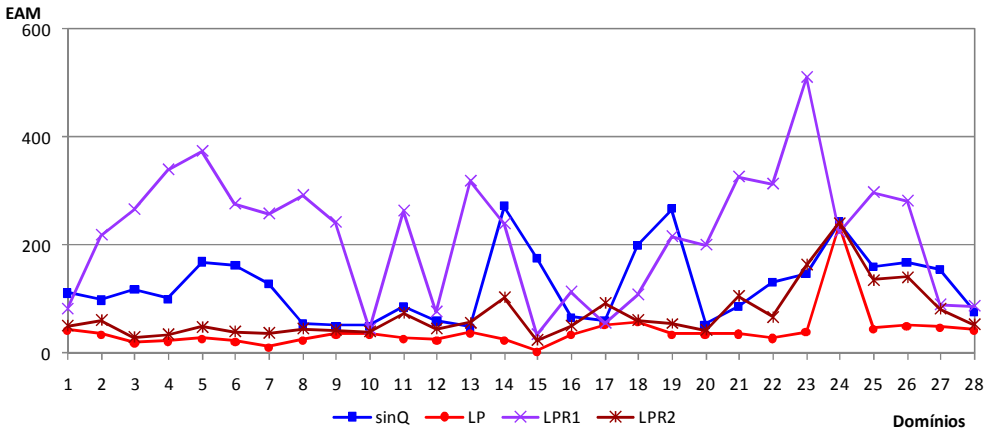


Gráfico 7.2.13: EQM médio dos estimadores do grupo B, ao nível de NUTSIII

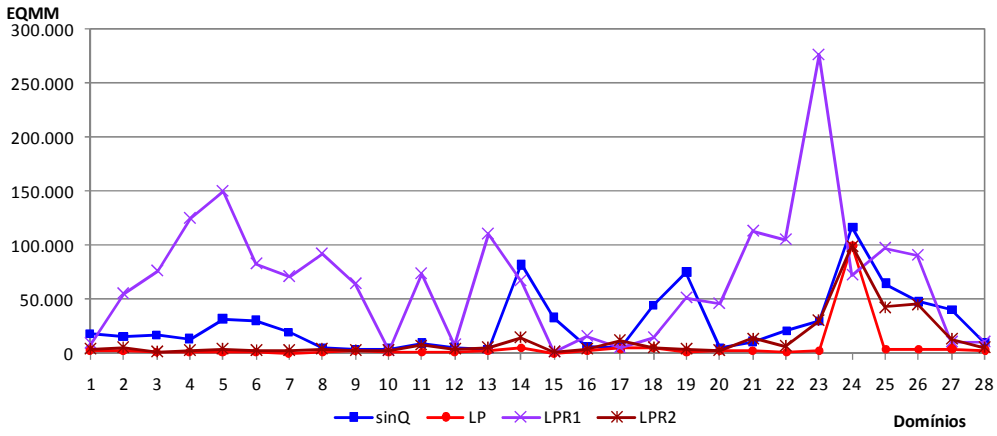
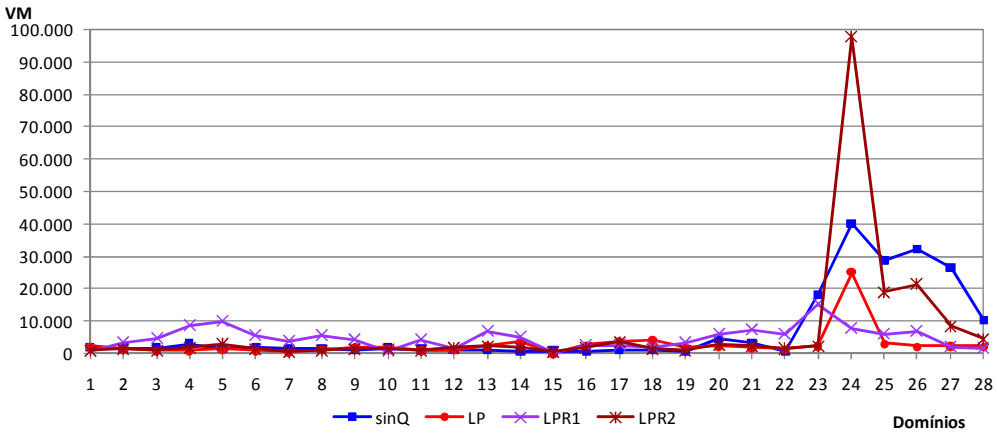


Gráfico 7.2.14: Variância média dos estimadores do grupo B, ao nível de NUTSIII



Comece-se pela análise do gráfico 7.2.14. Este gráfico ilustra que as variâncias médias dos estimadores que garantem a consistência interna são globalmente da mesma ordem de grandeza das variâncias do estimador EBLUP que não garante essa consistência interna. As excepções encontram-se nas variâncias dos estimadores que garantem a consistência interna ao nível de NUTSII (*sinQ* e *LPR*2), em NUTSIII que pertencem à NUTSII *Alentejo* (NUTSIII 24, 25, 26 e 27).

Pelos gráficos 7.2.12 e 7.2.13, pode observar-se que todos os estimadores que garantem a consistência interna apresentam maiores valores do erro absoluto médio e no EQM médio, para a generalidade das NUTSIII, do que o estimador EBLUP espaciotemporal sem restrições. Estes resultados devem-se sobretudo a maiores enviesamentos, uma vez que as variâncias dos quatro estimadores do grupo B são globalmente da mesma ordem de grandeza. Através da análise dos gráficos 7.2.12 e 7.2.13, destaca-se ainda que o melhor estimador, em termos de precisão, que garante a consistência interna na publicação das estimativas é o estimador EBLUP espaciotemporal LPR2.

Na tabela 7.2.31 estão expostas as medidas de precisão relativa média associadas aos estimadores do grupo B. Por conveniência de exposição, decidiu apresentar-se também nesta tabela as taxas médias de cobertura dos IC *design-based*.

Tabela 7.2.31: Medidas de precisão média e taxas médias de cobertura dos IC *design-based* dos estimadores do grupo B, por grupo de NUTSIII

Grupo	$\widehat{\mu}_{it}^{sinQ}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$	$\widehat{\mu}_{it}^{sinQ}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$
	ERAM (%)				EPRM (%)			
1	52,59	2,49	34,24	14,47	53,02	3,31	35,69	15,64
2	20,39	9,92	18,93	14,06	22,45	11,66	20,28	16,39
3	19,47	4,25	29,60	10,93	22,69	5,31	30,76	13,57
4	14,16	4,35	31,16	7,03	15,04	5,61	32,49	8,31
5	14,32	3,96	42,63	10,18	15,44	5,00	43,95	11,26
6	10,52	3,26	19,33	5,28	11,60	4,08	20,23	6,19
	CVM (%)				TCDBM (%)			
1	6,33	2,06	10,36	6,04	0,8	25,4	30,8	49,7
2	9,59	8,52	7,60	11,99	49,3	83,4	42,9	79,1
3	13,29	4,52	8,20	9,77	57,5	86,7	18,8	80,3
4	4,97	5,20	9,12	5,82	39,5	92,5	18,9	77,6
5	6,53	4,78	10,65	5,15	39,4	94,0	2,6	56,8
6	5,24	3,91	5,92	4,30	62,9	94,0	24,7	78,3

Nota: ERAM–erro relativo absoluto médio; EPRM–erro padrão relativo médio; CVM–coeficiente de variação médio; TCDBM-taxa de cobertura do intervalo de confiança *design-based* média.

A partir da tabela 7.2.31, pode observar-se que o estimador EBLUP espaciotemporal que garante a consistência interna ao nível de NUTSII apresenta menores valores no erro relativo absoluto médio e no erro padrão relativo médio, do que o estimador sintético pelo quociente. Isto indica que o estimador EBLUP modificado pelas restrições é mais eficiente do que o estimador sintético, apesar das pequenas variâncias médias (e coeficientes de variação médios) deste último estimador.

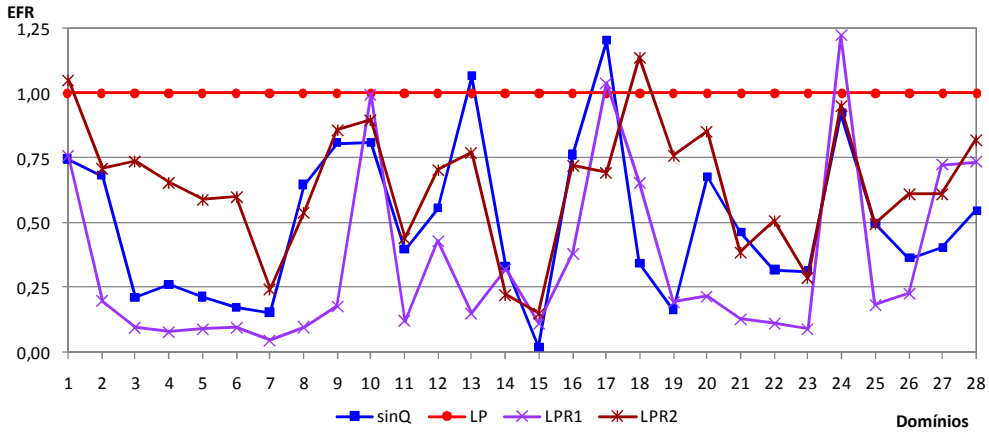
À semelhança do que se verifica com o enviesamento, a perda de precisão é também mais acentuada quando se exige a garantia da consistência interna ao nível de Portugal continental, sobretudo devido ao elevado enviesamento, uma vez que os coeficientes de variação médios nem chegam a atingir os 11%. Por seu lado, quando a garantia da consistência interna é exigida ao nível de NUTSII, observa-se apenas uma ligeira perda de precisão do estimador LPR2, quando comparado com uma situação sem garantia da consistência interna. De facto, os erros padrão relativos médios passam a situar-se entre os 6% e os 17%, enquanto numa situação sem restrições se situam entre os 4% e os 12%. É, também, de realçar que o estimador EBLUP que garante a consistência interna ao nível de NUTSII tende a apresentar melhores níveis de precisão com o aumento das dimensões amostrais nos domínios, à semelhança do que se verifica globalmente para todos os estimadores EBLUP sem restrições. Por último, é de destacar que nos três grupos de maiores dimensões amostrais (grupos 4, 5 e 6), os CV médios do estimador EBLUP sem restrições e do estimador EBLUP que garante a consistência interna ao nível de NUTSII são muito próximos, indicando que a perda de eficiência do último estimador nesses grupos se deve especialmente ao aumento do enviesamento.

Em jeito de conclusão, pode afirmar-se que a garantia da consistência entre a média ponderada das estimativas indirectas referentes às NUTSIII e as respectivas estimativas directas ao nível de NUTSII ou de Portugal continental, através da introdução de restrições no modelo que assiste a estimação, tem como resultado o aumento do enviesamento e a perda de precisão. Pode ainda afirmar-se que no grupo dos estimadores que garante a consistência interna, é ao estimador EBLUP espaciotemporal LPR2 que estão associados os menores enviesamentos e os melhores níveis de precisão.

C. Análise da medida de eficiência relativa dos estimadores do grupo B

A eficiência relativa média dos estimadores do grupo B face ao estimador EBLUP espaciotemporal sem restrições, por NUTSIII, é apresentada no gráfico 7.2.15.

Gráfico 7.2.15: Eficiência relativa média dos estimadores do grupo B face ao estimador EBLUP espaciotemporal sem restrições, ao nível de NUTSIII



O gráfico 7.2.15 ilustra que o estimador EBLUP espaciotemporal sem restrições é o estimador globalmente mais eficiente do grupo B. Este resultado indica que a garantia da consistência interna na publicação das estimativas conduz a perdas de precisão, independentemente do estimador utilizado.

A eficiência relativa média dos estimadores do grupo B face ao estimador EBLUP espaciotemporal sem restrições, por grupo de NUTSIII, é apresentada na tabela 7.2.32. Os resultados apresentados nesta tabela confirmam que a menor perda de precisão, quando se pretende garantir a consistência interna na publicação das estimativas, ocorre quando se utiliza o estimador EBLUP espaciotemporal com restrições ao nível de NUTSII (LPR2).

Tabela 7.2.32: Medida de eficiência relativa média dos estimadores do grupo B, por grupo de NUTSIII

Grupo	$\widehat{\mu}_{it}^{sinQ}$	$\widehat{\mu}_{it}^{LP}$	$\widehat{\mu}_{it}^{LPR1}$	$\widehat{\mu}_{it}^{LPR2}$
1	0,18	1,00	0,22	0,19
2	0,74	1,00	0,54	0,77
3	0,35	1,00	0,32	0,45
4	0,59	1,00	0,40	0,75
5	0,41	1,00	0,12	0,57
6	0,49	1,00	0,39	0,73

7.2.7.5 Síntese e discussão dos resultados

Os resultados obtidos no estudo por simulação *design-based* indicaram que o estimador directo é o que apresenta o melhor comportamento em termos de enviesamento. Na realidade, este estimador revelou-se aproximadamente centrado, tendo apresentado enviesamentos relativos muito pequenos, mesmo para os domínios de muito pequena dimensão. Contudo, os resultados indicaram um desempenho pobre do estimador directo em termos de precisão, devido às elevadas variâncias (próximas do EQM). Estes resultados empíricos estão, naturalmente, de acordo com os resultados teóricos, uma vez que este estimador é caracterizado por ser aproximadamente centrado, embora possa apresentar variâncias elevadas em alguns pequenos domínios (a este respeito veja, por exemplo, Cochran (1977), Särndal *et al.* (1992) ou Rao (2003)). Relativamente ao estimador directo, é ainda de salientar que as taxas médias de cobertura dos IC *design-based* observadas no estudo empírico estão muito próximas do nível de confiança definido, de acordo com o que refere Särndal *et al.* (1992).

Verificou-se, a partir dos resultados do estudo empírico, que os estimadores sintéticos apresentam os maiores enviesamentos, embora sejam, em geral, os estimadores mais competitivos em termos de variância. Por um lado, o estimador sintético pelo quociente revelou-se globalmente mais eficiente do que o estimador directo, tendo o ganho de eficiência sido obtido à custa de variâncias mais pequenas. Este facto conduziu a elevados rácios de enviesamento absolutos, que comprometeram a construção de IC *design-based* com boas taxas de cobertura. Por outro lado, apesar do estimador sintético pela regressão ter apresentado as variâncias mais pequenas, não conseguiu evidenciar ganhos de eficiência relativamente ao estimador directo, devido ao seu elevado enviesamento. Da mesma forma, os elevados rácios de enviesamento absolutos deste estimador levaram a que os IC baseados na variância *design-based* fossem pobres, revelando taxas médias de cobertura muito baixas para a generalidade dos domínios em análise. Estes resultados empíricos, para além de espelharem fielmente os comentários de Särndal *et al.* (1992, p. 410) acerca da estimação sintética, estão também em linha com os resultados empíricos obtidos por Coelho (2000), no âmbito de um estudo por simulação *design-based* com dados recolhidos através de um inquérito às estruturas agrícolas.

Ainda relativamente à estimação sintética no âmbito deste estudo empírico, admitiu-se que cada particular NUTSIII (domínio de interesse) apresenta um comportamento semelhante, em termos de preços médios da habitação, à respectiva NUTSII (no caso da estimação sintética pelo quociente) ou às restantes NUTSIII de Portugal continental (no caso da estimação sintética pela regressão). Como os resultados evidenciaram que os estimadores sintéticos são fortemente enviesados, então pode concluir-se que os pressupostos subjacentes a este tipo de estimação, designadamente através dos modelos implícitos que assistem a estimação, não se verificam no contexto da estimação do preço médio de transacção da habitação. Aliás, este resultado já era esperado, pois na análise exploratória da distribuição do preço médio de transacção da habitação na pseudo-população (subsecção 7.2.7.2), foi observado que esses preços apresentavam disparidades significativas entre as diferentes NUTSIII. Desta forma, parece reforçar-se a ideia da necessidade de utilização, no âmbito da aplicação prática, de estimadores assistidos por modelos de estimação em pequenos domínios que envolvam efeitos aleatórios específicos de domínio, de forma a acomodar a variabilidade existente entre as NUTSIII em termos de preços médios da habitação. Por último, é de salientar que a fragilidade de um estimador sintético pela regressão foi também evidenciada num estudo de Militino *et al.* (2007), no qual se comparou esse estimador com um estimador EBLUP, tendo como objectivo a estimação do Valor Acrescentado Bruto ao nível de comarcas e de províncias espanholas.

No subcapítulo 2.6 foram introduzidos os estimadores combinados, tendo sido definidos como estimadores que procuram equilibrar o enviesamento potencial do estimador sintético com a instabilidade do estimador directo, procurando desta forma evitar que a qualidade do estimador fique totalmente dependente da veracidade do modelo postulado (Rao, 2003). Os resultados alcançados para o caso particular do estimador combinado com pesos dependentes dos dados estão de acordo com a literatura, uma vez que se verificaram ganhos de enviesamento relativamente ao estimador sintético pela regressão e ganhos de variância relativamente ao estimador directo. É de notar que este estimador combinado, para além de apresentar ganhos em termos de enviesamento relativamente ao estimador sintético pela regressão, apresenta também ligeiros ganhos de precisão relativamente a esse estimador sintético. Os resultados obtidos para esse estimador combinado estão também em consonância total com os de outros estudos empíricos *design-based* envolvendo desenhos amostrais semelhantes. Destaca-se, em particular, os

estudos efectuados por Singh *et al.* (1994) e por Fabrizi *et al.* (2007). No primeiro estudo foram utilizados dados do *Statistics Canada's National Farm Survey*, enquanto no segundo foram utilizados dados do *European Community Household Panel Survey*. Pratesi e Salvati (2008) também obtiveram o mesmo tipo de resultados, embora no âmbito de um estudo do tipo *model-based*.

Os resultados obtidos neste estudo empírico por simulação também desvendaram que os estimadores EBLUP do grupo A (sem restrições) apresentam propriedades estatísticas de muito boa qualidade. De facto, verificou-se que este subgrupo de quatro estimadores é o que apresenta melhor desempenho em termos de precisão; apresenta ganhos expressivos de eficiência, sobretudo relativamente ao estimador directo; e apresenta, em termos de enviesamento, um desempenho globalmente muito melhor do que o dos estimadores sintéticos, mas apenas pouco pior do que o do estimador directo e do que o do estimador combinado com pesos dependentes dos dados. Estes resultados parecem anunciar que a combinação entre uma componente sintética e uma componente directa, reflectida nos estimadores EBLUP, consegue reduzir uma parte significativa do enviesamento do estimador sintético puro, trocando-a por um acréscimo da variância. É, aqui, de salientar que esta troca acaba por ser favorável do ponto de vista da precisão de todos os estimadores EBLUP, uma vez que o aumento na variância é mais do que compensado pela redução no enviesamento dos estimadores. Esta constatação, apesar de não ser de todo surpreendente, é, no entanto, melhor do que o esperado, uma vez que existem estudos na literatura onde os ganhos de precisão de estimadores EBLUP, relativamente a um estimador puramente sintético, só ocorrem quando se utiliza uma grande quantidade de "informação emprestada", normalmente através de dados temporais. Este tipo de resultados pode ser encontrado no estudo de Coelho (2000). No entanto, Militino *et al.* (2007) concluíram que um estimador EBLUP seccional é melhor do que um estimador sintético pela regressão, quando os pequenos domínios de interesse apresentam diferenças consideráveis entre si. Neste último estudo foram utilizados dados do *Business Survey of the Basque Country*.

Os resultados do estudo empírico parecem também confirmar que os estimadores EBLUP, assistidos por modelos de estimação em pequenos domínios que envolvem efeitos aleatórios específicos de domínio e/ou de domínio-tempo, são mais adequados do que os restantes estimadores quando existem discrepâncias significativas no

comportamento dos diferentes domínios de interesse, como é o caso nos preços médios de transacção da habitação em Portugal. Uma vez que os resultados obtidos para os quatro estimadores são globalmente uniformes, pode mesmo concluir-se que os estimadores EBLUP são mais adequados do que os restantes, independentemente do tipo de efeitos aleatórios incluídos no modelo e da natureza da informação neles introduzida (seccional, espacial, temporal). Mais uma vez, existe conformidade total entre os resultados obtidos neste estudo e os obtidos noutros estudos empíricos *design-based* envolvendo desenhos amostrais semelhantes. Veja-se, mais uma vez, os resultados obtidos nos estudos efectuados por Singh *et al.* (1994), por Coelho (2000) e por Fabrizi *et al.* (2007).

A constatação de que a utilização de uma maior quantidade de informação (temporal e/ou espacial) na estrutura de erro dos modelos de estimação em domínios que assistem os estimadores EBLUP tende a melhorar as propriedades do enviesamento, da precisão e da eficiência dos estimadores, foi outro resultado importante obtido no estudo empírico. Verificou-se, também, que a melhoria dessas propriedades é particularmente evidente no caso do estimador EBLUP espaciotemporal, o qual é o estimador mais eficiente no grupo dos estimadores avaliados e apresenta valores do enviesamento relativo absoluto médio pouco superiores aos do estimador directo, que é, como se sabe, teoricamente um estimador aproximadamente centrado. De facto, estes resultados estão em concordância com os resultados dos poucos estudos empíricos efectuados, que comparam simultaneamente um estimador EBLUP seccional com estimadores EBLUP espaciais, temporais ou espaciotemporais. Esses estudos são devidos a Coelho (2000) e a Singh *et al*. (2005). A este respeito, saliente-se agora o estudo empírico efectuado com dados reais por Singh *et al.* (2005) que, apesar de baseado em modelos de estimação em domínios do tipo *state space*, evidenciou também que o estimador que utiliza simultaneamente informação espacial e temporal apresenta ganhos significativos de eficiência relativamente aos restantes estimadores EBLUP, na estimação dos gastos mensais em consumo *per capita*, em pequenas regiões geográficas de um estado indiano. Em especial, aqueles autores também destacaram o facto do estimador EBLUP espaciotemporal apresentar vantagens em relação ao estimador EBLUP temporal, pelo facto de explorar a associação espacial existente entre os domínios de interesse.

Considerando agora o estimador EBLUP temporal, os resultados do estudo empírico mostraram que ele apresenta melhor precisão do que o estimador sintético e do que os estimadores EBLUP seccional e espacial, continuando a exibir reduções significativas nas medidas de enviesamento, sobretudo relativamente ao estimador sintético. Note-se, ainda, que os méritos do estimador EBLUP temporal foram mais evidentes nos domínios de menores dimensões amostrais. Este facto revela a importância de se utilizar "informação emprestada" de períodos passados na estimação em pequenos domínios, pois parece ser uma das melhores alternativas para aumentar a dimensão efectiva da amostra específica dos domínios, melhorando desta forma a qualidade dos estimadores. Alguns dos estudos pioneiros de referência, baseados em dados reais, que compararam estimadores EBLUP temporais com um estimador EBLUP seccional, devem-se a Choudhry e Rao (1989), Singh *et al.* (1994) e Datta *et al.* (1997, 2002). Todos estes estudos também confirmam a superioridade dos estimadores EBLUP temporais. Posteriormente, Singh *et al.* (2005) verificaram que um estimador EBLUP temporal é significativamente mais eficiente do que um estimador EBLUP seccional. Recentemente, Fabrizi *et al.* (2007) também concluíram que a utilização de modelos de estimação com informação auxiliar adequada e que "utilizem informação temporal emprestada", melhora significativamente a qualidade da estimação do rendimento médio das famílias italianas em pequenas regiões geográficas.

Relativamente ao estimador EBLUP espacial, os resultados obtidos mostraram que ele apresenta melhores propriedades do que os estimadores sintético e EBLUP seccional, sobretudo ao nível do enviesamento. Esta situação sugere a consideração de associação espacial entre os pequenos domínios, como uma importante via para obtenção de "informação emprestada", tendo em vista a redução do enviesamento. A generalidade dos estudos empíricos por simulação baseados em dados reais, que compararam estimadores EBLUP espaciais com um estimador EBLUP seccional, também evidencia ganhos em termos de eficiência quando se utiliza informação espacial. Veja-se, por exemplo, os estudos efectuados por Singh *et al.* (2005), Petrucci e Salvati (2006) e Pratesi e Salvati (2008). É, contudo, de realçar que, no estudo empírico de Coelho (2000), foi observado que a consideração de estruturas de covariância espacial nos efeitos aleatórios dos modelos de estimação em domínios, apesar de conduzir a reduções substanciais de enviesamento, leva a perdas moderadas de precisão (o que não ocorre neste estudo empírico). A este respeito, destaca-se ainda que no estudo empírico

de Chandra *et al.* (2007*b*), com dados do *Italian farm structure survey*, a superioridade do estimador EBLUP espacial é muito ténue ou mesmo nula.

Foi ainda possível extrair dos resultados deste estudo que, relativamente ao estimador EBLUP seccional de Fay-Herriot, a introdução de informação exclusivamente temporal melhora mais as propriedades do estimador do que a introdução de informação unicamente espacial. Por outras palavras, este facto revela que, na estimação do parâmetro de interesse numa NUTSIII particular, é melhor utilizar observações amostrais referentes a essa NUTSIII de períodos passados do que referentes a NUTSIII contíguas. Este resultado não é surpreendente, dada a natureza dos domínios de interesse definidos neste estudo. Na realidade, para além da própria localização geográfica dos domínios (NUTSIII), as próprias dinâmicas económicas divergem de domínio para domínio, pelo que é mais benéfico utilizar exclusivamente informação passada de um determinado domínio, do que informação de domínios vizinhos. Uma conclusão semelhante a esta foi obtida por Singh *et al*. (2005). Estes autores verificaram que o estimador EBLUP temporal é globalmente mais eficiente do que o estimador EBLUP espacial, quando comparados com um estimador EBLUP seccional.

Os resultados do estudo empírico indicaram que o recurso a IC baseados em estimativas do EQMP *model-based* resultou num aumento generalizado das taxas médias de cobertura dos IC, quando comparadas com aquelas que estavam associadas aos IC baseados em estimativas da variância *design-based*. Em particular, verificou-se, para a generalidade dos estimadores, que as taxas médias de cobertura dos IC *model-based* estão muito mais próximas do nível de confiança definido e estão muito menos relacionadas com a dimensão amostral dos domínios, do que as taxas médias de cobertura dos IC *design-based*. Estes factos parecem ser suportados pela boa qualidade dos estimadores do EQMP *model-based* dos EBLUP utilizados, os quais são estimadores não enviesados até à segunda ordem. Os resultados obtidos neste estudo empírico estão em linha com os obtidos no estudo realizado por Coelho (2000), tendo este autor sublinhado que, neste tipo de situações, os IC *model-based* podem constituir uma excelente alternativa aos IC *design-based*, mesmo sob uma perspectiva de amostragem repetida.

Ainda no que diz respeito às taxas médias de cobertura dos IC produzidos pelos estimadores EBLUP, os resultados indicaram que as diferenças entre esses dois tipos de

taxas médias de cobertura diminuem com o aumento da informação presente na estrutura de erro dos modelos que assistem a estimação. Estes resultados poderão ser devidos à redução do enviesamento nos estimadores EBLUP, que também ocorre à medida que é introduzida mais informação nos modelos que assistem a estimação. Note-se que aquela situação é particularmente evidente no caso do estimador EBLUP espaciotemporal que, como se verificou, é um estimador aproximadamente centrado.

É, também, de salientar que a metodologia seguida no diagnóstico aos modelos de estimação em pequenos domínios sugeriu, de entre os modelos utilizados, como o modelo que se ajusta melhor aos dados, aquele que assiste o estimador EBLUP que apresenta melhores propriedades sob amostragem repetida. Este facto reforça a importância e a adequabilidade da metodologia seguida no diagnóstico aos modelos de estimação em pequenos domínios. Em particular, essa metodologia de diagnóstico permite seleccionar, não só os modelos que descrevem melhor a população finita, mas sobretudo, permite indicar os modelos que assistem os estimadores dos parâmetros de interesse em pequenos domínios com melhor qualidade.

Os resultados empíricos ilustraram que a garantia da consistência interna na publicação das estimativas conduz ao aumento do enviesamento e à perda de precisão dos estimadores, comparativamente a um estimador que não verifica essa consistência. Pode interpretar-se esta perda de qualidade dos estimadores como o preço a pagar pela garantia da consistência interna, normalmente tão desejada pelos produtores de estatísticas oficiais aquando da publicação de estimativas do mesmo parâmetro de interesse para diferentes níveis de desagregação. A este respeito, os resultados também mostraram que o estimador que assegura a consistência interna com melhores propriedades é o estimador EBLUP espaciotemporal com restrições ao nível de NUTSII. Tendo em conta a deterioração das suas propriedades, relativamente ao estimador EBLUP espaciotemporal sem restrições, pode alvitrar-se que a garantia da consistência interna não confere robustez ao estimador ajustado, devido ao facto do modelo que assiste a estimação se encontrar bem especificado, descrevendo adequadamente a população finita.

Os resultados indicaram também que a garantia da consistência interna ao nível de Portugal continental tem um efeito mais prejudicial sobre as propriedades do estimador EBLUP que garante essa consistência, do que sobre as propriedades do estimador

EBLUP que garante a consistência apenas ao nível de NUTSII. Este resultado era esperado, uma vez que será mais difícil garantir que a média ponderada das estimativas EBLUP ao nível de NUTSIII, que apresentam naturalmente diferenças entre si, igualem uma única estimativa para todo o país, do que igualem estimativas ao nível de NUTSII. Acredita-se que os preços médios de transacção da habitação apresentam variabilidades intra-NUTSII menores do que a variabilidade total, devido a dinâmicas económicas regionais diferentes que afectam aqueles preços. Como tal, o enviesamento e a variância associados ao estimador que garante a consistência ao nível de Portugal continental serão maiores.

Por último, foi ainda possível observar pelo estudo empírico que a garantia da consistência interna afecta muito pouco a variância dos estimadores, desde que a estimativa directa do parâmetro de interesse para o nível de agregação superior tenha uma boa precisão. No caso em que essa estimativa directa apresenta uma precisão inaceitável, então essa variância aumenta significativamente. Relativamente a este assunto, Coelho (2000) refere que existem riscos de se assegurar a consistência interna das estimativas a um nível de agregação para o qual o estimador que é usado como padrão apresenta fraca precisão, o que se veio a confirmar neste estudo empírico.

Perante estes resultados, recomenda-se que a consistência interna na publicação das estimativas do preço médio de transacção da habitação seja garantida para níveis pouco agregados (neste caso ao nível de NUTSII).

7.3 AVALIAÇÃO DO DESEMPENHO DOS ESTIMADORES DO EQMP PROPOSTOS

7.3.1 Introdução

Neste subcapítulo é apresentado um estudo por simulação *model-based*, implementado para avaliar o desempenho dos estimadores do EQMP dos EBLUP sem restrições. Este estudo encontra-se dividido em duas partes. Na primeira parte, é apresentado um estudo para avaliar o desempenho dos estimadores por reamostragem (*jackknife* e *bootstrap*) do

EQMP do EBLUP temporal, face ao estimador analítico desse EQMP, no contexto do modelo de Rao-Yu (secção 7.3.2). Na segunda parte, é apresentado um estudo de avaliação do desempenho dos estimadores propostos (analítico, *jackknife* e *bootstrap*) para medição da incerteza associada ao EBLUP espaciotemporal (secção 7.3.3). Em ambos os casos, é apresentado em primeiro lugar o desenho do respectivo estudo, seguido da apresentação e discussão dos principais resultados obtidos.

7.3.2 Avaliação do desempenho dos estimadores do EQMP do EBLUP temporal

7.3.2.1 Desenho do estudo por simulação *model-based*

Começa por apresentar-se a forma como os dados foram gerados. Por conveniência de ordem prática, neste estudo empírico foram considerados m=28 domínios e admitiu-se a existência de dados temporais referentes a T=7 períodos de tempo. O vector de dados iniciais do parâmetro de interesse, $\mathbf{y}$, foi gerado de acordo com a especificação do modelo de estimação em pequenos domínios de Rao-Yu. Para tal, foi considerado um modelo com p=2, ou seja, com uma variável auxiliar, X_{it}, e o termo independente, $\mathbf{x}_{it} = (1, x_{it})'$. Os mT valores de X_{it} foram gerados a partir de uma distribuição uniforme no intervalo [0,1]. Admitiu-se que os verdadeiros valores dos efeitos fixos são $\boldsymbol{\beta} = (1; 2)'$, as verdadeiras variâncias dos efeitos aleatórios específicos de domínio são $\sigma_v^2 \in \{0{,}5; 1{,}0\}$, as verdadeiras variâncias dos efeitos aleatórios específicos de domínio-tempo são $\sigma^2 \in \{0{,}00; 0{,}25; 0{,}50; 1{,}00\}$, e os valores dos parâmetros de autocorrelação são $\rho \in \{0{,}0; 0{,}2; 0{,}4; 0{,}8\}$. Tal como em Rao e Yu (1994), assumiu-se que $\varepsilon_{it} \overset{iid}{\sim} N(0,1)$ e que ρ é conhecido, apesar dos métodos de reamostragem propostos também poderem contemplar a situação de ρ desconhecido. Todos os erros do modelo foram gerados de forma mutuamente independente a partir de distribuições normais de média nula. Por último, o vector dos mT valores do parâmetro de interesse foi gerado através do modelo (4.3.4), o qual se admite ser o verdadeiro modelo.

Note-se que se está perante um estudo que envolve 32 combinações de parâmetros de variância, obtendo-se o modelo de Fay-Herriot (4.2.3) como caso particular quando se

considera $\sigma^2 = 0$ e $\rho = 0$, para $t = 1, ..., 7$. No contexto deste modelo particular, e para o caso de estimação da componente de variância pelo estimador dos momentos de Prasad-Rao (4.2.13), utilizou-se o estimador analítico do EQMP do EBLUP de Prasad-Rao (4.2.27) (EQMP-PR), o estimador *bootstrap* paramétrico de Butar-Lahiri (4.2.35) e o estimador *jackknife* ponderado de Chen-Lahiri (4.2.32).

O estudo por simulação de Monte Carlo, desenhado para comparar o desempenho dos estimadores do EQMP do EBLUP temporal, foi efectuado de acordo com o seguinte algoritmo:

1. Gerar L=1.000 conjuntos de dados iniciais, $\mathbf{y}^{(l)} = \left(y_{11}^{(l)}, ..., y_{it}^{(l)}, ..., y_{mT}^{(l)}\right)$, tal como descrito acima, l=1, …, L.

2. Calcular as estimativas das componentes de variância, $\hat{\sigma}_v^{2(l)}$ e $\hat{\sigma}^{2(l)}$, pelo método dos momentos com base nos dados iniciais, $\mathbf{y}^{(l)}$, e ajustar o modelo (4.3.4) de forma a determinar as estimativas dos efeitos fixos $\hat{\boldsymbol{\beta}}^{(l)} = \hat{\boldsymbol{\beta}}\left(\mathbf{y}^{(l)}; \hat{\boldsymbol{\psi}}^{(l)}\right)$, onde $\hat{\boldsymbol{\psi}}^{(l)} = \left(\hat{\sigma}_v^{2(l)}, \hat{\sigma}^{2(l)}, \rho\right)'$, para cada l=1, …, L.

3. Calcular as estimativas do parâmetro de interesse, $\hat{\theta}_{it}^{(l)}\left(\hat{\boldsymbol{\psi}}^{(l)}\right)$, e as suas estimativas analíticas do EQMP (EQMP-RY), $eqmp^{RY(l)}(\hat{\theta}_{it}^{(l)})$, para cada l=1, …, L. Para além disso, calcular as estimativas *jackknife* do EQMP do EBLUP utilizando dois tipos de pesos: $eqmp^{J1(l)}\left(\hat{\theta}_{it}^{(l)}\right)$ com $w_{1et} = (m-1)/m$ (EQMP-J1) e $eqmp^{J2(l)}\left(\hat{\theta}_{it}^{(l)}\right)$ com $w_{2et} = 1 - \mathbf{x}_{et}'\left(\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}\mathbf{x}_{et}$, (EQMP-J2), para cada l=1, …, L.

4. Gerar B=250 conjuntos de dados *bootstrap* com base nas estimativas $\hat{\sigma}_v^{2(l)}$ e $\hat{\sigma}^{2(l)}$, tal como descrito na secção 4.3.6, e depois calcular as estimativas *bootstrap* do EQMP do EBLUP (EQMP-B1): $eqmp^{B1(l)}(\hat{\theta}_{it}^{(l)})$, para cada l=1, …, L.

5. Calcular as aproximações aos verdadeiros valores do EQMP do EBLUP para cada i-ésimo domínio em cada t-ésimo período de tempo, $EQMP_{it}$, os quais servem como termo de comparação. Esses valores foram calculados através de simulação de Monte

Carlo, com base em R=5.000 conjuntos de dados independentes, de forma a assegurar uma melhor precisão nos resultados.

A avaliação da qualidade dos estimadores do EQMP do EBLUP foi efectuada com base no enviesamento relativo (BR) e no EQM relativo (EQMR)[106] dos estimadores em análise. Para um estimador $eqmp_{it}^{f}$ do parâmetro $EQMP_{it}$, estas medidas são definidas, respectivamente, como:

$$BR_{it} = L^{-1}\sum_{l=1}^{L}\frac{eqmp_{it}^{f(l)} - EQMP_{it}}{EQMP_{it}}\times 100, \tag{7.3.1}$$

$$EQMR_{it} = L^{-1}\sum_{l=1}^{L}\frac{\left(eqmp_{it}^{f(l)} - EQMP_{it}\right)^2}{EQMP_{it}}\times 100, \tag{7.3.2}$$

onde $f \in \{PR, RY, B1, J1, J2\}$ denota os diferentes estimadores do EQMP e l o l-ésimo conjunto de dados, l=1, …, 1.000. Estas medidas são calculadas ao nível individual (domínio-tempo). Com o objectivo de sumariar os resultados, são utilizadas três medidas globais sobre os mT domínios de interesse: a percentagem de domínios onde o enviesamento relativo é negativo (BRN), o enviesamento relativo absoluto médio (BRAM), $BRAM = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\left|BR_{it}\right|$, e o EQM relativo médio (EQMRM), $EQMRM = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}EQMR_{it}$. Para além disso, foram também utilizados diagramas em caixas de bigodes para representar as distribuições do enviesamento relativo dos estimadores do EQMP do EBLUP temporal.

A realização desta parte do estudo empírico *model-based* exigiu a programação em linguagem SAS de programas com as seguintes funções: (*i*) geração de conjuntos de dados com as características apresentadas acima e cálculo das estimativas dos parâmetros de interesse através do estimador EBLUP; (*ii*) cálculo das estimativas analíticas do EQMP; (*iii*) cálculo das estimativas *jackknife* do EQMP; (*iv*) cálculo das estimativas *bootstrap* do EQMP; e (*v*) cálculo das medidas de avaliação da qualidade

[106] Em rigor, a medida utilizada é o erro quadrático relativo médio. Contudo, por facilidade de linguagem decidiu utilizar-se a designação EQM relativo. No primeiro caso, a medida global sobre os mT domínios seria designada por erro quadrático relativo médio médio, o que poderia gerar confusão ao leitor.

dos estimadores. Todos estes programas estão reunidos no apêndice 23. O tempo total de processamento computacional deste estudo por simulação foi de cerca de 90 dias.

Tal como observado por Bell (2001) no contexto do estimador *jackknife* do EQMP do EBLUP, e posteriormente por Jiang e Lahiri (2006), o estimador do termo de correcção de enviesamento dos estimadores *jackknife* e *bootstrap* paramétricos do EQMP do EBLUP pode produzir estimativas negativas, podendo conduzir a estimativas negativas do EQMP do EBLUP. Neste estudo por simulação não ocorreram estimativas negativas do EQMP, embora tivessem sido obtidas algumas estimativas negativas do termo de correcção de enviesamento. Em média verificaram-se 0,46%, 0,57% e 0,58% de estimativas negativas do termo de correcção de enviesamento nos estimadores EQMP-B1, EQMP-J1 e EQMP-J2, respectivamente.

7.3.2.2 Análise dos resultados do estudo

As tabelas 7.3.1 a 7.3.4 mostram as medidas percentagem de domínios onde o enviesamento relativo é negativo, enviesamento relativo absoluto médio, EQM relativo médio, associadas a cada um dos quatro estimadores do EQMP, para ρ=0,0, ρ=0,2, ρ=0,4 e ρ=0,8, respectivamente. Todas as medidas são apresentadas em termos percentuais. Da mesma forma, as figuras 7.3.1 a 7.3.4, contêm os diagramas em caixas de bigodes do enviesamento relativo desses estimadores. Como se pode observar *a priori*, o desempenho dos diferentes estimadores do EQMP depende significativamente da combinação dos parâmetros de autocorrelação e de variância.

A tabela 7.3.1 apresenta o desempenho dos quatro estimadores do EQMP do EBLUP para o caso particular do modelo de Fay-Herriot. Através da comparação desses estimadores, pode observar-se que os estimadores *jackknife* têm um desempenho melhor do que o estimador analítico, quer em termos de enviesamento quer em termos de precisão. Pelo contrário, o desempenho do estimador *bootstrap* é pior do que o do estimador analítico, apesar de tender a subestimar o verdadeiro EQMP na mesma percentagem de domínios.

A partir da figura 7.3.1 pode observar-se que os estimadores *jackknife* têm uma propensão para apresentar os intervalos de variação de enviesamento relativo com as

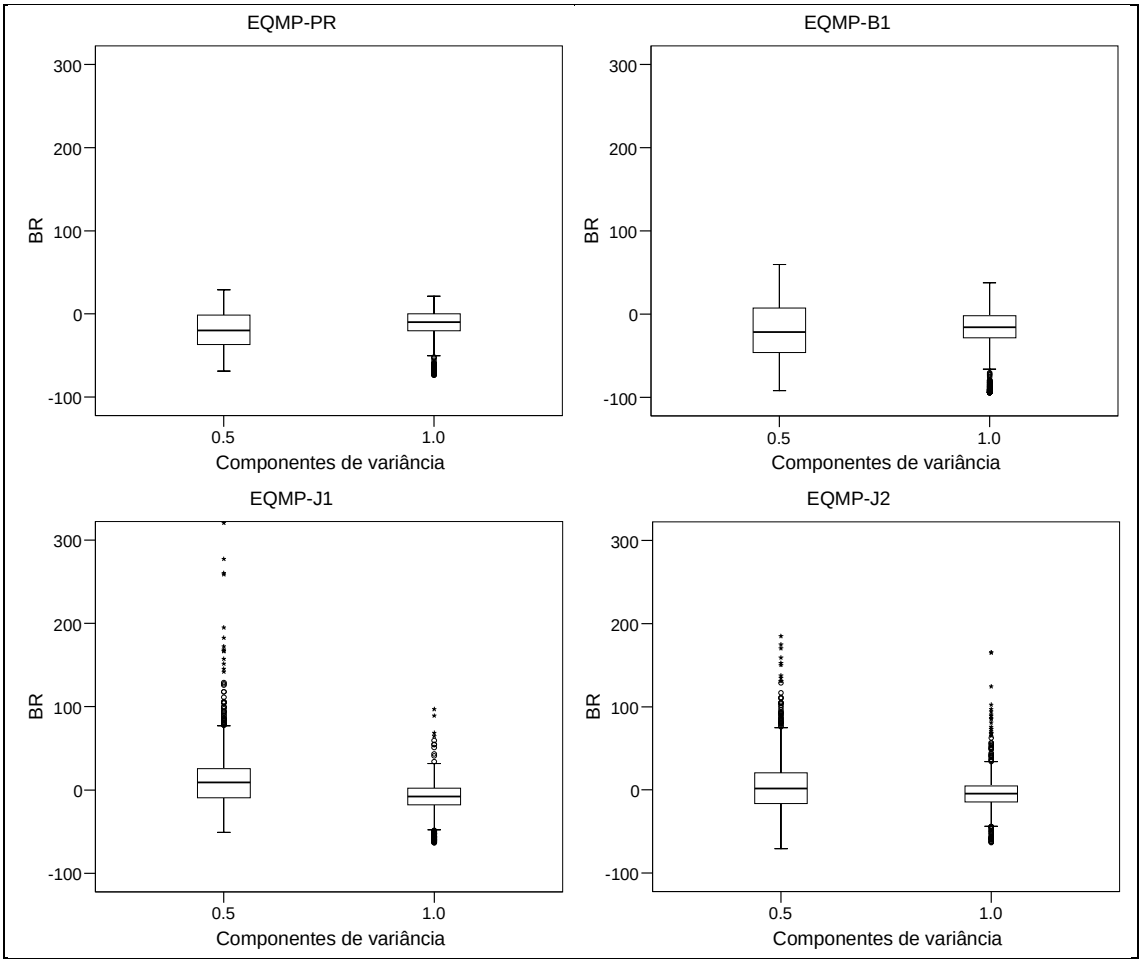
maiores amplitudes, apesar do enviesamento relativo médio ser aproximadamente igual a zero.

Tabela 7.3.1: Medidas de qualidade dos estimadores do EQMP do EBLUP seccional, para ρ=0,0

σ_v^2	**EQMP-PR**	**EQPM-B1**	**EQMP-J1**	**EQMP-J2**
		BRN (%)		
0,5	77,679	68,821	38,000	47,500
1,0	74,464	77,821	69,143	61,750
		BRAM (%)		
0,5	25,042	33,766	22,617	22,575
1,0	16,127	22,294	13,781	12,851
		EQMRM (%)		
0,5	4,464	7,898	4,313	3,888
1,0	3,089	5,663	2,093	2,012

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

Figura 7.3.1: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP seccional, para ρ=0,0



A tabela 7.3.2 ilustra que o estimador analítico é o melhor quando se considera ρ=0,2. A partir desta tabela pode verificar-se que as estimativas produzidas pelo estimador analítico apresentam menores valores percentuais nas medidas erro relativo absoluto médio e EQM relativo médio (embora da mesma ordem de magnitude) do que as estimativas produzidas através dos estimadores por reamostragem. A tabela 7.3.2 também evidencia que o estimador *bootstrap* apresenta um desempenho ligeiramente melhor do que ambos os estimadores *jackknife*, em termos de enviesamento (BRAM) e de precisão (EQMRM), para valores mais pequenos de σ^2. Para além disso, e para essas medidas, verifica-se que as diferenças entre o estimador *bootstrap* e o estimador analítico são muito pequenas. Na tabela 7.3.2 pode ainda descortinar-se um comportamento globalmente semelhante para todos os estimadores, no que se refere às três medidas utilizadas, designadamente, (*i*) a percentagem de domínios com estimativas do EQMP que apresentam um enviesamento positivo tende a aumentar para valores elevados das componentes de variância; e (*ii*) o enviesamento tende a diminuir e a precisão a melhorar para valores mais elevados de σ^2.

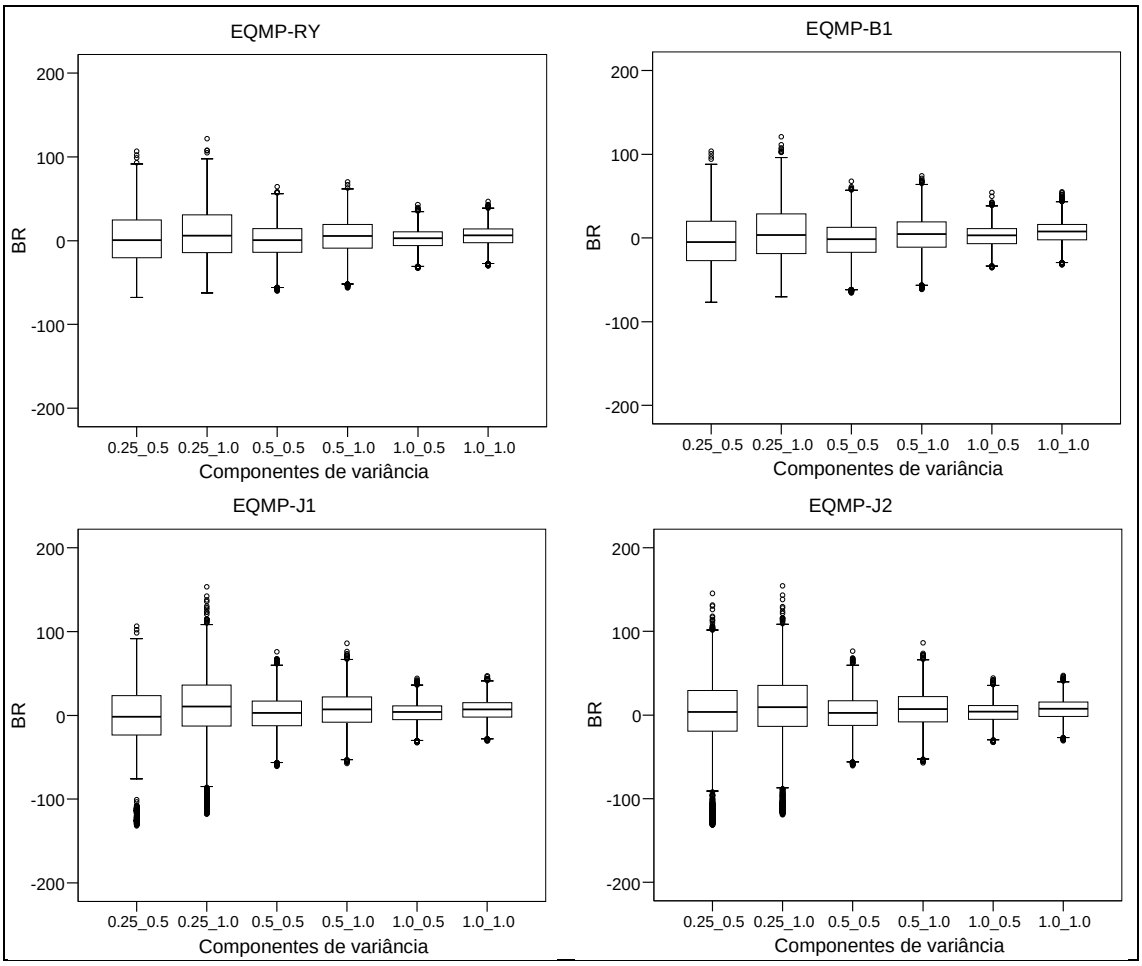
Tabela 7.3.2: Medidas de qualidade dos estimadores do EQMP do EBLUP temporal, para ρ=0,2

σ^2	σ_v^2	**EQMP-RY**	**EQMP-B1**	**EQMP-J1**	**EQMP-J2**
			BRN (%)		
0,25	0,5	49,408	55,224	51,449	46,235
	1,0	44,184	46,898	39,408	39,663
0,50	0,5	48,490	52,878	44,867	44,673
	1,0	39,837	41,867	37,347	36,939
1,00	0,5	41,133	41,347	38,694	38,204
	1,0	31,500	30,459	30,010	29,286
			BRAM (%)		
0,25	0,5	26,388	27,901	30,168	31,100
	1,0	26,700	27,946	32,105	31,985
0,50	0,5	16,401	17,404	17,374	17,278
	1,0	17,176	18,027	18,333	18,286
1,00	0,5	9,770	10,690	10,156	10,122
	1,0	10,854	12,169	11,399	11,431
			EQMRM (%)		
0,25	0,5	2,886	3,245	4,466	4,678
	1,0	2,999	3,256	4,786	4,774
0,50	0,5	1,677	1,922	1,864	1,838
	1,0	1,762	1,955	2,006	1,993
1,00	0,5	0,752	0,908	0,811	0,805
	1,0	0,903	1,146	0,995	1,001

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

A partir da figura 7.3.2 é possível observar que todos os estimadores do EQMP apresentam enviesamentos relativos médios aproximadamente iguais a zero e que a amplitude dos intervalos de variação dos enviesamentos relativos diminui para valores elevados das componentes de variância. Contudo, este decréscimo na amplitude dos intervalos de variação é mais pronunciado nos estimadores do tipo *jackknife*, do que no estimador *bootstrap*.

Figura 7.3.2: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP temporal, para ρ=0,2



A partir da análise da tabela 7.3.3 é possível observar que, quando se considera ρ=0,4, todos os estimadores subestimam o verdadeiro EQMP do EBLUP em mais de metade dos pequenos domínios e que esta subestimação tende a diminuir com o aumento da variância associada ao processo AR(1), σ^2. Neste contexto, ou seja, considerando ρ=0,4, os estimadores por reamostragem do tipo *jackknife* podem competir em termos de enviesamento com o estimador analítico, o qual volta a ser claramente o melhor estimador. Para além disso, os dois estimadores *jackknife* apresentam um

comportamento muito semelhante entre si, sendo ligeiramente menos enviesados e mais precisos do que o estimador *bootstrap*. Saliente-se, ainda, que o estimador analítico é também o mais preciso para ρ=0,4, uma vez que é o que apresenta os menores valores no EQM relativo médio, sendo tanto mais preciso quanto maiores valores assumirem as componentes de variância. Por último, a partir da tabela 7.3.3 pode ainda descortinar-se um comportamento globalmente semelhante para todos os estimadores. Neste caso verifica-se, para todos os estimadores, que o enviesamento diminui e a precisão melhora para valores mais elevados de ambas as componentes de variância.

Tabela 7.3.3: Medidas de qualidade dos estimadores do EQMP do EBLUP temporal, para ρ=0,4

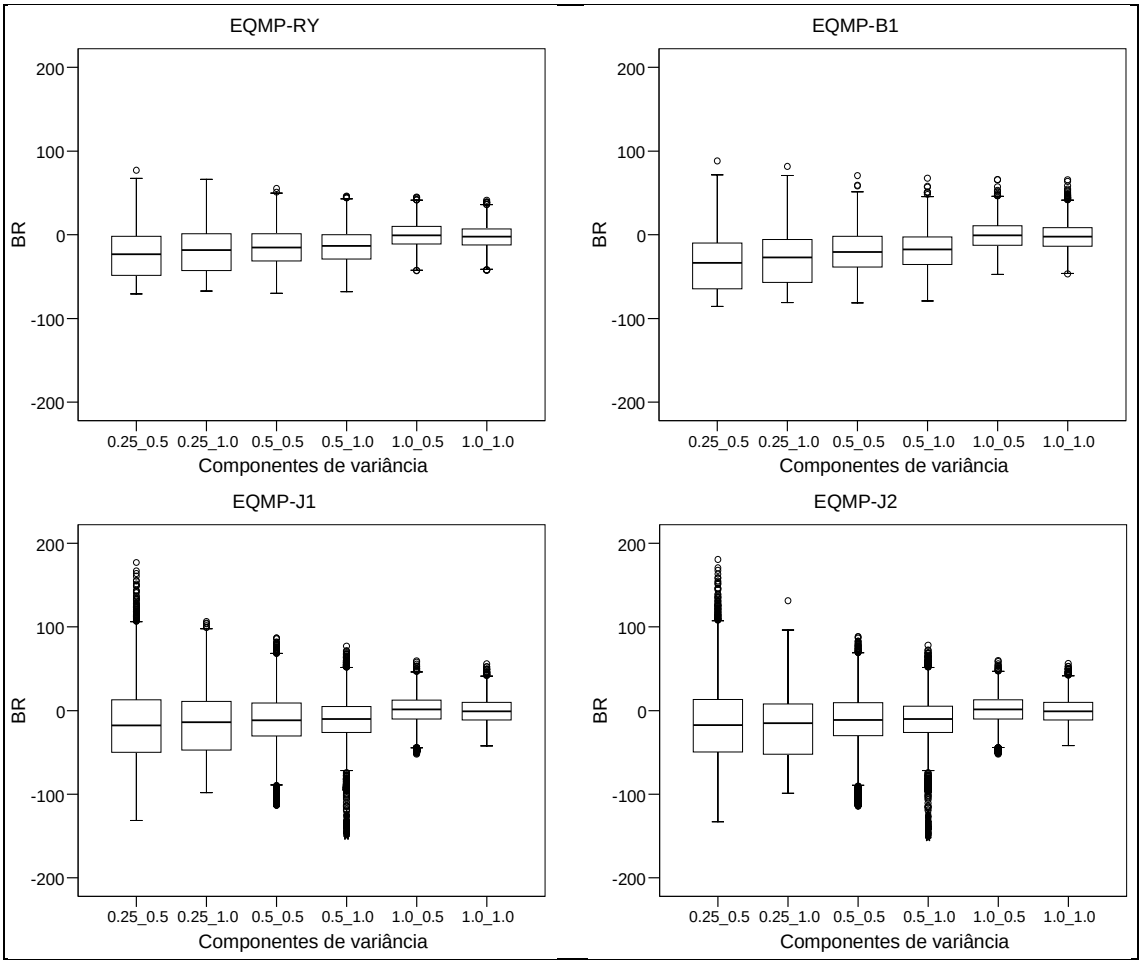
σ^2	σ_v^2	**EQMP-RY**	**EQMP-B1**	**EQMP-J1**	**EQMP-J2**
			BRN (%)		
0,25	0,5	76,541	80,714	67,612	67,367
	1,0	73,898	78,878	65,296	67,449
0,50	0,5	73,337	77,092	64,337	63,908
	1,0	74,857	77,878	67,949	67,612
1,00	0,5	51,418	51,745	47,235	47,020
	1,0	56,816	55,939	52,735	52,449
			BRAM (%)		
0,25	0,5	31,607	40,665	38,391	38,503
	1,0	28,119	35,889	33,028	33,086
0,50	0,5	22,858	27,326	25,819	25,864
	1,0	20,858	24,920	23,231	23,248
1,00	0,5	12,236	13,828	13,477	13,508
	1,0	11,549	13,163	12,758	12,785
			EQMRM (%)		
0,25	0,5	4,587	7,538	7,559	7,632
	1,0	3,872	6,286	5,812	6,005
0,50	0,5	3,394	4,840	4,854	4,881
	1,0	3,002	4,253	4,809	4,849
1,00	0,5	1,176	1,510	1,439	1,446
	1,0	1,108	1,442	1,351	1,357

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

A figura 7.3.3 ilustra claramente que todos os estimadores subestimam sistematicamente o verdadeiro EQMP, principalmente para valores mais baixos de σ^2. Contudo, o enviesamento relativo médio tende para zero para valores elevados das componentes de variância. Mais uma vez se verifica que a amplitude dos intervalos de variação dos enviesamentos relativos diminui para valores elevados das componentes de variância. Na figura 7.3.3 pode ainda observar-se que o estimador *jackknife* apresenta intervalos de variação do enviesamento relativo com as maiores amplitudes, bem como

um grande número de domínios com valores de enviesamento relativo considerados *outliers* quando $\sigma^2 = 0{,}50$, sobretudo na aba esquerda das distribuições. Esta situação indica, nesses casos, uma subestimação mais acentuada do verdadeiro valor do EQMP do EBLUP, do que os restantes estimadores.

Figura 7.3.3: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP temporal, para ρ=0,4



A partir da leitura da tabela 7.3.4, pode concluir-se que os estimadores *jackknife* subestimam o verdadeiro EQMP do EBLUP na maioria dos domínios e esta subestimação aumenta para valores mais elevados de σ^2, ao contrário do que se verifica para ρ=0,2 e ρ=0,4. Também o estimador *bootstrap* manifesta forte subestimação do verdadeiro EQMP do EBLUP, embora neste caso esta subestimação diminua para valores mais elevados de ambas as componentes de variância. Pelo contrário, o estimador analítico sofre do problema de sobreestimação para valores baixos de σ^2. Estes padrões também podem ser observados na figura 7.3.4.

Tabela 7.3.4: Medidas de qualidade dos estimadores do EQMP do EBLUP temporal, para ρ=0,8

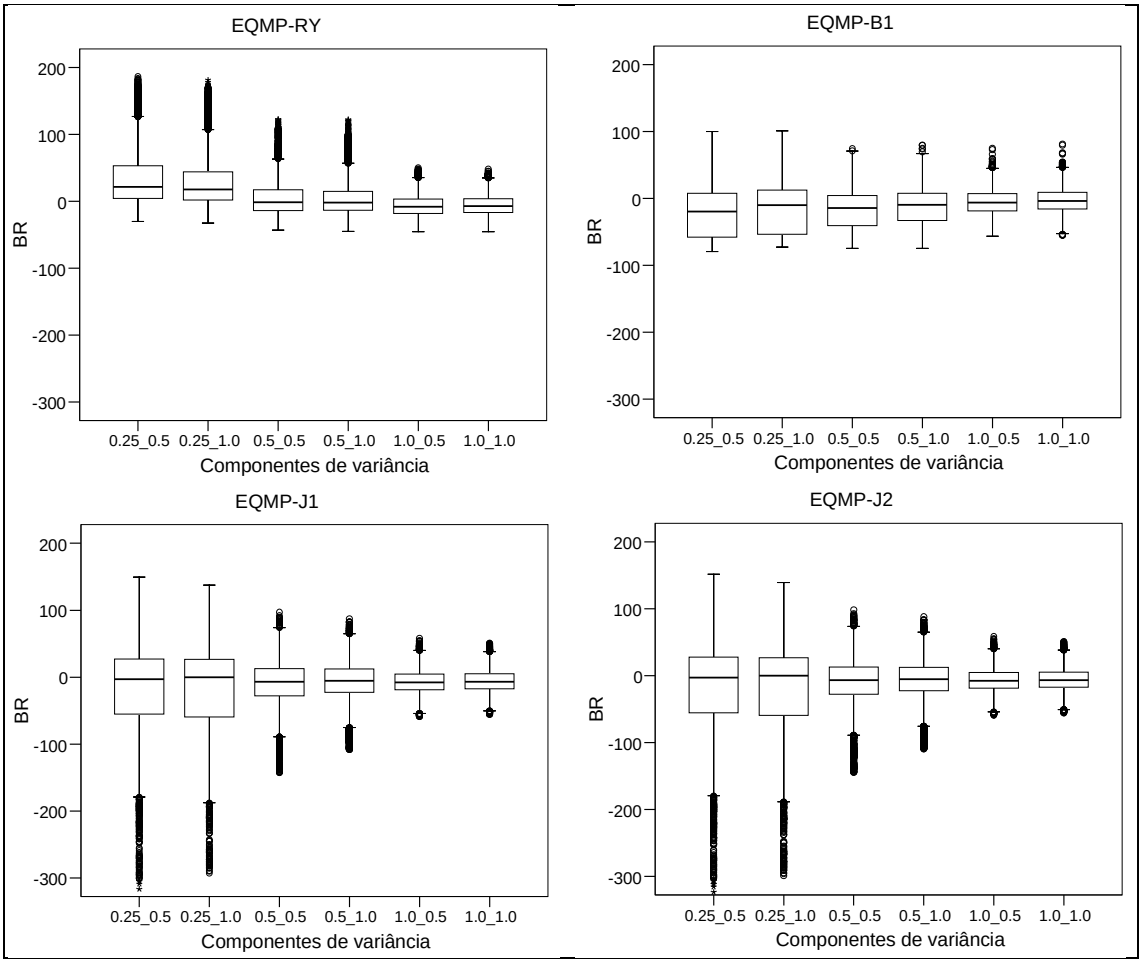
σ^2	σ_v^2	EQMP-RY	EQMP-B1	EQMP-J1	EQMP-J2
			BRN (%)		
0,25	0,5	17,296	68,980	52,327	52,010
	1,0	20,847	62,316	50,133	49,816
0,50	0,5	51,949	70,245	59,980	59,582
	1,0	53,459	65,571	59,031	58,602
1,00	0,5	68,265	63,184	66,837	66,500
	1,0	67,265	58,194	65,847	65,429
			BRAM (%)		
0,25	0,5	37,056	35,711	45,188	45,577
	1,0	32,635	32,311	46,097	46,542
0,50	0,5	18,577	27,145	27,673	27,753
	1,0	17,256	23,960	23,709	23,751
1,00	0,5	14,381	16,371	15,616	15,581
	1,0	13,531	15,159	14,537	14,507
			EQMRM (%)		
0,25	0,5	10,982	7,085	14,806	15,143
	1,0	9,369	6,403	16,707	17,110
0,50	0,5	2,891	5,654	6,800	6,859
	1,0	2,601	4,674	5,200	5,231
1,00	0,5	1,852	2,536	2,282	2,272
	1,0	1,671	2,215	2,008	1,999

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

A tabela 7.3.4 evidencia também que os estimadores por reamostragem do EQMP são globalmente mais enviesados do que o estimador analítico, para diferentes combinações das componentes de variância. Porém, o estimador *bootstrap* pode competir com o estimador analítico em termos de enviesamento. Na figura 7.3.4 é possível observar que o estimador *bootstrap* apresenta enviesamentos relativos médios negativos para todas as combinações das componentes de variância. Para além disso, estes enviesamentos relativos médios distam mais de zero do que os correspondentes enviesamentos relativos médios dos estimadores *jackknife*. De facto, estes últimos dois estimadores apresentam enviesamentos relativos que, em média, estão consistentemente próximos de zero para qualquer combinação das componentes de variância. Esta parece constituir uma vantagem desses estimadores relativamente ao estimador analítico, o qual evidencia uma tendência para apresentar enviesamentos relativos médios positivos para valores baixos de σ^2. É, ainda, de salientar que os estimadores *jackknife* tendem a apresentar amplitudes inter-quartil dos enviesamentos relativos menores ou iguais do que as correspondentes amplitudes do estimador *bootstrap*, mas apresentam as maiores amplitudes nos respectivos intervalos de variação. Para além disso, os estimadores

jackknife apresentam um elevado número de domínios com valores de enviesamento relativo considerados *outliers* nas abas esquerdas das distribuições, quando $\sigma^2 = 0{,}25$ e $\sigma^2 = 0{,}50$, ao contrário do que se verifica para os outros dois estimadores. Por todas estas razões, pode-se concluir que, em termos de enviesamento, os estimadores por reamostragem podem competir com o estimador analítico, sendo preferível o estimador *bootstrap* para valores baixos de σ^2 e um dos estimadores *jackknife* para valores elevados de σ^2. Por último, é ainda de salientar que o enviesamento relativo absoluto médio de todos os estimadores do EQMP diminui com o aumento dos valores de ambas as componentes de variância. No que se refere à precisão, os resultados obtidos a partir de estimadores baseados em métodos por reamostragem são globalmente semelhantes, embora com clara vantagem para o estimador *bootstrap* para $\sigma^2 = 0{,}25$. Em todo o caso, o estimador analítico é o mais preciso para valores elevados de σ^2, ou seja, para $\sigma^2 = 0{,}50$ e $\sigma^2 = 1{,}00$.

Figura 7.3.4: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP temporal, para ρ=0,8



7.3.2.3 Síntese e discussão dos resultados

Os resultados obtidos neste estudo por simulação *model-based* indicaram, para o caso particular do modelo de Fay-Herriot ($\sigma^2 = 0$ e $\rho = 0$), que o estimador *jackknife* tem um desempenho melhor do que o estimador analítico, enquanto o estimador *bootstrap* tem um comportamento pior do que esse estimador analítico. Estes resultados são concordantes com os resultados obtidos por Jiang *et al*. (2002) e por Butar e Lahiri (2003).

Após uma comparação global dos estimadores obtidos por métodos de reamostragem, para diferentes níveis de autocorrelação, pode concluir-se que os dois estimadores do tipo *jackknife* apresentam os mesmos níveis de enviesamento e de precisão para todos os casos, uma vez que as medidas BRAM e EQMRM apresentam sempre valores muito próximos. Por simplicidade de implementação, sugere-se a utilização dos pesos $w_{1et} = (m-1)/m$ nas aplicações práticas. Pode também concluir-se que o estimador *bootstrap* apresenta globalmente os mesmos (ou ligeiramente piores) níveis de enviesamento, do que os estimadores *jackknife*, tal como Fabrizi *et al.* (2007) concluíram no contexto de estimação em pequenos domínios com modelos temporais de nível unidade. Contudo, os estimadores do tipo *jackknife* revelam a desvantagem de poderem apresentar enviesamentos mais acentuados do que o estimador *bootstrap*, para valores mais elevados do coeficiente de autocorrelação. Em termos de precisão, verifica-se que o estimador *bootstrap* apresenta níveis de precisão muito semelhantes aos dos estimadores *jackknife,* ou eventualmente melhores para valores baixos de σ^2. Estes resultados são extremamente importantes, uma vez que o estimador *bootstrap* é o estimador por reamostragem mais fácil de implementar porque não exige a dedução de uma expressão fechada, mesmo no contexto de modelos de estimação em domínios mais complexos[107]. Pelo exposto acima, sugere-se que o estimador *bootstrap* seja o estimador por reamostragem preferido nas aplicações práticas.

[107] É, contudo, de relembrar que a metodologia *jackknife* geral adaptada por Jiang *et al*. (2002) para a medição da incerteza dos estimadores EBLUP no contexto da estimação em domínios, não exige a dedução de expressões fechadas para os estimadores do EQMP. No entanto, esta metodologia tem a

Numa análise transversal a todos os estimadores, para diferentes níveis de autocorrelação, é de notar que o enviesamento diminui e a precisão melhora para valores mais elevados de ambas as componentes de variância. A única excepção ocorre para ρ=0,2, em que o aumento de σ_v^2 faz piorar, se bem que ligeiramente, o enviesamento e a precisão de todos os estimadores. Verifica-se também que não existe um único estimador que seja uniformemente melhor do que todos os outros em termos de enviesamento e de precisão, apesar de alguma instabilidade nos resultados. Fabrizi *et al.* (2007) obtiveram conclusões semelhantes, após a comparação de diferentes estimadores do EQMP no contexto de modelos de nível unidade.

Em jeito de conclusão pode dizer-se que os resultados alcançados neste estudo empírico mostram que os estimadores *bootstrap* e *jackknife* apresentam um desempenho muito bom, quando comparado com o estimador analítico, na medição da incerteza associada às estimativas EBLUP temporal, mesmo no contexto de um pequeno número de períodos de tempo e um moderado número de pequenos domínios. Desta forma, parece ser perfeitamente adequado o uso de estimadores baseados em métodos de reamostragem e, em particular, o estimador *bootstrap*, na estimação da incerteza associada ao estimador EBLUP temporal, como alternativa aos estimadores baseados em longos desenvolvimentos analíticos. A utilização deste tipo de estimadores baseados em métodos de reamostragem é promissora no contexto de modelos longitudinais de estimação em pequenos domínios mais complexos, para os quais é impossível deduzir aproximações analíticas para o EQMP do EBLUP. A complexidade do modelo pode ser resultante, por exemplo, da consideração de estruturas de covariância dos efeitos aleatórios que envolvam a estimação de várias componentes de variância (como, por exemplo, as estruturas auto-regressiva de primeira ordem heterocedástica – ARH(1) e composta simétrica heterocedástica – CSH) ou da introdução de restrições na estimação.

---

desvantagem de ser mais exigente em termos de cálculo computacional e de algumas estimativas do EQMP do EBLUP poderem assumir valores negativos.

7.3.3 Avaliação do desempenho dos estimadores do EQMP do EBLUP espaciotemporal

7.3.3.1 Desenho do estudo por simulação *model-based*

Uma vez que os objectivos deste estudo por simulação são idênticos aos do estudo apresentado na secção 7.3.2, embora no contexto de um estimador EBLUP assistido agora por um modelo espaciotemporal, então o desenho do estudo também é idêntico. Por conveniência de ordem prática, foram considerados m=28 domínios e admitiu-se a existência de dados temporais referentes a T=7 períodos de tempo. No que se refere à geração de dados, o vector de dados iniciais do parâmetro de interesse, $\mathbf{y}$, foi gerado de acordo com a especificação do modelo espaciotemporal de estimação em pequenos domínios. Para tal, foi também considerado um modelo com p=2, ou seja, com uma variável auxiliar, X_{it}, e o termo independente, $\mathbf{x}_{it} = (1, x_{it})'$. Os mT valores de X_{it} foram gerados a partir de uma distribuição uniforme no intervalo [0,1]. Admitiu-se que os verdadeiros valores dos efeitos fixos são $\boldsymbol{\beta} = (1; 2)'$, as verdadeiras variâncias dos efeitos aleatórios específicos de domínio são $\sigma_u^2 \in \{0{,}5; 1{,}0\}$, as verdadeiras variâncias dos efeitos aleatórios específicos de domínio-tempo são $\sigma^2 \in \{0{,}25; 0{,}50; 1{,}00\}$, os valores dos parâmetros de autocorrelação temporal são $\rho \in \{0{,}2; 0{,}4; 0{,}8\}$ e os valores dos parâmetros de associação espacial são $\phi \in \{0{,}25; 0{,}50\}$. Utilizou-se uma matriz fixa de pesos espaciais, $\mathbf{W} = \{w_{ij}^*\}$, com pesos estandardizados por linhas da forma $w_{ij}^* = w_{ij} / w_{i\bullet}$, constituída com base em dados referentes à vizinhança por adjacência das NUTSIII de Portugal continental[108]. Admitiu-se que $\varepsilon_{it} \overset{iid}{\sim} N(0,1)$ e que ρ e ϕ são conhecidos. Note-se que, apesar dos métodos de reamostragem propostos poderem contemplar a situação de desconhecimento destes parâmetros, não seria possível deduzir uma aproximação analítica do EQMP admitindo que ρ e ϕ são conhecidos. Todos os erros do modelo foram gerados de forma mutuamente independente a partir de distribuições normais de média nula. Por último, o vector dos mT valores do parâmetro

[108] Chandra *et al.* (2007*a*, 2007*b*) utilizaram um procedimento semelhante para a escolha de uma matriz de pesos espaciais num estudo de simulação *model-based*, tendo utilizado informação sobre a contiguidade de domínios no âmbito do *The Environmental Monitoring and Assessment Program Survey*.

de interesse foi gerado através do modelo (5.2.5). Note-se que se está perante um estudo que envolve 36 combinações de componentes de variância.

O estudo por simulação de Monte Carlo, desenhado para comparar o desempenho dos estimadores do EQMP do EBLUP espaciotemporal, foi efectuado de acordo com o seguinte algoritmo:

1. Gerar L=1.000 conjuntos de dados iniciais, $\mathbf{y}^{(l)} = \left(y_{11}^{(l)},...,y_{it}^{(l)},...,y_{mT}^{(l)}\right)$, tal como descrito acima, l=1, …, L.

2. Calcular as estimativas das componentes de variância, $\hat{\sigma}_u^{2(l)}$ e $\hat{\sigma}^{2(l)}$, pelo método dos momentos com base nos dados iniciais, $\mathbf{y}^{(l)}$, e ajustar o modelo (5.2.5) de forma a determinar as estimativas dos efeitos fixos $\hat{\boldsymbol{\beta}}^{(l)} = \hat{\boldsymbol{\beta}}\left(\mathbf{y}^{(l)}; \hat{\boldsymbol{\psi}}^{(l)}\right)$, onde $\hat{\boldsymbol{\psi}}^{(l)} = \left(\hat{\sigma}_u^{2(l)}, \hat{\sigma}^{2(l)}, \rho, \phi\right)'$, para cada l=1, …, L.

3. Calcular as estimativas do parâmetro de interesse, $\hat{\theta}_{it}^{(l)}\left(\hat{\boldsymbol{\psi}}^{(l)}\right)$, e as suas estimativas analíticas do EQMP (EQMP-A), $eqmp^{A(l)}(\hat{\theta}_{it}^{(l)})$, para cada l=1, …, L. Para além disso, calcular as estimativas *jackknife* do EQMP do EBLUP (EQMP-J), $eqmp^{J(l)}\left(\hat{\theta}_{it}^{(l)}\right)$, utilizando os pesos[109] $w_{et} = (m-1)/m$, para cada l=1, …, L.

4. Gerar B=250 conjuntos de dados *bootstrap* com base nas estimativas $\hat{\sigma}_u^{2(l)}$ e $\hat{\sigma}^{2(l)}$, tal como descrito no subcapítulo 5.7, e depois calcular as estimativas *bootstrap* do EQMP do EBLUP (EQMP-B), $eqmp^{B(l)}(\hat{\theta}_{it}^{(l)})$, para cada l=1, …, L.

5. Calcular aproximações aos verdadeiros valores do EQMP do EBLUP para cada i-ésimo domínio em cada t-ésimo período de tempo, $EQMP_{it}$, os quais servem como termo de comparação. Esses valores foram calculados através de simulação de Monte

[109] No âmbito deste estudo empírico decidiu utilizar-se apenas os pesos mais simplistas, $w_{1et} = (m-1)/m$, porque se concluiu, no âmbito do estudo empírico apresentado na secção 7.3.2, que a utilização dos pesos $w_{2et} = 1 - \mathbf{x}_{et}'\left(\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}'\right)^{-1}\mathbf{x}_{et}$ não apresenta vantagens significativas, em termos de precisão e enviesamento, no desempenho dos estimadores *jackknife* do EQMP do EBLUP temporal.

Carlo, com base em R=5.000 conjuntos de dados independentes, de forma a assegurar uma melhor precisão nos resultados.

A avaliação da qualidade dos estimadores do EQMP do EBLUP espaciotemporal foi efectuada com base nas mesmas três medidas globais sobre os mT domínios de interesse: percentagem de domínios onde o enviesamento relativo é negativo (BRN), enviesamento relativo absoluto médio (BRAM), e EQM relativo médio (EQMRM). As medidas ao nível individual (domínio-tempo) que suportam o cálculo destas medidas globais são dadas pelas expressões (7.3.1) e (7.3.2.), onde $f \in \{A, B, J\}$ denota os diferentes estimadores do EQMP e l o l-ésimo conjunto de dados, l=1, …, 1.000. Foram igualmente utilizados diagramas em caixas de bigodes para representar as distribuições do enviesamento relativo dos estimadores do EQMP do EBLUP espaciotemporal.

Os programas desenvolvidos em SAS que permitiram a realização desta parte do estudo empírico estão reunidos no apêndice 24. As funções desempenhadas por aqueles programas são semelhantes às funções exercidas pelos programas desenvolvidos na primeira parte do estudo *model-based*, mas agora no âmbito do modelo espaciotemporal. Contudo, dada a maior complexidade da estrutura de covariância dos efeitos aleatórios do modelo, os tempos de processamento computacional associados a cada estimador do EQMP, em cada conjunto de dados e para cada combinação de parâmetros, são superiores aos anteriores. O tempo total de processamento computacional deste estudo por simulação foi de cerca de 140 dias.

Tal como se verificou no estudo empírico apresentado na secção 7.3.2, também neste estudo por simulação não ocorreram estimativas negativas do EQMP do EBLUP espaciotemporal, embora tivessem sido obtidas algumas estimativas negativas do termo de correcção de enviesamento. Em média verificaram-se 1,37% e 1,85% estimativas negativas do termo de correcção de enviesamento nos estimadores EQMP-B e EQMP-J, respectivamente.

7.3.3.2 Análise dos resultados do estudo

As tabelas 7.3.5 a 7.3.7 apresentam os resultados do desempenho dos diferentes estimadores do EQMP do EBLUP espaciotemporal, de acordo com as medidas BRN, BRAM e EQMRM, quando os dados da população artificial seguem o modelo postulado. Por sua vez, as figuras 7.3.5 a 7.3.7, contêm os diagramas em caixas de bigodes do enviesamento relativo desses estimadores. Numa análise prévia, pode notar-se que o desempenho dos três estimadores do EQMP depende significativamente dos valores da combinação das componentes de variância, σ^2 e σ_u^2.

A partir da análise da tabela 7.3.5 é possível observar que o estimador analítico é o que apresenta globalmente o melhor desempenho em termos de enviesamento (BRAM) e de precisão (EQMRM), quando se considera ρ=0,2. Contudo, os resultados apresentados nessa tabela também indicam que ambos os estimadores baseados em métodos de reamostragem podem competir com o estimador analítico, quer em termos de enviesamento (BRAM) quer de precisão (EQMRM), pois os valores obtidos nas respectivas medidas de qualidade estão muito próximos dos valores obtidos para o estimador analítico. Na realidade, o estimador *bootstrap* até é ligeiramente melhor do que o estimador analítico quando $\sigma^2 = 0{,}50$, enquanto o estimador *jackknife* apresenta pequenos ganhos de enviesamento e de precisão, relativamente ao estimador analítico quando $\sigma^2 = 1{,}00$ e $\phi = 0{,}5$. Relativamente aos estimadores baseados em métodos de reamostragem, note-se que o estimador *bootstrap* tende a ser levemente melhor do que o estimador *jackknife* para a generalidade de combinações de componentes de variância. A única excepção encontra-se quando $\sigma^2 = 1{,}00$, para a qual o estimador *jackknife* é melhor. Na tabela 7.3.5 pode, ainda, observar-se um comportamento globalmente semelhante para todos os estimadores, no que se refere às medidas BRN, BRAM e EQMRM, designadamente, (*i*) o valor do coeficiente de associação espacial não tem um impacto significativo na qualidade dos estimadores; (*ii*) a qualidade dos estimadores tende a melhorar, em termos de enviesamento e de precisão, para valores mais elevados dos parâmetros de variância (principalmente de σ^2); e (*iii*) a percentagem de estimativas do EQMP com enviesamento negativo diminui para valores mais elevados dos parâmetros de variância.

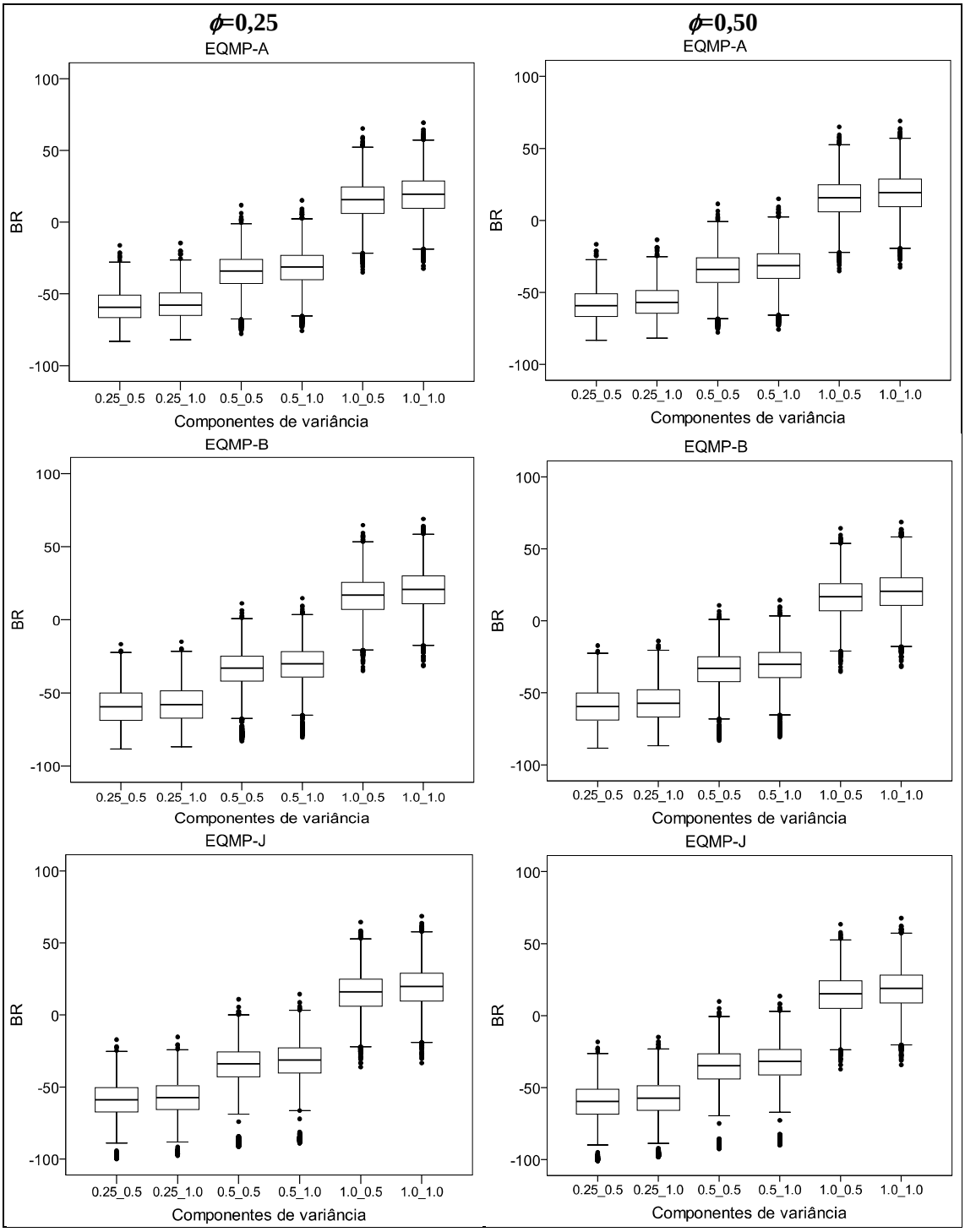
Tabela 7.3.5: Medidas de qualidade dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,2

σ^2	σ_u^2	**EQMP-A**	**EQMP-B**	**EQMP-J**	**EQMP-A**	**EQMP-B**	**EQMP-J**
			ϕ=0,25			**ϕ=0,50**	
				BRN (%)			
0,25	0,5	100,000	100,000	100,000	100,000	100,000	100,000
	1,0	100,000	100,000	100,000	100,000	100,000	100,000
0,50	0,5	99,908	99,867	99,898	99,908	99,878	99,949
	1,0	99,755	99,571	99,684	99,755	99,582	99,786
1,00	0,5	13,520	12,133	13,745	13,959	12,480	15,143
	1,0	9,153	8,051	8,990	9,296	8,245	10,010
				BRAM (%)			
0,25	0,5	58,674	59,872	60,027	58,753	59,995	60,817
	1,0	57,213	58,409	58,534	56,528	57,788	58,410
0,50	0,5	34,770	34,063	35,121	34,856	34,169	35,952
	1,0	31,991	31,246	32,346	32,121	31,421	33,037
1,00	0,5	17,287	18,090	17,392	17,294	18,124	16,935
	1,0	20,449	21,504	20,515	20,303	21,313	19,898
				EQMRM (%)			
0,25	0,5	26,138	27,592	27,770	26,261	27,757	28,541
	1,0	24,899	26,317	26,465	23,984	25,434	26,008
0,50	0,5	8,637	8,501	9,039	8,696	8,568	9,440
	1,0	7,374	7,250	7,758	7,442	7,332	8,064
1,00	0,5	1,925	2,087	1,960	1,933	2,103	1,880
	1,0	2,537	2,775	2,571	2,516	2,741	2,451

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

A figura 7.3.5 também ilustra que a subestimação do EQMP do EBLUP espaciotemporal diminui para valores mais elevados dos parâmetros de variância. Contudo, o enviesamento relativo médio não converge claramente para zero para valores mais elevados dos parâmetros de variância. Em particular, é de salientar que todas as estimativas do EQMP apresentam enviesamento relativo negativo quando $\sigma^2 = 0{,}25$, enquanto que cerca de 90% dessas estimativas apresentam enviesamento relativo positivo quando $\sigma^2 = 1{,}00$. Neste último caso, o enviesamento relativo médio é superior a zero. Na figura 7.3.5 observam-se ainda amplitudes dos intervalos de variação do enviesamento relativo da mesma ordem de grandeza para todos os estimadores.

Figura 7.3.5: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,2



A tabela 7.3.6 evidencia que o estimador analítico é novamente o melhor, para a generalidade dos casos, quando se considera ρ=0,4. A partir desta tabela, é também possível observar que, apesar do pior desempenho do estimador *bootstrap* quando comparado com o estimador analítico, aquele estimador consegue apresentar níveis de enviesamento e de precisão muito próximos dos observados para o estimador analítico.

De facto, ele chega mesmo a ser melhor do que o estimador analítico em termos de enviesamento (BRAM) quando $\sigma^2 = 0{,}50$. Por seu lado, o estimador *jackknife* é claramente o pior quando se considera $\sigma^2 = 0{,}50$. Contudo, o estimador *jackknife* consegue ser ligeiramente menos enviesado e mais preciso do que os outros dois estimadores quando $\sigma^2 = 1{,}00$. Por último, note-se que o comportamento globalmente semelhante observado para todos os estimadores no caso ρ=0,2, volta a estar claramente ilustrado na tabela 7.3.6, para o caso ρ=0,4.

Tabela 7.3.6: Medidas de qualidade dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,4

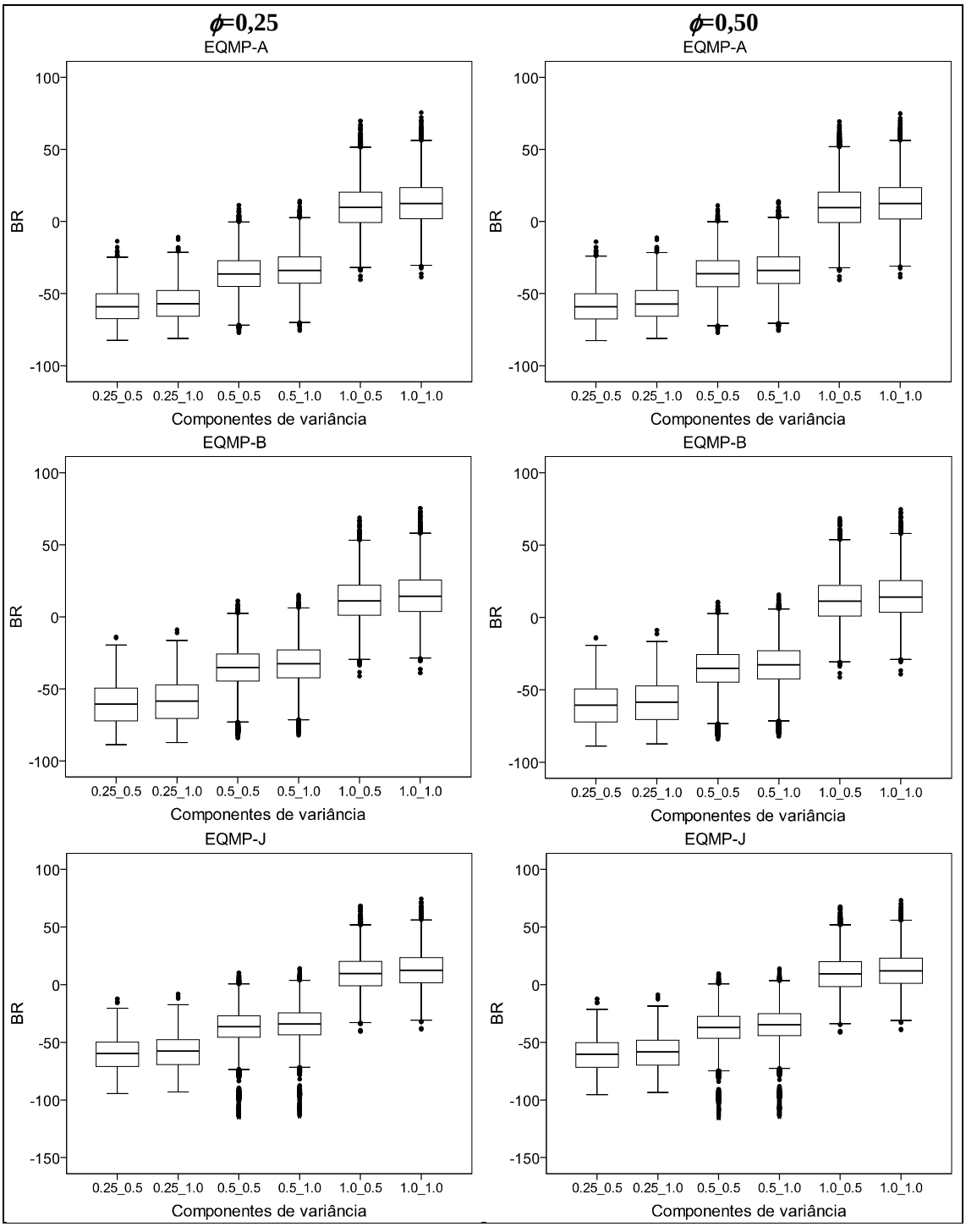
σ^2	σ_u^2	**EQMP-A**	**EQMP-B**	**EQMP-J**	**EQMP-A**	**EQMP-B**	**EQMP-J**
		ϕ=0,25			**ϕ=0,50**		
				BRN (%)			
0,25	0,5	100,000	100,000	100,000	100,000	100,000	100,000
	1,0	100,000	100,000	100,000	100,000	100,000	100,000
0,50	0,5	99,755	99,592	99,714	99,786	99,561	99,776
	1,0	99,429	99,010	99,327	99,408	99,051	99,439
1,00	0,5	26,316	22,939	27,235	26,347	22,908	28,143
	1,0	21,173	17,602	21,786	21,531	17,898	22,969
				BRAM (%)			
0,25	0,5	58,363	60,733	60,659	58,417	60,808	61,362
	1,0	56,305	58,634	58,606	56,385	58,757	59,187
0,50	0,5	36,561	36,026	38,342	36,595	36,037	38,970
	1,0	34,199	33,548	35,998	34,293	33,679	36,585
1,00	0,5	15,019	15,818	14,963	15,107	15,959	14,902
	1,0	16,875	18,092	16,886	16,853	18,052	16,592
				EQMRM (%)			
0,25	0,5	26,036	28,725	28,636	26,128	28,852	29,338
	1,0	23,948	26,509	26,449	24,058	26,661	27,005
0,50	0,5	9,741	9,829	11,562	9,779	9,864	11,911
	1,0	8,592	8,658	10,357	8,649	8,727	10,651
1,00	0,5	1,645	1,812	1,641	1,665	1,846	1,624
	1,0	2,025	2,294	2,035	2,024	2,288	1,971

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

A partir da figura 7.3.6 verifica-se que todos os estimadores têm uma propensão para apresentar intervalos de variação de enviesamento relativo com maiores amplitudes, para valores mais elevados de σ^2. Verifica-se também, à semelhança do que foi observado para o caso ρ=0,2, que a subestimação do EQMP do EBLUP diminui para valores mais elevados dos parâmetros de variância, apesar do enviesamento relativo médio não parecer convergir para zero para valores mais elevados desses parâmetros.

Por último, a análise da figura 7.3.6 permite, ainda, validar o pior desempenho do estimador *jackknife* quando $\sigma^2 = 0{,}50$, uma vez que apresenta um avultado número de domínios com valores de enviesamento relativo considerados *outliers*, sobretudo na aba esquerda das distribuições. Esta situação indica, nesses casos, uma subestimação mais acentuada do verdadeiro valor do EQMP do EBLUP, do que os restantes estimadores.

Figura 7.3.6: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,4



A tabela 7.3.7 mostra que o estimador analítico é globalmente o melhor quando se considera ρ=0,8. Com efeito, nessa tabela pode observar-se que esse estimador apresenta, nesta situação, ganhos relevantes em termos de enviesamento (BRAM) e de precisão (EQMRM) para valores mais baixos de σ^2, face aos estimadores por reamostragem. Porém, para valores elevados desse parâmetro de variância, os níveis de enviesamento e de precisão de todos os estimadores são semelhantes, sobressaindo o estimador *jackknife* como o que apresenta um desempenho ligeiramente melhor. Na tabela 7.3.7 pode ainda descortinar-se um comportamento globalmente semelhante, em termos de enviesamento e de precisão, entre os dois estimadores por reamostragem, sendo o estimador *bootstrap* melhor apenas quando $\sigma^2 = 0{,}50$. A tabela 7.3.7 evidencia ainda que todos os estimadores apresentam os mesmos níveis de subestimação do verdadeiro EQMP do EBLUP, para qualquer combinação de parâmetros. É de salientar que essa subestimação ocorre na maioria dos domínios, apesar de diminuir com o aumento do valor dos parâmetros de variância. A partir da análise da tabela 7.3.7 ressalta também que o coeficiente de associação espacial não parece influenciar a qualidade do estimador do EQMP, à semelhança dos casos anteriores (ρ=0,2 e ρ=0,4). Por último, é de notar que a qualidade de todos os estimadores melhora para valores mais elevados dos parâmetros de variância (σ^2 e σ_u^2), também à semelhança dos casos anteriores.

Tabela 7.3.7: Medidas de qualidade dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,8

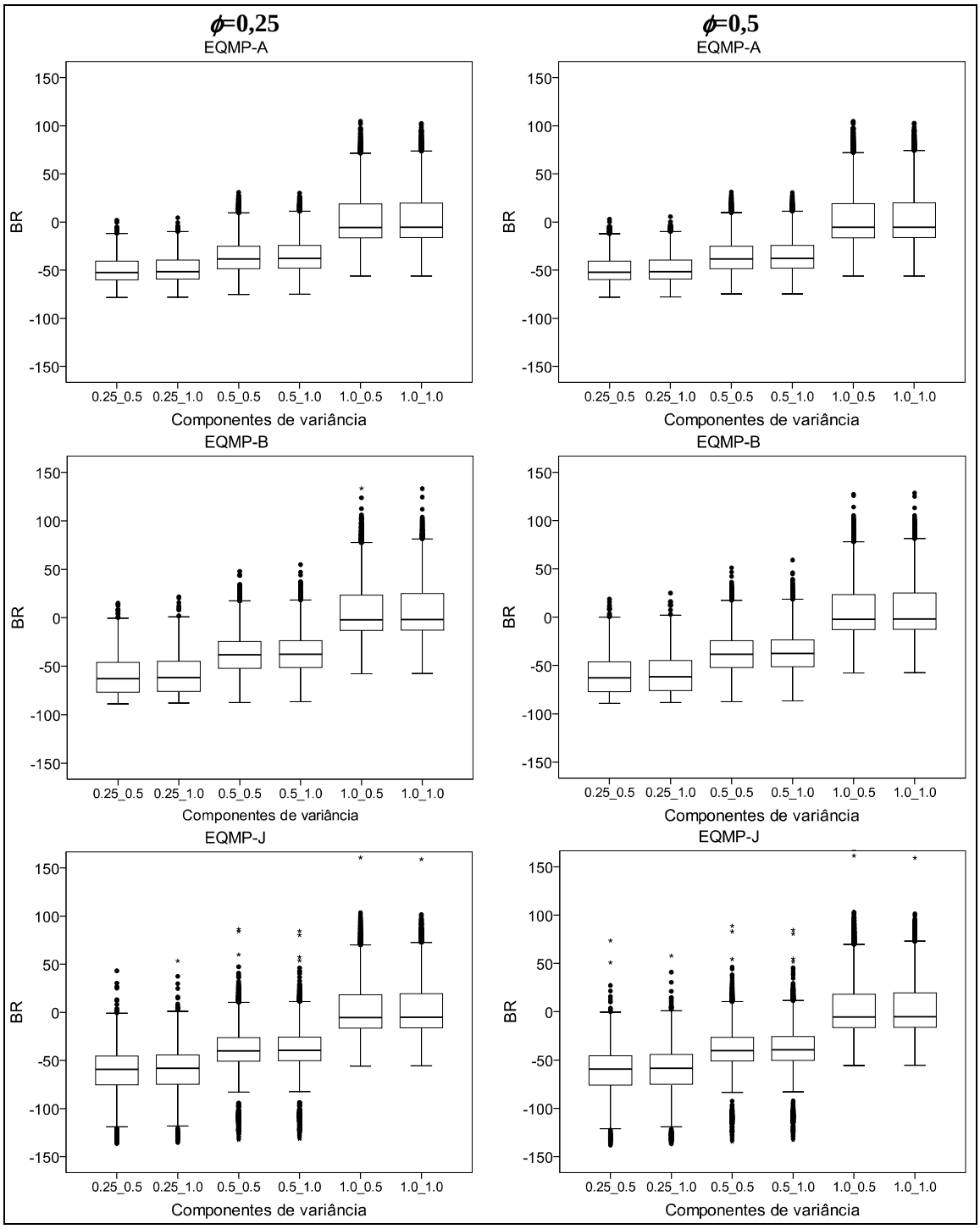
σ^2	σ_u^2	EQMP-A	EQMP-B	EQMP-J	EQMP-A	EQMP-B	EQMP-J
		ϕ=0,25			ϕ=0,50		
				BRN (%)			
0,25	0,5	99,990	99,898	99,878	99,980	99,878	99,867
	1,0	99,990	99,878	99,827	99,980	99,837	99,827
0,50	0,5	96,378	94,061	95,827	96,051	93,969	95,827
	1,0	96,163	93,469	95,602	95,990	93,184	95,245
1,00	0,5	59,765	54,153	59,224	59,571	54,122	59,255
	1,0	59,469	53,673	58,929	59,306	53,561	59,061
				BRAM (%)			
0,25	0,5	50,048	60,057	59,451	49,776	60,065	59,744
	1,0	49,143	58,869	58,279	49,062	58,844	58,490
0,50	0,5	36,395	38,984	40,039	36,257	38,966	40,244
	1,0	35,730	38,231	39,359	35,662	38,218	39,489
1,00	0,5	21,665	21,954	21,534	21,789	21,964	21,533
	1,0	21,666	22,110	21,612	21,773	22,173	21,628
				EQMRM (%)			
0,25	0,5	19,944	29,258	28,975	19,774	29,329	29,340
	1,0	19,173	28,053	27,831	19,131	28,100	28,107
0,50	0,5	10,696	12,913	13,590	10,636	12,927	13,751
	1,0	10,350	12,460	13,192	10,320	12,467	13,305
1,00	0,5	3,502	3,883	3,484	3,555	3,889	3,484
	1,0	3,494	3,943	3,503	3,539	3,969	3,504

Nota: BRN-enviesamento relativo negativo; BRAM-enviesamento relativo absoluto médio; EQMRM-EQM relativo médio.

Na figura 7.3.7 é possível observar que todos os estimadores subestimam o verdadeiro EQMP na maioria dos domínios, para todas as combinações de componentes de variância. Observa-se, também, que o estimador *jackknife* apresenta as maiores amplitudes nos intervalos de variação dos enviesamentos relativos. Para além disso, este estimador manifesta um elevado número de domínios com valores de enviesamento relativo considerados *outliers* em ambas as abas das distribuições ($\sigma^2 = 0{,}25$ e $\sigma^2 = 0{,}50$), ao contrário do que se verifica para os outros dois estimadores. Esta situação indica, que o estimador *jackknife* apresenta uma maior propensão para subestimar ou sobreestimar, de forma severa, o verdadeiro valor do EQMP do EBLUP, do que os restantes estimadores. Pela análise da figura 7.3.7 verifica-se, ainda, que os estimadores analítico e *bootstrap* apresentam distribuições de enviesamento relativo muito semelhantes. Por todas estas razões, pode concluir-se que, em termos de enviesamento, o estimador *bootstrap* é o único que consegue competir com o estimador

analítico, apesar da clara vantagem deste último estimador, sobretudo quando $\sigma^2 = 0{,}25$.

Figura 7.3.7: Diagramas em caixa de bigodes do BR dos estimadores do EQMP do EBLUP espaciotemporal, para ρ=0,8



7.3.3.3 Síntese e discussão dos resultados

Numa análise transversal a todos os estimadores, para diferentes níveis de autocorrelação, pode concluir-se que:

(*i*) o enviesamento tende a diminuir e a precisão a melhorar para valores mais elevados de ambos os parâmetros de variância (σ^2 e σ_u^2);

(*ii*) o coeficiente de associação espacial não tem um impacte significativo na qualidade dos estimadores, embora se observe uma ténue perda de qualidade para valores mais elevados desse coeficiente;

(*iii*) o coeficiente de autocorrelação temporal também não tem um impacte significativo na qualidade dos estimadores;

(*iv*) os dois estimadores por reamostragem apresentam níveis de enviesamento e de precisão próximos dos obtidos para o estimador analítico, sendo em alguns casos até ligeiramente melhores;

(*v*) os dois estimadores por reamostragem apresentam níveis de enviesamento e de precisão muito próximos entre si, não existindo um estimador por reamostragem que seja invariavelmente melhor do que o outro.

Perante estes resultados, verifica-se que não existe um único estimador que seja uniformemente melhor do que os outros, em termos de enviesamento e de precisão. Tal situação já seria de esperar, tendo em conta os resultados obtidos na primeira parte deste estudo empírico, na qual também não foi possível identificar um estimador do EQMP do EBLUP temporal sistematicamente melhor do que os restantes.

Através de uma análise comparativa global dos estimadores por reamostragem, pode verificar-se que apesar da grande semelhança entre o desempenho desses dois estimadores, o estimador *bootstrap* parece apresentar uma ligeira desvantagem, em termos de enviesamento (BRAM) e de precisão (EQMRM), relativamente ao estimador *jackknife*. Contudo, o estimador *jackknife* revela a desvantagem de poder apresentar enviesamentos mais acentuados do que o estimador *bootstrap*, para valores mais elevados do coeficiente de autocorrelação (ρ=0,4 e ρ=0,8). Estes resultados, aliados ao

facto do estimador *bootstrap* ser mais fácil implementar porque não exige a dedução de uma expressão fechada, mesmo no contexto de modelos de estimação em domínios mais complexos, indica que ele deva ser o estimador por reamostragem preferido na mediação da incerteza associada ao EBLUP espaciotemporal. Tal alvitre é concordante com a sugestão da primeira parte deste estudo empírico, na qual se avaliou o desempenho de estimadores do EQMP do EBLUP temporal.

É altura de relembrar que o estimador *jackknife* do EQMP do EBLUP espaciotemporal, avaliado no âmbito desta parte do estudo empírico do tipo *model-based*, está a ser utilizado num contexto em que se admite a hipótese de existência de dependência espacial entre os domínios de interesse. Tal como foi referido anteriormente, a extracção repetida de um conjunto de T observações pertencentes a cada domínio pode provocar alterações na estrutura de associação inter-domínio, gerando estimativas inconsistentes. Tendo em conta os resultados obtidos nas duas partes do estudo empírico *model-based*, designadamente, (*i*) o estimador *jackknife* do EQMP do EBLUP espaciotemporal apresenta um comportamento globalmente semelhante ao dos outros estimadores, à imagem do que ocorre com os estimadores *jackknife* do EQMP do EBLUP temporal (nos quais não se verifica essa hipotética quebra de estrutura); e (*ii*) todos os estimadores *jackknife* do EQMP dos EBLUP temporal e espaciotemporal apresentam a mesma tendência para subestimar de forma mais acentuada o verdadeiro valor do EQMP, para valores mais elevados do coeficiente de autocorrelação (ρ=0,4 e ρ=0,8); então não parece existir evidência empírica que permita excluir os estimadores do tipo *jackknife* em situações nas quais poderá eventualmente existir quebra da estrutura de associações.

Em jeito de conclusão, os resultados alcançados neste estudo empírico mostram que os estimadores *bootstrap* e *jackknife* apresentam um desempenho muito bom quando comparados com o estimador analítico, na medição da incerteza associada às estimativas do EBLUP espaciotemporal. Mais uma vez, existe evidência empírica que permite considerar como adequado o uso de estimadores baseados em métodos por reamostragem, e em particular o estimador *bootstrap*, na estimação da incerteza associada aos estimadores EBLUP, como alternativa aos estimadores baseados em longos desenvolvimentos analíticos.

7.4 ESTIMATIVAS DO PREÇO MÉDIO DE TRANSACÇÃO DA HABITAÇÃO

7.4.1 Introdução

O último subcapítulo do estudo empírico é dedicado à apresentação das estimativas do preço médio de transacção da habitação em Portugal continental. As estimativas foram produzidas com base em dados reais obtidos através do IPTH, realizado nos quatro trimestres de 2002 e nos três primeiros trimestres de 2003, e através do IABH realizado no último trimestre de 2001, nos quatro trimestres de 2002 e nos dois primeiros trimestres de 2003.

Para além das estimativas produzidas pelo melhor estimador no contexto deste problema real, identificado no estudo empírico por simulação apresentado no subcapítulo 7.2, decidiu também apresentar-se as estimativas produzidas através de todos os outros estimadores em avaliação nesse estudo. Em particular, será dado especial destaque às estimativas produzidas pelo estimador EBLUP espaciotemporal com restrições ao nível de NUTSII, as quais, como se concluiu no referido estudo empírico, apesar de não apresentarem os melhores níveis de enviesamento nem de precisão, garantem a consistência interna na publicação das estimativas.

Para todas as estimativas, foram calculadas também as respectivas estimativas dos EPR: $epr\left(\hat{\mu}_{ti}^{f}\right)=\sqrt{\hat{V}\left(\hat{\mu}_{ti}^{f}\right)}\Big/\hat{\mu}_{ti}^{f}$, onde *f=dir, sinR, sinQ, com* e $epr\left(\hat{\mu}_{ti}^{f}\right)=\sqrt{eqmp\left(\hat{\mu}_{ti}^{f}\right)}\Big/\hat{\mu}_{ti}^{f}$, onde $f=FH,NS,RY,LP,LPR2$ [110]. As expressões analíticas dos estimadores utilizados, bem como a metodologia de estimação das variâncias ou dos EQMP, respectivamente dos estimadores tradicionais e dos estimadores EBLUP, encontram-se nas subsecções 7.2.4.2 e 7.2.4.3. É de notar que as estimativas dos EPR dos estimadores sintético pelo quociente, sintético pela regressão e combinado estão subestimadas pelo facto de não contemplarem o enviesamento desses estimadores, o qual não é possível estimar no contexto de problemas reais para os quais a população finita é desconhecida.

[110] Note-se que no caso dos estimadores centrados o EPR é igual ao CV.

Para tornar este texto menos denso, vão apresentar-se apenas as estimativas referentes ao último trimestre conhecido (terceiro trimestre de 2003), remetendo-se as estimativas dos restantes trimestres para apêndices.

7.4.2 Estimativas que não garantem a consistência interna

Nesta secção são apresentadas estimativas do preço médio de transacção da habitação produzidas através de estimadores que não garantem a consistência interna na publicação das estimativas. Começa por apresentar-se as estimativas directas do preço médio de transacção ao nível de Portugal continental e das respectivas NUTSII (a divisão de Portugal em NUTS, de acordo com o Decreto-lei n.º 244/2002, pode ser consultada no apêndice 25). A tabela 7.4.1 apresenta, para estes níveis de agregação, as dimensões amostrais de empresas (a_i) e de transacções (n_i), as estimativas directas do parâmetro de interesse ($\hat{\mu}_i^{dir}$) e as respectivas estimativas dos EPR $\left[epr\left(\hat{\mu}_i^{dir}\right)\right]$, referentes ao terceiro trimestre de 2003. As estimativas dos outros seis trimestres encontram-se no apêndice 26.

Tabela 7.4.1: Estimativas directas do preço médio de transacção da habitação ao nível de Portugal continental e das respectivas NUTSII (valores em euros/m^2), referentes ao terceiro trimestre de 2003

Regiões		a_i	n_i	$\hat{\mu}_i^{dir}$	$epr\left(\hat{\mu}_i^{dir}\right)$
NUTSII	Norte	124	564	970	6,2%
	Centro	110	387	854	2,6%
	Lisboa	178	751	1.313	2,7%
	Alentejo	32	121	1.046	11,1%
	Algarve	34	89	1.189	5,6%
Portugal continental		**458**	**1.912**	**1.125**	**2,6%**

Tal como seria de esperar, o estimador directo produz estimativas do preço médio de transacção da habitação com muito boa precisão para Portugal continental, e com uma boa precisão ao nível de NUTSII. A única excepção refere-se à NUTSII *Alentejo*, como resultado de uma maior variabilidade e de um pequeno número de observações. Estas são as estimativas utilizadas no âmbito do estimador EBLUP espaciotemporal com restrições[111], de forma a garantir a consistência interna na publicação das estimativas.

[111] De acordo com Rao (2003), se as estimativas directas forem de boa qualidade para determinados

A partir da tabela 7.4.1 verifica-se que no terceiro trimestre de 2003, a estimativa directa do preço médio de transacção da habitação em Portugal continental é de 1.125 €/m^2. Por regiões NUTSII do continente, verifica-se que a estimativa do preço médio mais baixa se refere à região *Centro* (854 €/m^2), enquanto a mais alta se refere à região de *Lisboa* (1.313 €/m^2).

Em seguida são apresentadas as estimativas directas do preço médio de transacção ao nível de NUTSIII. Na tabela 7.4.2 é apresentada, para cada uma dessas regiões, a dimensão amostral de empresas (a_i) e de transacções (n_i), a estimativa directa do parâmetro de interesse ($\hat{\mu}_i^{dir}$) e as respectivas estimativas dos EPR $\left[epr\left(\hat{\mu}_i^{dir}\right)\right]$, referentes ao terceiro trimestre de 2003. As estimativas relativas aos outros seis trimestres encontram-se no apêndice 27.

Tabela 7.4.2: Estimativas directas do preço médio de transacção da habitação ao nível de NUTSIII (valores em euros/m^2), referentes ao terceiro trimestre de 2003

NUTSIII	a_i	n_i	$\hat{\mu}_i^{dir}$	$epr\left(\hat{\mu}_i^{dir}\right)$	NUTSIII	a_i	n_i	$\hat{\mu}_i^{dir}$	$epr\left(\hat{\mu}_i^{dir}\right)$
111	9	23	769	9,8%	167	1	1	500	-
112	10	31	652	6,2%	168	2	6	726	2,8%
113	12	39	736	4,4%	169	5	12	782	10,7%
114	78	405	1.054	7,1%	16A	5	17	648	6,4%
115	4	19	625	6,8%	16B	29	77	948	5,6%
116	11	22	686	4,2%	171	111	488	1.389	3,0%
117	1	7	626	-	172	70	263	1.091	4,9%
118	3	18	744	0,3%	16C	7	34	817	5,6%
161	14	49	860	6,7%	185	10	40	817	4,1%
162	25	90	948	4,2%	181	4	26	1.357	9,1%
163	13	56	730	4,8%	182	7	19	783	12,5%
164	8	27	651	3,3%	183	7	24	970	7,5%
165	7	17	852	4,1%	184	5	12	1.074	9,4%
166	1	1	377	-	150	34	89	1.189	5,6%

Quando se efectua a estimação do preço médio de transacção da habitação ao nível das NUTSIII do continente através do estimador directo, verifica-se que a qualidade da estimação se degrada substancialmente. O EPR médio das 25 NUTSIII para as quais as estimativas da variância não são omissas é igual a 6,0%[112], ultrapassando em dois casos os 10,0%.

níveis de agregação, então elas podem ser utilizadas de forma a garantir a consistência interna.

[112] Note-se que, tal como apresentado na subsecção 7.2.4.2, as estimativas da variância do estimador directo baseadas no método delta são omissas se $a_{th'} = 1$ para h'=1, …, H.

Ao nível das regiões NUTSIII, a análise das estimativas directas do valor médio de transacção da habitação revela que em apenas seis das 28 regiões esse valor se situa acima dos 1.000 €/m^2, designadamente no *Grande Porto, Grande Lisboa, Península de Setúbal, Alentejo Litoral, Baixo Alentejo* e *Algarve*. Pelo contrário, o valor mais baixo é de 377 €/m^2 e refere-se ao *Pinhal Interior Sul*.

Em seguida, nas tabelas 7.4.3 e 7.4.4, são apresentadas as estimativas dos parâmetros de interesse e as respectivas estimativas dos EPR, ao nível de NUTSIII, produzidas respectivamente através dos estimadores tradicionais e dos estimadores EBLUP, referentes ao terceiro trimestre de 2003. As estimativas relativas aos outros seis trimestres encontram-se, respectivamente, nos apêndices 28 e 29.

Tabela 7.4.3: Estimativas do preço médio de transacção da habitação ao nível de NUTSIII (valores em euros/m^2) produzidas através dos estimadores tradicionais, referentes ao terceiro trimestre de 2003

NUTSIII	$\hat{\mu}_i^{dir}$	$\hat{\mu}_i^{sinQ}$	$\hat{\mu}_i^{sinR}$	$\hat{\mu}_i^{com}$	$epr(\hat{\mu}_i^{dir})$	$epr(\hat{\mu}_i^{sinQ})$	$epr(\hat{\mu}_i^{sinR})$	$epr(\hat{\mu}_i^{com})$
111	769	968	792	792	9,8%	6,2%	3,3%	3,3%
112	652	885	708	681	6,2%	6,2%	4,2%	3,7%
113	736	859	682	717	4,4%	6,2%	4,7%	3,3%
114	1.054	1.149	973	985	7,1%	6,2%	3,5%	3,1%
115	625	839	662	662	6,8%	6,2%	5,1%	5,1%
116	686	902	726	707	4,2%	6,2%	4,0%	2,9%
117	626	791	615	626	-	6,2%	6,2%	-
118	744	767	590	744	0,3%	6,2%	6,8%	0,3%
161	860	838	835	835	6,7%	2,6%	3,2%	3,2%
162	948	955	976	976	4,2%	2,6%	3,5%	3,5%
163	730	809	800	747	4,8%	2,6%	3,3%	3,6%
164	651	707	677	669	3,3%	2,6%	4,8%	3,4%
165	852	749	727	842	4,1%	2,6%	4,0%	3,8%
166	377	729	703	377	-	2,6%	4,3%	-
167	500	643	600	500	-	2,6%	6,6%	-
168	726	677	640	721	2,8%	2,6%	5,6%	2,7%
169	782	716	687	708	10,7%	2,6%	4,6%	4,3%
16A	648	760	740	666	6,4%	2,6%	3,8%	5,0%
16B	948	663	997	997	5,6%	2,6%	3,6%	3,6%
171	1.389	1.376	1.411	1.411	3,0%	2,7%	6,0%	6,0%
172	1.091	1.138	1.135	1.135	4,9%	2,7%	4,5%	4,5%
16C	817	612	908	841	5,6%	2,6%	3,2%	4,1%
185	817	736	889	833	4,1%	11,1%	3,2%	3,2%
181	1.357	1.209	1.033	1309	9,1%	11,1%	3,8%	8,1%
182	783	936	760	760	12,5%	11,1%	3,6%	3,6%
183	970	1.098	922	922	7,5%	11,1%	3,2%	3,2%
184	1.074	927	751	1.043	9,4%	11,1%	3,7%	8,7%
150	1.189	1.189	1.168	792	5,6%	5,6%	4,7%	3,3%

Pela análise da tabela 7.4.3 verifica-se que os estimadores sintéticos, que como foi verificado anteriormente são estimadores enviesados e pouco precisos do ponto de vista *design-based*, tendem a suavizar as estimativas dos parâmetros de interesse, ou seja, não produzem estimativas muito baixas nem muito altas desses parâmetros. É, também, de salientar, que as estimativas dos EPR associadas aos estimadores sintéticos e ao estimador combinado podem sugerir uma boa qualidade das estimativas, quando na realidade isso não ocorre. Relembre-se que estas estimativas não têm em consideração o enviesamento desses estimadores.

Tabela 7.4.4: Estimativas do preço médio de transacção da habitação ao nível de NUTSIII (valores em euros/m^2) produzidas através dos estimadores EBLUP, referentes ao terceiro trimestre de 2003

NUTSIII	$\widehat{\mu}_i^{FH}$	$\widehat{\mu}_i^{NS}$	$\widehat{\mu}_{it}^{RY}$	$\widehat{\mu}_{it}^{LP}$	$epr(\widehat{\mu}_i^{FH})$	$epr(\widehat{\mu}_i^{NS})$	$epr(\widehat{\mu}_{it}^{RY})$	$epr(\widehat{\mu}_{it}^{LP})$
111	775	757	823	832	8,6%	8,7%	5,0%	4,4%
112	657	670	738	752	5,9%	5,7%	5,0%	4,7%
113	733	734	742	755	4,3%	4,3%	4,5%	4,3%
114	1.034	1.012	1.038	1.040	6,4%	6,3%	3,6%	3,4%
115	628	648	656	696	6,5%	6,2%	5,4%	4,9%
116	687	691	709	729	4,2%	4,1%	4,3%	4,2%
117	626	631	631	637	3,5%	3,5%	4,2%	4,1%
118	744	743	744	742	1,0%	1,0%	1,1%	1,1%
161	855	860	877	880	6,3%	6,0%	4,4%	4,2%
162	951	957	952	960	4,0%	4,0%	3,7%	3,7%
163	734	732	737	754	4,7%	4,6%	4,7%	4,5%
164	651	645	663	640	3,3%	3,3%	3,7%	3,9%
165	843	838	800	791	4,0%	4,0%	4,3%	4,3%
166	386	390	404	447	5,7%	5,6%	6,4%	5,8%
167	503	507	496	513	4,4%	4,3%	5,2%	5,0%
168	724	723	722	706	2,8%	2,8%	3,2%	3,4%
169	754	701	789	673	9,6%	9,6%	5,9%	5,6%
16A	656	647	607	652	6,1%	6,0%	6,0%	5,5%
16B	956	947	987	1003	5,2%	5,1%	3,9%	3,6%
171	1.391	1.383	1.403	1.405	3,0%	3,0%	2,6%	2,6%
172	1.098	1.112	1.094	1.149	4,6%	4,4%	3,6%	3,3%
16C	828	807	790	826	5,3%	5,3%	4,7%	4,3%
185	822	821	816	857	4,0%	4,0%	3,8%	3,8%
181	1.201	1.223	1.194	1.121	7,8%	6,8%	5,0%	3,3%
182	775	716	796	752	10,4%	10,2%	5,4%	4,8%
183	958	991	1.015	990	6,8%	6,2%	4,2%	3,8%
184	951	963	931	881	8,6%	8,3%	4,3%	4,1%
150	1.185	1.220	1.145	1.203	5,2%	5,0%	3,7%	3,2%

Da análise da tabela 7.4.4 ressalta que as estimativas produzidas pelo estimador espaciotemporal proposto são aquelas que apresentam os menores EPR. Note-se que cerca de 90% dessas estimativas apresentam EPR não superiores a 5%. Relativamente

às estimativas do preço médio de transacção da habitação, é de salientar a aparente concordância que se verifica entre as estimativas EBLUP espaciotemporais e as estimativas directas, uma vez que os valores mais elevados se estimam igualmente para o *Grande Porto, Grande Lisboa, Península de Setúbal, Alentejo Litoral, Baixo Alentejo* e *Algarve*, enquanto os mais baixos se estimam para o *Pinhal Interior Sul* e para a *Serra da Estrela*.

Por último, apresenta-se na tabela 7.4.5 a contribuição de cada componente do estimador do EQMP na medição da incerteza associada ao EBLUP seccional e ao EBLUP espaciotemporal. Os resultados relativos aos estimadores do EQMP dos EBLUP espacial e temporal encontram-se no apêndice 30. Relembre-se que o estimador do EQMP do EBLUP sem restrições é decomposto em três componentes: $g_1(\widehat{\boldsymbol{\psi}})$, que mede a variabilidade devida à estimação dos efeitos aleatórios; $g_2(\widehat{\boldsymbol{\psi}})$, que mede a variabilidade devida à estimação dos efeitos fixos; e $2g_3(\widehat{\boldsymbol{\psi}})$, que, numa parte, mede a variabilidade devida à estimação das componentes de variância e, na outra parte, inclui uma correcção de enviesamento, repartidas em partes iguais.

Quer pela análise dos resultados apresentados na tabela 7.4.5, quer pela análise dos resultados sintetizados no apêndice 30, sobressai que: (*i*) a variabilidade devida à estimação dos efeitos fixos é insignificante, não ultrapassando os 2% do EQMP total, em qualquer dos estimadores EBLUP utilizados; e (*ii*) a variabilidade devida à estimação das componentes de variância associada aos estimadores EBLUP assistidos por modelos longitudinais (RY e LP) chega a representar cerca de 20% do EQMP total, enquanto no caso dos outros dois estimadores nem ultrapassa 1% do respectivo EQMP. É também interessante notar, que a variabilidade devida à estimação dos efeitos aleatórios representa uma proporção muito menor do EQMP total no caso dos estimadores EBLUP de RY e de LP, do que no caso dos outros dois estimadores EBLUP. Estes resultados ilustram bem a importância de se utilizarem estimadores do EQMP dos EBLUP de RY e de LP, que considerem a variabilidade devida à estimação das componentes de variância. Caso essa componente não seja considerada, *i.e.*, caso se utilize um estimador simplista, então a subestimação das estimativas do EQMP pode ser severa, comprometendo desta forma uma boa avaliação da qualidade das estimativas EBLUP do preço médio de transacção da habitação ao nível de NUTSIII.

Tabela 7.4.5: Estimativas do EQMP dos estimadores EBLUP de Fay-Herriot e espaciotemporal do preço médio de transacção da habitação, referentes ao terceiro trimestre de 2003

NUTSIII	$eqmp(\hat{\mu}_i^{FH})$	$g_1(\hat{\psi})$	$g_2(\hat{\psi})$	$2g_3(\hat{\psi})$	$eqmp(\hat{\mu}_{it}^{LP})$	$g_1(\hat{\psi})$	$g_2(\hat{\psi})$	$2g_3(\hat{\psi})$
111	4.462	95,2%	1,0%	3,8%	1.331	60,8%	2,9%	36,4%
112	1.498	97,9%	0,5%	1,7%	1.253	47,8%	0,2%	52,0%
113	1.001	98,5%	0,4%	1,2%	1.068	45,9%	0,0%	54,1%
114	4.380	94,6%	1,6%	3,7%	1.233	61,0%	2,2%	36,8%
115	1.685	97,5%	0,7%	1,8%	1.178	49,2%	0,1%	50,8%
116	817	98,8%	0,2%	1,0%	926	46,8%	0,1%	53,1%
117	490	99,2%	0,3%	0,6%	680	45,6%	0,0%	54,4%
118	60	99,9%	0,0%	0,1%	71	79,0%	0,0%	20,9%
161	2.872	96,5%	0,7%	2,8%	1.340	54,6%	0,6%	44,8%
162	1.472	97,8%	0,6%	1,6%	1.255	47,6%	1,6%	50,7%
163	1.167	98,4%	0,3%	1,3%	1.131	47,2%	0,0%	52,8%
164	465	99,3%	0,2%	0,6%	630	49,9%	0,0%	50,1%
165	1.146	98,4%	0,3%	1,3%	1.175	46,8%	0,5%	52,7%
166	489	99,2%	0,2%	0,6%	665	47,8%	0,1%	52,1%
167	490	99,1%	0,3%	0,6%	660	48,1%	0,2%	51,8%
168	410	99,3%	0,2%	0,5%	564	51,3%	0,1%	48,7%
169	5.223	94,2%	1,7%	4,1%	1.409	63,2%	0,8%	36,0%
16A	1.585	97,8%	0,4%	1,7%	1.307	48,1%	0,6%	51,3%
16B	2.505	96,4%	1,1%	2,5%	1.330	53,1%	1,2%	45,7%
171	1.702	94,3%	4,0%	1,7%	1.357	45,9%	8,1%	46,0%
172	2.588	95,3%	2,2%	2,5%	1.400	52,8%	2,0%	45,2%
16C	1.940	97,4%	0,6%	2,1%	1.253	50,5%	0,9%	48,6%
185	1.071	98,5%	0,3%	1,2%	1.060	50,0%	0,1%	49,9%
181	8.725	91,1%	4,1%	4,8%	1.394	71,6%	2,5%	25,9%
182	6.503	93,8%	1,6%	4,6%	1.300	68,6%	0,1%	31,4%
183	4.220	95,1%	1,2%	3,7%	1.425	60,7%	0,1%	39,2%
184	6.721	93,7%	1,7%	4,6%	1.332	66,4%	1,2%	32,4%
150	3.749	93,2%	3,5%	3,3%	1.491	54,0%	7,4%	38,6%
Média	---	96,8%	1,1%	2,1%	---	54,1%	1,2%	44,7%

7.4.3 Estimativas que garantem a consistência interna

Na tabela 7.4.6 são apresentadas as estimativas do preço médio de transacção da habitação ao nível de NUTSIII, que garantem a consistência interna ao nível de NUTSII, ou seja, que garantem que a média ponderada dessas estimativas pertencentes a cada NUTSII iguala a estimativa directa do preço médio de transacção da habitação relativa à NUTSII correspondente[113]. A avaliação da precisão das estimativas, que

[113] É de salientar que existem pequenas diferenças entre as estimativas resultantes da média ponderada das estimativas ao nível de NUTSIII nas várias NUTSII e as respectivas estimativas directas do preço médio de transacção da habitação ao nível de NUTSII. A diferença relativa média entre essas estimativas

garantem essa consistência interna, foi efectuada com base no estimador *bootstrap* do EQMP do EBLUP com restrições (6.3.26). À semelhança dos casos anteriores, as estimativas apresentadas na tabela 7.4.6 são referentes apenas ao terceiro trimestre de 2003. As estimativas referentes aos outros seis trimestres encontram-se no apêndice 31. As estimativas directas ao nível de NUTSII, utilizadas nesta estimação, são as apresentadas na tabela 7.4.1, enquanto os ponderadores utilizados podem ser consultados no apêndice 32. O programa em SAS produzido para fazer esta estimação encontra-se no apêndice 33.

Tabela 7.4.6: Estimativas EBLUP do preço médio de transacção da habitação ao nível de NUTSIII (valores em euros/m^2) que garantem a consistência interna ao nível de NUTSII, referentes ao terceiro trimestre de 2003

NUTSIII	$\widehat{\mu}_{it}^{LPR2}$	$epr(\widehat{\mu}_{it}^{LPR2})$	NUTSIII	$\widehat{\mu}_{it}^{LPR2}$	$epr(\widehat{\mu}_{it}^{LPR2})$
111	835	10,9%	167	504	17,2%
112	758	11,3%	168	695	14,3%
113	760	10,4%	169	671	13,9%
114	1.057	7,3%	16A	645	12,3%
115	694	8,5%	16B	1.000	9,0%
116	727	12,5%	171	1.392	8,1%
117	629	14,9%	172	1.163	8,8%
118	743	13,7%	16C	858	10,7%
161	847	10,8%	185	967	10,1%
162	932	9,6%	181	1.287	12,7%
163	777	12,4%	182	875	10,7%
164	648	13,1%	183	1.114	12,9%
165	775	14,2%	184	959	13,6%
166	471	14,8%	150	1.189	9,3%

A partir da leitura da tabela 7.4.6, pode verificar-se que a maior parte das estimativas apresenta EPR superiores a 10%. Quando se comparam as estimativas EBLUP espaciotemporais que não garantem a consistência interna (tabela 7.4.4), com as respectivas estimativas EBLUP espaciotemporais que garantem a consistência interna ao nível de NUTSII (tabela 7.4.6), observa-se uma concordância global nas estimativas do preço médio de transacção da habitação, mas verifica-se uma deterioração da qualidade destas últimas, sendo o seu EPR cerca de três vezes superior. É, contudo, de salientar que mesmo verificando-se esta perda de qualidade das estatísticas apresentadas na tabela 7.4.6, elas ainda apresentam níveis de precisão aceitáveis para publicação

é de 0,238%. Estas diferenças negligenciáveis devem-se ao facto de ter sido utilizada a inversa generalizada da matriz **C**, em vez da inversa da matriz **C**, pelo facto desta matriz ser singular.

pelos produtores de estatísticas oficiais, o que as torna bastante apelativas pelo facto de garantirem a tão desejada consistência interna.

8. CONCLUSÃO

8.1 PRINCIPAIS CONCLUSÕES

O trabalho de investigação apresentado nesta tese teve como um dos principais objectivos, o desenvolvimento de uma metodologia de estimação em pequenos domínios, no âmbito de inquéritos por amostragem, com dados de nível área de natureza espacial e/ou cronológica. O outro principal objectivo deste trabalho consistiu em utilizar esses desenvolvimentos metodológicos na resolução de um problema de estimação de índole prático. Mais especificamente, pretendeu (*i*) propor-se uma metodologia de estimação, ao nível de NUTSIII, do preço médio de transacção da habitação em Portugal, com precisão aceitável; (*ii*) garantir-se a consistência aritmética na publicação dessas estimativas com as disponíveis para níveis mais agregados (NUTSII e Portugal continental); (*iii*) propor-se estimadores das componentes de variância do modelo espaciotemporal que assiste a estimação; (*iv*) desenvolver-se estimadores do EQMP *model-based* dos EBLUP temporal e espaciotemporal; e (*v*) avaliar-se a qualidade dos estimadores propostos para estimar o preço médio de transacção da habitação, bem como das medidas de precisão propostas para estimadores combinados sem restrições.

Após uma recensão teórica sobre conceitos fundamentais na estimação em domínios, sobre predição em modelos lineares mistos e sobre estimadores combinados habitualmente utilizados na estimação em pequenos domínios, foi proposta, no capítulo quinto, uma metodologia de estimação em pequenos domínios utilizando um modelo linear misto com dados espaciais e cronológicos. No âmbito deste modelo espaciotemporal, foram efectuados os seguintes trabalhos: (*i*) foi deduzida a expressão explícita do estimador (E)BLUP espaciotemporal; (*ii*) foram propostos estimadores

centrados dos parâmetros de variância, utilizando uma abordagem pelo método dos momentos; e (*iii*) foram propostos três estimadores do EQMP *model-based* do estimador EBLUP espaciotemporal, utilizando a abordagem delta e as abordagens por reamostragem (métodos *jackknife* e *bootstrap*).

O modelo espaciotemporal proposto, para além de incorporar simultaneamente na estimação dados de natureza espacial e cronológica, tira também partido de informação auxiliar relativa a outras variáveis conhecidas na população, bem como de estruturas de covariância espacial e cronológica entre os dados amostrais, mesmo que exógenos aos domínios de interesse. Tendo em conta a generalidade e a flexibilidade do modelo de estimação em domínios proposto, envolvendo efeitos aleatórios de domínio espacialmente dependentes através de um processo SAR e efeitos aleatórios de domínio-tempo temporalmente autocorrelacionados através de um processo AR(1), julga-se possível a sua aplicação num vasto leque de problemas práticos nas áreas da economia, gestão, ambiente, saúde, entre outras. Na maior parte deste tipo de problemas é razoável assumir que os efeitos aleatórios associados a domínios vizinhos estejam associados, assim como efeitos aleatórios, associados a um determinado domínio, referentes a períodos de tempo próximos estejam correlacionados e que essa autocorrelação decaia para zero com o aumento da distância temporal.

Para além disso, uma característica muito apelativa do estimador EBLUP espaciotemporal, assistido pelo modelo proposto, é que é possível evitar que o estimador se reduza a um estimador sintético "puro", mesmo quando a dimensão amostral observada no domínio que é alvo de inferência é nula. Para tal, basta que esse domínio apresente pelo menos uma relação de dependência espacial com outro domínio de dimensão amostral não nula, e/ou que existam observações amostrais referentes ao domínio em pelo menos um dos períodos de tempo anteriores. Esta característica desse estimador espaciotemporal constitui uma grande vantagem em relação aos estimadores combinados revistos no capítulo quarto.

No sentido de se alcançarem os objectivos definidos, foi proposta, no capítulo sexto, uma metodologia geral que permitiu a introdução de restrições em modelos de estimação em pequenos domínios de nível área, considerados como casos particulares do modelo linear misto. A metodologia proposta, inspirada nos princípios da econometria clássica sobre estimação dos parâmetros de uma regressão sob restrições

lineares, teve como objectivo principal garantir a consistência aritmética na publicação das estimativas referentes a pequenos domínios com as estimativas referentes a níveis mais agregados. Complementarmente, através desta metodologia, pretendeu conferir-se aos estimadores uma maior robustez contra possíveis falhas ou más especificações dos modelos que assistem a estimação. Para além de terem sido deduzidos resultados gerais sobre a estimação com restrições, cujos resultados de Ugarte *et al*. (2009) constituem um caso particular, nesta tese foi também deduzida a expressão explícita do preditor com restrições e foram propostas duas metodologias inovadoras de estimação do seu EQMP. O capítulo sexto terminou com a apresentação do modelo de estimação em pequenos domínios com dados espaciais e cronológicos, considerando restrições na estimação, de forma a garantir a consistência interna na publicação das estimativas.

No âmbito da metodologia proposta de estimação com restrições, foi verificado que o preditor com restrições é dado pela soma de duas parcelas: (*i*) o preditor sem restrições e (*ii*) um factor de correcção definido como uma proporção da diferença entre as estimativas do parâmetro de interesse ao nível de uma dada região e a média ponderada das estimativas produzidas para os pequenos domínios dessa região. Neste contexto, foi demonstrado que esse preditor com restrições é BLUP. Para além disso, foi ainda verificado que cada uma das duas componentes do EQMP do BLUP com restrições é igual à respectiva componente do EQMP do BLUP sem restrições, mais um acréscimo devido à variabilidade gerada pela introdução das restrições na estimação.

Nesta tese foram ainda propostos dois estimadores do EQMP *model-based* do estimador EBLUP temporal sem restrições, utilizando abordagens por reamostragem (métodos *jackknife* e *bootstrap*). Apesar destes desenvolvimentos não terem feito parte dos objectivos centrais da tese, eles revelaram-se importantes pelo facto de terem permitido avaliar o desempenho deste tipo de estimadores do EQMP, face a um estimador analítico estabelecido na literatura (Rao e Yu, 1994).

O último grande objectivo desta tese consistiu na avaliação da qualidade: (*i*) dos estimadores propostos para estimar o preço médio de transacção da habitação, relativamente a diversos outros estimadores directos e indirectos habitualmente utilizados na estimação desse tipo de parâmetros de interesse em pequenos domínios; e (*ii*) dos estimadores do EQMP propostos para medir a incerteza dos estimadores combinados de parâmetros de interesse sem restrições. Essa avaliação foi efectuada com

base em dois estudos empíricos por simulação do tipo *design-based* e *model-based*, respectivamente.

No estudo empírico do tipo *design-based*, no qual foram avaliadas as propriedades dos estimadores dos parâmetros de interesse sob uma perspectiva de amostragem repetida, foram utilizados dados reais de painel, fornecidos pelo IPTH e pelo IABH, ambos da responsabilidade do INE. Os resultados obtidos neste estudo empírico permitem extrair um conjunto de conclusões, sendo os principais resultados e as respectivas ilações enumerados de seguida.

1. O estimador directo é aproximadamente centrado, embora seja um estimador pouco preciso devido à sua elevada variância.

2. Os estimadores sintéticos apresentam enviesamentos severos, embora apresentem variâncias muito pequenas. Os elevados enviesamentos destes estimadores parecem indicar que os pressupostos subjacentes a este tipo de estimação não se verificam no contexto da estimação do preço médio de transacção da habitação. Este facto reforçou a convicção da necessidade de utilização de estimadores assistidos por modelos que envolvessem efeitos aleatórios específicos de domínio, de forma a acomodar a variabilidade existente entre as NUTSIII em termos de preços de transacção da habitação.

3. O estimador combinado com pesos dependentes dos dados apresenta ganhos de enviesamento relativamente ao estimador sintético pela regressão e ganhos de variância relativamente ao estimador directo. Para além disso, este estimador combinado consegue ainda apresentar alguns ganhos de precisão relativamente àquele estimador sintético.

4. Os quatro estimadores EBLUP sem restrições são aqueles que apresentam globalmente melhores propriedades, uma vez que são os mais precisos e só são mais enviesados do que o estimador directo. Estes resultados parecem indicar que este tipo de combinação entre uma componente sintética e uma componente directa, consegue reduzir uma parte significativa do enviesamento do estimador sintético puro, trocando-a por um pequeno acréscimo de variância, sendo esta troca favorável do ponto de vista da precisão dos estimadores EBLUP. Estes resultados parecem também confirmar que os estimadores EBLUP, assistidos por modelos que envolvem

efeitos aleatórios, são mais adequados do que os estimadores sintéticos, quando existem discrepâncias significativas no comportamento da variável de interesse nos diferentes pequenos domínios.

5. Os estimadores EBLUP que tiram partido da associação espacial entre as observações, tendem a apresentar melhores propriedades do que os estimadores que não utilizam tal informação, sobretudo ao nível do enviesamento. Este facto indica que a consideração de associação espacial é uma excelente via de obtenção de "informação emprestada" tendo em vista a redução do enviesamento.

6. Os estimadores EBLUP que tiram partido da autocorrelação cronológica entre as observações, tendem a apresentar melhores propriedades do que os estimadores que não utilizam tal informação, sendo os ganhos mais acentuados ao nível do enviesamento. Desta forma, parece existir evidência que a consideração de autocorrelação cronológica é também uma excelente via de obtenção de "informação emprestada" tendo em vista a melhoria das propriedades dos estimadores.

7. A introdução de informação exclusivamente temporal na estimação melhora mais as propriedades dos estimadores do que a introdução de informação unicamente espacial, comparativamente a um estimador que não utiliza tal tipo de informação. Contudo, a melhor alternativa consiste em utilizar simultaneamente esses dois tipos de informação. De facto, o estimador EBLUP espaciotemporal revelou-se o mais eficiente e um estimador aproximadamente centrado.

8. Os estimadores EBLUP apresentam taxas médias de cobertura dos IC *model-based* mais próximas do nível de confiança definido e menos dependentes das dimensões amostrais dos domínios, do que as taxas médias de cobertura dos IC *design-based*. Estes factos parecem confirmar a boa qualidade dos estimadores do EQMP dos EBLUP, também sob uma perspectiva de amostragem repetida, pelo que os IC *model-based* podem constituir uma excelente alternativa aos IC *design-based*, sob essa perspectiva de estimação.

9. A garantia da consistência interna na publicação das estimativas conduz ao aumento do enviesamento e à perda de precisão dos estimadores EBLUP com restrições, comparativamente a um estimador que não verifica essa consistência. Porém, as perdas de precisão desses estimadores EBLUP são geralmente pouco significativas,

desde que a estimativa directa do parâmetro de interesse para o nível de agregação superior tenha uma boa precisão. Este facto parece confirmar que a garantia da consistência interna não confere robustez ao estimador ajustado, devido ao facto do modelo que assiste a estimação se encontrar provavelmente bem especificado, descrevendo adequadamente a população finita.

10. A garantia da consistência interna ao nível de Portugal continental tem um efeito mais significativo sobre as propriedades do estimador EBLUP que garante essa consistência, do que sobre as propriedades do estimador EBLUP que garante a referida consistência apenas ao nível de NUTSII. No entanto, se a consistência interna for garantida a um nível geográfico pouco agregado (NUTSII), então o efeito sobre a variância do estimador é quase insignificante.

Perante estes resultados e estas ilações, parece ficar claro que a utilização de informação auxiliar pode melhorar significativamente as propriedades dos estimadores de parâmetros de interesse em pequenos domínios. Em particular, o modelo de estimação em pequenos domínios proposto parece fornecer um bom enquadramento para a consideração de tal informação no processo de estimação, nomeadamente através da utilização de estruturas de covariância dos efeitos aleatórios auto-regressivas.

Face ao exposto, recomenda-se que a estimação do preço médio de transacção da habitação em Portugal seja efectuada através da utilização da metodologia proposta nesta tese. Se não for exigida a garantia da consistência interna na publicação das estimativas do preço médio de transacção da habitação ao nível de NUTSIII, então recomenda-se a utilização do estimador EBLUP espaciotemporal sem restrições e a avaliação da sua precisão através do estimador analítico do EQMP, ambos propostos no quinto capítulo. Neste caso, o erro padrão relativo médio das estimativas é de 5,8%. Pelo contrário, se for exigida essa garantia da consistência interna, então recomenda-se a utilização do estimador EBLUP espaciotemporal com restrições, derivado a partir da metodologia de estimação com restrições proposta no sexto capítulo. Neste caso, deve ser exigido que a estimativa directa do parâmetro de interesse ao nível de cada NUTSII seja igual à soma ponderada das estimativas indirectas desse parâmetro de interesse das NUTSIII que fazem parte da referida NUTSII, tendo como resultado um erro padrão relativo médio das estimativas de 10,6%.

Desta forma, é possível construir-se um índice de preços de transacção da habitação em Portugal, ao nível de NUTSIII, baseado nas estimativas do preço médio de transacção da habitação com boa precisão, produzidas através da metodologia proposta nesta tese.

No que se refere ao estudo empírico do tipo *model-based*, no qual foram avaliadas as propriedades dos estimadores do EQMP propostos para medir a incerteza dos estimadores combinados de parâmetros de interesse sem restrições, os resultados permitem extrair as seguintes conclusões:

1. Não existe um único estimador do EQMP, quer do EBLUP temporal quer do EBLUP espaciotemporal, que seja uniformemente melhor do que os restantes, em termos de enviesamento e de precisão.

2. Os estimadores por reamostragem apresentam um desempenho muito bom quando comparados com o respectivo estimador analítico (temporal ou espaciotemporal), na medição da incerteza associada às estimativas dos EBLUP. De facto, verificou-se que os estimadores *bootstrap* e *jackknife* chegaram a apresentar um desempenho melhor do que o respectivo estimador analítico em muitos casos.

3. Os estimadores por reamostragem apresentaram um comportamento muito semelhante entre si, com uma pequena vantagem, sobretudo em termos de enviesamento relativo absoluto médio, para os estimadores do tipo *jackknife*. Contudo, a sua tendência para subestimador ou sobreestimar os verdadeiros valores do EQMP de forma mais acentuada em alguns domínios, aliada à sua mais trabalhosa aplicação prática, favorecem os estimadores do tipo *bootstrap*.

Desta forma, conclui-se que existe evidência empírica que permite considerar como adequado, o uso de estimadores baseados em métodos por reamostragem, e em particular o estimador *bootstrap*, na estimação da incerteza associada aos EBLUP, como alternativa aos estimadores baseados em longos desenvolvimentos analíticos. Em particular, a utilização deste tipo de estimadores baseados em métodos por reamostragem é promissora no contexto de modelos longitudinais de estimação em pequenos domínios mais complexos, para os quais é impossível deduzir aproximações analíticas para o EQMP do EBLUP.

8.2 LIMITAÇÕES DO ESTUDO

Embora todos os objectivos deste trabalho de investigação tenham sido atingidos, tendo-se proposto uma metodologia de estimação em pequenos domínios que permite estimar o preço médio de transacção da habitação em Portugal com precisão aceitável, este estudo levanta preocupações de natureza metodológica que merecem ser investigadas. Algumas dessas preocupações são enumeradas de seguida.

Uma das limitações deste estudo prende-se com o facto de ter sido utilizado apenas um tipo de forma funcional nas estruturas de covariância espacial e cronológica dos efeitos aleatórios do modelo. Apesar de ser razoável assumir que os preços de transacção da habitação associados a um determinado domínio apresentam uma correlação cronológica decrescente ao longo do tempo, e que os preços de transacção da habitação associados a domínios vizinhos estão correlacionados de forma simultânea, independentemente da distância entre os domínios, existem muitas outras estruturas de covariância que também poderiam ser plausíveis. De facto, num contexto onde as diferenças entre o litoral e o interior, bem como entre o norte e o sul, podem ser significativas, o recurso a estruturas de covariância espacial isotrópicas ou anisotrópicas (Littell *et al*., 2006), poderá permitir uma melhor descrição da população finita. Por outro lado, a consideração de uma estrutura de covariância cronológica auto-regressiva heterogénea também poderá permitir uma melhor representação dessa realidade (Wolfinger, 1996). É, ainda, de salientar que até mesmo a estrutura espacial do tipo CAR poderá ser uma opção válida neste contexto. Contudo, o elevado tempo computacional exigido pela utilização da generalidade desses tipos de estruturas de covariância na presença de dados provenientes de sete vagas do inquérito e a necessidade de estimação de um maior número de componentes de variância na maior parte dessas estruturas, tornaram inviável ou desaconselhada a investigação do desempenho de estimadores assistidos por modelos com esses tipos de estruturas de covariância. Naturalmente que num contexto em que a velocidade de processamento computacional é cada vez maior, parte desta limitação tenderá a diminuir num futuro próximo. Permanece, contudo, o problema dessas estruturas de covariância serem pouco parcimoniosas, introduzindo por esta via variabilidade adicional na estimação das componentes de variância, a qual se reflecte na qualidade das estimativas dos

parâmetros de interesse em pequenos domínios, designadamente através da terceira componente do EQMP dos EBLUP.

Por outro lado, não existe uma única forma de definir a matriz de pesos espaciais, podendo os resultados ser sensíveis à escolha dessa matriz. Apesar de se ter utilizado um tipo de matriz de pesos normalmente utilizado no âmbito da estimação em pequenos domínios, será muito útil realizarem-se estudos para investigar o desempenho do estimador EBLUP espacial/espaciotemporal com matrizes de pesos espaciais mais complexas, por exemplo, como uma função das distâncias entre os pequenos domínios (Cliff e Ord, 1981; Getis e Aldstadt, 2004).

Outra limitação deste estudo prende-se com o facto de se ter admitido, no âmbito do modelo de estimação em pequenos domínios proposto, que os parâmetros de autocorrelação temporal e de associação espacial são conhecidos. Como se sabe, todos os parâmetros são geralmente desconhecidos nas aplicações práticas com dados reais, e estes não são uma excepção. Note-se que foi admitido que estes parâmetros são conhecidos devido à dificuldade em estimá-los, pelo método dos momentos, de forma simultaneamente consistente e admissível (Yu, 1993; Rao, 2003), *i.e.*, assumindo valores no intervalo [-1; 1] no contexto do modelo de Rao-Yu, bem como no contexto do modelo proposto. Dificuldades semelhantes foram identificadas por Fuller (1987) sob modelos de erros de medida. Por exemplo, será interessante explorar modificações adequadas ao estimador consistente deduzido pelo método dos momentos proposto por Rao e Yu (1994), de forma a tomar valores no intervalo admissível [-1; 1]. Acredita-se que este problema de investigação possa constituir, ele próprio, matéria para uma investigação autónoma.

Neste trabalho de investigação foi proposto um estimador EBLUP espaciotemporal para estimação de parâmetros em pequenos domínios, assistido por um modelo linear misto. Embora se tenha avaliado a qualidade desse estimador face a outros estimadores EBLUP estabelecidos na literatura, que utilizam informação seccional, espacial ou temporal, não se pode deixar de referir que o estimador proposto não foi comparado com o estimador EBLUP espaciotemporal assistido por um modelo do tipo *state space*, apresentado no subcapítulo 4.6. Esta situação pode ser considerada outra limitação deste trabalho de investigação. Todavia, nesta tese decidiu investigar-se metodologias de estimação em pequenos domínios assistidas por modelos suficientemente flexíveis na

descrição de diferentes tipos de realidades, como é o caso do modelo linear misto proposto, em oposição a modelos que exigem a especificação de uma equação de transição rígida para especificar completamente o modelo. Naturalmente que será interessante estudar-se no futuro o desempenho do estimador proposto face a esse outro estimador espaciotemporal.

Outra possível limitação deste estudo deve-se ao facto da estimação ter sido efectuada ao nível de NUTSIII e não ao nível de concelho, ou pelo menos ao nível de concelho nas áreas metropolitanas de Lisboa e do Porto, e ao nível de NUTSIII para o restante continente. Apesar da metodologia proposta permitir a estimação ao nível de concelho, tal não foi efectuada no estudo empírico *design-based* devido ao elevado tempo computacional exigido para a sua concretização. Contudo, se um produtor de estatísticas oficiais decidir fazê-lo, tal será possível, não só porque a metodologia o permite, mas sobretudo porque os recursos informáticos não serão seguramente um entrave.

Nesta tese foi proposta uma metodologia que permite a introdução de restrições em modelos de estimação em pequenos domínios. Apesar de se ter avaliado o desempenho dos estimadores EBLUP modificados pelas restrições no âmbito do estudo por simulação *design-based*, não fez parte dos objectivos deste trabalho de investigação avaliar a qualidade dos estimadores por reamostragem do EQMP do EBLUP com restrições. Esta limitação da tese dá origem a outro caminho de investigação futura.

Por último, no âmbito do estudo empírico por simulação *model-based*, decidiu avaliar-se a qualidade dos estimadores do EQMP dos EBLUP tendo em consideração apenas valores positivos para os coeficientes de associação espacial e autocorrelação temporal. De facto, a escolha de valores positivos para aqueles coeficientes foi efectuada de forma intencional, tendo em conta os objectivos práticos deste estudo, os resultados obtidos na análise exploratória de dados dos preços de transacção da habitação e o tempo computacional necessário para a realização de um estudo deste tipo. Porém, a consideração de valores negativos ou nulos nos referidos coeficientes deve ser objecto de investigação, uma vez que tal pode ocorrer no contexto de outros problemas reais.

8.3 DESENVOLVIMENTOS FUTUROS

Um trabalho de investigação nunca está terminado. Efectivamente, durante o seu desenvolvimento, e antes mesmo de se aproximar e antever a sua conclusão, é comum os investigadores identificarem áreas que gostariam de investigar no futuro, pontos que ficaram menos explorados ou assuntos que não conseguiram resolver, muitas vezes por não fazerem parte dos objectivos do trabalho, outras vezes por falta de tempo efectivo durante a elaboração dos seus trabalhos. Este trabalho de investigação não foge à regra. De facto, são muitos os problemas sobre os quais se pretende investigar no futuro, alguns deles identificados como possíveis limitações deste estudo.

Em primeiro lugar, pretende investigar-se no futuro imediato o desempenho do estimador EBLUP espacial e espaciotemporal, assistido por um modelo que envolve matrizes de pesos espaciais como função das distâncias entre os pequenos domínios.

Em segundo lugar, pretende estudar-se o desempenho do modelo proposto utilizando outros tipos de estruturas de covariância parcimoniosas. Em particular, pretende testar-se uma estrutura de covariância do tipo CAR, bem como uma estrutura de covariância espacial exponencial, tendo em conta as distâncias entre os pequenos domínios (isotrópica). A dedução de um estimador analítico do EQMP do EBLUP, sob esse tipo de estruturas de covariância, constitui outro problema não trivial que se pretende investigar no futuro.

Em terceiro lugar, pretende avaliar-se a robustez das metodologias de reamostragem para estimação do EQMP do EBLUP temporal e espaciotemporal, sob situações de não normalidade dos erros do modelo. Um caminho a seguir poderá passar por realizar um estudo por simulação *model-based* em que os erros do modelo são gerados através de distribuições diferentes da normal, como por exemplo a *t*-student (que tem caudas mais largas do que a normal) ou a distribuição gumbel (que é assimétrica), as quais já foram utilizadas num estudo semelhante no âmbito do modelo espacial de Salvati.

Em quarto lugar, e ainda no que se refere às referidas metodologias de reamostragem para estimação do EQMP do EBLUP espaciotemporal, pretende testar-se o seu desempenho na situação em que os parâmetros de autocorrelação temporal e de associação espacial são desconhecidos. Neste caso, deverá utilizar-se um método de

verosimilhança na estimação das componentes de variância, uma vez que não existem estimadores, pelo método dos momentos, que permitam a estimação dos parâmetros de autocorrelação temporal e de associação espacial de forma simultaneamente consistente e admissível.

Por último, pretende avaliar-se o desempenho dos estimadores do EQMP do EBLUP com restrições, através de um estudo *model-based*. Planeia-se que este estudo de avaliação, o qual constitui um interessante tema para a comunidade científica, seja realizado num futuro próximo, utilizando um procedimento semelhante ao que foi seguido na avaliação da qualidade dos estimadores do EQMP dos EBLUP sem restrições.

De facto, são muitas as linhas por onde se pode desenrolar uma investigação futura neste campo da estimação em pequenos domínios…

BIBLIOGRAFIA

ALI, M. M. (1979) Analysis of stationary spatial-temporal processes: Estimation and prediction. *Biometrika,* 66**,** 513-518.

ANDERSON, T. W. (1973) Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure. *The Annals of Statistics,* 1**,** 135-141.

ANSELIN, L. (1992) *Spatial Econometrics: Methods and Models.,* Boston, Kluwer Academic Publishers.

ANSLEY, C. F. & KOHN, R. (1986) Prediction mean squared error for state space models with estimated parameters. *Biometrika,* 73**,** 467-473.

ARORA, V. & LAHIRI, P. (1997) On the superiority of the bayes method over the BLUP in small area estimation problem. *Statistica Sinica,* 7**,** 1053-1063.

ARORA, V., LAHIRI, P. & MUKHERJEE, K. (1997) Empirical Bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association,* 92**,** 1555-1562.

ASSEMBLEIA DA REPÚBLICA (1989) Lei n.º 6/89 de 15 de Abril - Sistema Estatístico Nacional. Lisboa, Diário da República-I Série.

ASSEMBLEIA DA REPÚBLICA (2002) Decreto-Lei n.º 244/2002 de 5 de Novembro. Lisboa, Diário da República-I Série-A.

BAILAY, T. C. & GATRELL, A. C. (1995) *Interaction Spatial Data Analysis,* London, Longman.

BAILEY, M. J., MUTH, R. F. & NOURSE, H. O. (1963) A Regression Method for Real Estate Price Construction. *Journal of the American Statistical Association,* 58**,** 933-942.

BALTAGI, B. (2005) *Econometric Analysis of Panel Data.,* England, John Wiley &

Sons.

BALTAGI, B. H., SONG, S. H., JUNG, B. C. & KOH, W. (2007) Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics,* 140**,** 5-51.

BANERJEE, S., CARLIN, B. & GELFAND, A. (2004) *Hierarchical Modelling and Analysis for Spatial Data,* New York, Chapman & Hall.

BARNDORFF-NIELSEN, O. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika,* 70**,** 343-365.

BATTESE, G. E., HARTER, R. M. & FULLER, W. A. (1988) An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association,* 83**,** 28-36.

BELL, J. (2002) *Como realizar um projecto de investigação,* Lisboa, Gradiva.

BELL, W. (2001) Discussion with "Jackknife in the Fay-Herriot model with an example". *Proceedings of the Seminar on Funding Opportunity in Survey Research.* Federal Committee on Statistical Methodology.

BESAG, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society: Series B,* 36**,** 192-136.

BESAG, J. & KOOPERBERG, C. (1995) On conditional and intrinsic autoregression. *Biometrika,* 82**,** 733-46.

BINDER, D. (1998) Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodology,* 24**,** 101-108.

BOOTH, J. G. & HOBERT, J. P. (1998) Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association,* 93 262-272.

BROWN, K. G. (1976) Asymptotic behavior of MINQUE-type estimators of variance components. *The Annals of Statistics,* 4**,** 746-754.

BURNHAM, K. P. & ANDERSON, D. R. (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach,* New York, Springer.

BURRIDGE, P. (1980) On the Cliff-Ord test for spatial correlation. *Journal of the Royal Statistical Society: Series B,* 107-108.

BUTAR, F. B. & LAHIRI, P. (2003) On measures of uncertainty of empirical Bayes small area estimators. *Journal of Statistical Planning and Inference,* 112**,** 63-76.

CARVALHO, M. L. & NATÁRIO, I. C. (2008) *Análise de Dados Espaciais,* Lisboa, Sociedade Portuguesa de Estatística.

CHANDRA, H. & CHAMBERS, R. (2006*a*) Improved direct estimators for small areas. *Working Paper M06/07*. Southampton, United Kingdom, Southampton Statistical Sciences Research Institute.

CHANDRA, H. & CHAMBERS, R. (2006*b*) Multipurpose small area estimation. *Working Paper M06/06.* Southampton, United Kingdom, Southampton Statistical Sciences Research Institute.

CHANDRA, H. & CHAMBERS, R. (2006*c*) Small area estimation with skewed data. *Working Paper M06/05.* Southampton, United Kingdom, Southampton Statistical Sciences Research Institute.

CHANDRA, H., SALVATI, N. & CHAMBERS, R. (2007*a*) Small area estimation for spatially correlated populations - a comparison of direct and indirect model-based methods. *Working Paper M07/09.* Southampton, United Kingdom, Southampton Statistical Sciences Research Institute.

CHANDRA, H., SALVATI, N. & CHAMBERS, R. (2007*b*) Small area estimation for spatially correlated populations - a comparison of direct and indirect model-based methods. *Statistics in Transition - new series,* 8**,** 331-350.

CHATTERJEE, S., LAHIRI, P. & LI, H. (2008) Parametric Bootstrap approximation to the distribution of EBLUP and related prediction intervals in liner mixed models. *The Annals of Statistics,* 36**,** 1221-1245.

CHEN, S. & LAHIRI, P. (2002) A weighted jackknife MSPE estimator in small area estimation. *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

CHEN, S. & LAHIRI, P. (2003) A comparison of different MSPE estimators of EBLUP for the Fay-Herriot model. *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

CHEN, S. & LAHIRI, P. (2005) On mean squared prediction error estimation in small area estimation problems. *Proceedings of the Section on Survey Research*

Methods. American Statistical Association.

CHEN, S. & LAHIRI, P. (2008) On mean squared prediction error estimation in small area estimation problems. *Communications in Statistics-Theory and Methods,* 37**,** 1792-1798.

CHOLETTE, P. A. & DAGUM, E. B. (1994) Benchmarking Time Series with Autocorrelated Survey Errors. *International Statistical Review,* 62**,** 365-377.

CHOUDHRY, G. H. & RAO, J. N. K. (1989) Small area estimation using models that combine time series and cross-sectional data. *Proceedings of the Statistics Canada, Symposium on Analysis of Data in Time.* Statistics Canada.

CLAYTON, D. G. & BERNARDINELLI, L. (1996) Bayesian methods for mapping disease risks. IN ELLIOT, P., CUZICK, J., ENGLISH, D. & STERN, R. (Eds.) *Geographical and Environment Epidemiology: Methods for Small-Area Studies.* Oxford, Oxford University Press.

CLIFF, A. D. & ORD, J. K. (1973) *Spatial autocorrelation,* London, Pion Press.

CLIFF, A. D. & ORD, J. K. (1975) Space-time modeling with an application to regional forecasting. *Transactions of the Institute of British Geographers,* 66**,** 119-128.

CLIFF, A. D. & ORD, J. K. (1981) *Spatial Processes: Models and Applications,* London, Pion Press.

COCHRAN, W. (1977) *Sampling techniques,* New York, John Wiley & Sons.

COELHO, P. S. (2000) Estimação em domínios sob o modelo linear geral misto com informação cronológica e espacial. Tese de Doutoramento, Lisboa, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.

CONSIGLIO, L. D., FALORSI, P. D., FALORSI, S. & RUSSO, A. (2003) Conditional and unconditional analysis of some small area estimators in complex sampling. *Survey Methodology,* 29**,** 53-61.

COOPER, D. M. & THOMPSON, R. (1977) A Note on the Estimation of the Parameters of the Autoregressive-Moving Average Process. *Biometrika,* 64**,** 625-628.

CRESSIE, N. (1989) Empirical bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association,* 84**,** 1033-1044.

CRESSIE, N. (1991) Small-area prediction of undercount using the general linear

model. *Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality.* Statistics Canada.

CRESSIE, N. (1992) REML estimation in empirical bayes smoothing of census undercount. *Survey Methodology,* 18**,** 75-94.

CRESSIE, N. (1993) *Statistics for Spatial Data,* New York, John Wiley & Sons.

DAS, K., JIANG, J. & RAO, J. N. K. (2004) Mean squared error of empirical predictor. *The Annals of Statistics,* 32**,** 818-840.

DATTA, G. S., DAY, B. & BASAWA, I. (1999) Empirical best linear unbiased and empirical bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference,* 75**,** 269-279.

DATTA, G. S. & GHOSH, M. (1991) Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics,* 19**,** 1748-1770.

DATTA, G. S. & LAHIRI, P. (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica,* 10**,** 613-627.

DATTA, G. S., LAHIRI, P. & MAITI, T. (1997) Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data *Technical Report n.º 97-19.* Athens, Department of Statistics, University of Georgia.

DATTA, G. S., LAHIRI, P. & MAITI, T. (2002) Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference,* 102**,** 83-97.

DATTA, G. S., RAO, J. N. K. & SMITH, D. D. (2005) On measuring the variability of small area estimators under a basic area level model. *Biometrika,* 92**,** 183-196.

DAVISON, A. C. & HINKLEY, D. V. (1997) *Bootstrap methods and their application,* Cambridge, Cambridge University Press.

DE LUNA, X. & GENTON, M. G. (2002) Simulation-based Inference for Simultaneous Processes on Regular Lattices. *Statistics and Computing,* 12**,** 125-134.

DE LUNA, X. & GENTON, M. G. (2005) Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica,* 15**,** 547-568.

DEMIDENKO, E. (2004) *Mixed Models: Theory and Applications,* New York, Wiley.

DIEWERT, E. (2006) Conclusions and Future Directions. *OCDE-IMF Workshop on Real Estate Price Indexes.* Paris, OCDE.

DIGGLE, P. J. (1988) An approach to the analysis of repeated measurements. *Biometrics,* 44**,** 959-971.

DIGGLE, P. J., LIANG, K.-Y. & ZEGER, S. L. (1996) *Analysis of Longitudinal Data,* Oxford, Oxford Universitiy Press.

DREW, D., SINGH, M. P. & CHOUDHRY, G. H. (1982) Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology,* 8**,** 17-47.

EFRON, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics,* 7**,** 1-26.

EFRON, B. & TIBSHIRANI, R. (1993) *An introduction to the bootstrap,* New York, Chapman & Hall.

EFRON, N. (1993) *Statistics for Spatial Data,* New York, John Wiley & Sons.

FABRIZI, E., FERRANTE, M. R. & PACEI, S. (2007) Small area estimation of average houshold income based on unit level models for panel data. *Survey Methodology,* 33**,** 187-198.

FAI, A. H. T. & CORNELIUS, P. L. (1996) Approximate F-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-plot Experiments. *Journal of Statistical Computation and Simulation,* 54**,** 363-378.

FALORSI, P. D., FALORSI, S. & RUSSO, A. (1994) Empirical comparison of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology,* 20**,** 171-176.

FALORSI, P. D., FALORSI, S. & RUSSO, A. (1999) Small area estimation at provincial level in the Italian Labour Force Survey. *Journal of the Italian Statistical Society,* 1**,** 93-109.

FAY, R. E. (1987) Application of multivariate regression to small domain estimation. IN PLATEK, R., RAO, J. N. K., SÄRNDAL, C.-E. & SINGH, M. P. (Eds.) *Small Area Statistics: An international symposium.* New York, John Wiley & Sons.

FAY, R. E. & HERRIOT, R. A. (1979) Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74**,** 269-277.

FISHER, R. A. (1992) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A,* 222**,** 309-368.

FISHMAN, G. S. (1996) *Monte Carlo: Concepts, algorithms, and applications,* New York, Springer-Verlag.

FRADA, J. J. C. (2001) *Guia prático para a elaboração e apresentação de trabalhos científicos,* Lisboa, Microcosmos.

FULLER, W. A. (1987) *Measurement Error Models,* New York, John Wiley & Sons.

FULLER, W. A. & BATTESE, G. E. (1973) Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association,* 68**,** 626-632.

FULLER, W. A. & HARTER, R. M. (1987) The multivariate components of variance model for small area estimation. IN PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E. & SINGH, M. P. (Eds.) *Small Area Statistics.* New York, John Wiley & Sons.

GEARY, R. C. (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician,* 5**,** 115-145.

GETIS, A. & ALDSTADT, J. (2004) Constructing the spatial weights matrix using a local statistic. *Geographical Analysis,* 32**,** 90-104.

GHOSH, M. & NANGIA, N. (1993) Estimation of median income of four-person families: A bayesian time series approach. *Technical Report.* Gainsville, Department of Statistics, University of Florida.

GHOSH, M., NANGIA, N. & KIM, D. (1996) Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association,* 91**,** 1423-1431.

GHOSH, M. & RAO, J. N. K. (1994) Small area estimation: An appraisal. *Statistical Science,* 9**,** 55-93.

GIESBRECHT, F. G. & BURNS, J. C. (1985) Two-Stage Analysis Based on a Mixed Model: Large-sample Asymptotic Theory and Small-Sample Simulation Results.

Biometrics, 41**,** 477-486.

GIROUARD, N. & BLONDAL, S. (2001) House prices and economic activity. *OECD Economics Department Working Papers n.º 279.* Paris, OCDE.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. & SANTAMARÍA, L. (2005) Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Working Paper 05-49.* Madrid, Departamento de Estadística, Universidad Carlos III de Madrid.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. & SANTAMARÍA, L. (2008) Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation,* 78**,** 443-462.

GONZALEZ, M. E. (1973) Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section.* American Statistical Association.

GREENE, W. (2003) *Econometric Analysis,* New Jersey, USA, Prentice Hall.

GUYON, P. (1982) Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika,* 69**,** 95-105.

HAINING, R. (1990) *Spatial Data Analysis in the Social and Environmental Sciences,* Cambridge, Cambridge University Press.

HALL, P. (1992) *The Bootstrap and Edgeworth Expansion,* New York, Springer Verlag.

HALL, P. & MAITI, T. (2006*a*) Nonparametric estimation of mean squared prediction error in nested-error regression models. *The Annals of Statistics,* 34**,** 1733-1750.

HALL, P. & MAITI, T. (2006*b*) On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B,* 68**,** 221-238.

HAMADA, M. & SITTER, R. (2004) Statistical research: some advice for beginners. *The American Statistician,* 58**,** 93-101.

HANSEN, M. H., HURWITZ, W. N. & MADOW, W. G. (1953) *Sample Survey Methods and Theory,* New York, John Wiley & Sons.

HARTLEY, H. O. & RAO, J. N. K. (1967) Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika,* 54**,** 93-108.

HARVEY, A. C. (1989) *Forecasting, structural time series models and the kalman*

filter, Cambridge, Cambridge University Press.

HARVILLE, D. A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika,* 61**,** 383-385.

HARVILLE, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association,* 72**,** 320-340.

HARVILLE, D. A. (1985) Decomposition of prediction error. *Journal of the American Statistical Association,* 80**,** 132-138.

HARVILLE, D. A. (1990) BLUP (best linear unbiased prediction) and beyond. IN GIANOLA., D. & HAMMOND, K. (Eds.) *Advances in Statistical Methods for Genetic Improvement of Livestock.* New York, Springer-Verlag.

HARVILLE, D. A. (1991) Comment on Robinson-That BLUP is a good thing: The estimation of random effects. *Statistical Science,* 6**,** 35-39.

HARVILLE, D. A. & JESKE, D. R. (1992) Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association,* 87**,** 724-731.

HENDERSON, C. R. (1953) Estimation of variance and covariance components. *Biometrics,* 9**,** 226-252.

HENDERSON, C. R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics,* 31**,** 423-447.

HEYDE, C. C. (1994) A quasi-likelihood approach to the REML estimating equations. *Statistics & Probability Letters,* 21**,** 381-384.

HEYDE, C. C. (1997) *Quasi-likelihood and its application,* New York, Springer-Verlag.

HILLMER, A. F. & TRABELSI, A. (1987) Benchmarking of Economic Time Series. *Journal of the American Statistical Association,* 82**,** 1064-1071.

HOCKING, R. R. & KUTNER, M. H. (1975) Some analytical and numerical comparisons of estimators for the mixed A.O.V. model. *Biometrics,* 31**,** 19-28.

HORVITZ, D. G. & THOMPSON, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47**,** 663-685.

HURVICH, C. M. & TSAI, C. L. (1989) Regression and time series model selection in small samples. *Biometrika,* 79**,** 297-307.

INE (1998) *Tipologia de Áreas Urbanas,* Lisboa, Instituto Nacional de Estatística.

INE (2001*a*) Inquérito aos Preços de Avaliação Bancária na Habitação: Metodologia – Acção 5.3. Porto, Direcção Regional do Norte, Instituto Nacional de Estatística (documento não publicado).

INE (2001*b*) Inquérito aos Preços de Transacção na Habitação: Metodologia – Acção 5.1. Porto, Direcção Regional do Norte, Instituto Nacional de Estatística (documento não publicado).

ISAKI, C. T., TSAY, J. H. & FULLER, W. A. (2000) Estimation of census adjusted factors. *Survey Methodology,* 26**,** 31-42.

JENNRICH, R. & SCHLUCHTER, M. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics,* 42**,** 805-820.

JIANG, J. (1996) REML estimation: Asymptotic behaviour and related topics. *The Annals of Statistics,* 24**,** 255-286.

JIANG, J. (1997*a*) A derivation of BLUP - best linear unbiased predictor. *Statistics & Probability Letters,* 25**,** 321-324.

JIANG, J. (1997*b*) Wald consistency and the model of sieves in REML estimation. *The Annals of Statistics,* 25**,** 1781-1803.

JIANG, J. (1998) Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association,* 93**,** 720-729.

JIANG, J. (2005) Partially observed information and inference about non-Gaussian mixed linear models. *The Annals of Statistics,* 33**,** 2695-2731.

JIANG, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications,* New York, Springer.

JIANG, J. & LAHIRI, P. (2006) Mixed Model Prediction and Small Area Estimation. *Test,* 15**,** 1-96.

JIANG, J., LAHIRI, P. & WAN, S.-M. (1999) Jackknifing the mean squared error of empirical best predictor. *Proceedings of the 52th session of the International Statistical Institute.* Helsinki, International Statistical Institute [disponível em http://www.stat.fi/isi99/proceedings/arkisto/invited/47.html, acedido em

30/05/2007].

JIANG, J., LAHIRI, P. & WAN, S.-M. (2002) A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics,* 30**,** 1782-1810.

JIANG, J., LAHIRI, P., WAN, S.-M. & WU, C.-H. (2001) Jackknifing in the Fay-Herriot model with an example. *Technical Report.* Lincoln, Department of Mathematics and Statistics, University of Nebraska.

JUDGE, G. G., GRIFFITHS, W. E., HILL, R. C., LÜTKEPOHL, H. & LEE, T.-C. (1985) *The Theory and Practice of Econometrics,* New York, John Wiley & Sons.

KACKAR, R. N. & HARVILLE, D. A. (1981) Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-Theory and Methods,* 10**,** 1249-1261.

KACKAR, R. N. & HARVILLE, D. A. (1984) Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association,* 79**,** 853-862.

KENWARD, M. G. & ROGER, J. H. (1997) Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics,* 53**,** 983-997.

KHURI, A. I. & SAHAI, H. (1985) Variance components analysis: a selective literature survey. *International Statistical Review,* 53**,** 279-300.

KISH, L. (1995) *Survey Sampling,* New York, John Wiley & Sons.

KNOTTNERUS, P. (2003) *Sample survey theory - Some pythagorean perspective,* New York, Springer-Verlag.

LAHIRI, P. (2003) On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation. *Statistical Science,* 18**,** 199-210.

LAHIRI, P. & MAITI, T. (2002) Empirical Bayes estimation of relative risks in disease mapping. *Calcutta Statistical Association Bulletin,* 55**,** 211-212.

LAHIRI, P. & RAO, J. N. K. (1995) Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association,* 90**,** 758-766.

LAHIRI, S. N., MAITI, T., KATZOFF, M. & PARSONS, V. (2007) Resampling-based empirical prediction: an application to small area estimation. *Biometrika,* 94**,** 469-485.

LAIRD, N. M. & LOUIS, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association,* 82**,** 739-757.

LAIRD, N. M. & WARE, J. H. (1982) Random-effects models for longitudinal data. *Biometrics,* 38**,** 963-974.

LANGE, N. & RYAN, L. (1989) Assessing normality in random effects models. *The Annals of Statistics,* 24**,** 255-286.

LEHTONEN, R., SÄRNDAL, C.-E. & VEIJANEN, A. (2003) The effect of model choice in estimation for domains, including small domains. *Survey Methodology,* 29**,** 33-44.

LESAGE, J. P. & PACE, R. K. (2004) *Spatial and Spatiotemporal Econometrics,* Oxford, Elsevier.

LI, H. (2007) Small Area Estimation: an empirical best linear unbiased prediction approach. PhD thesis, Maryland, College Park, University of Maryland [disponível em http://www.lib.umd.edu/drum/bitstream/1903/7600/1/umi-umd-4867.pdf, acedido em 20/10/2008].

LITTEL, R., MILLIKEN, G., STROUP, W., WOLFINGER, R. & SCHABENBERGER, O. (2006) *SAS® for Mixed Models,* Cary, SAS Institute Inc.

LITTELL, R. C., PENDERGAST, J. & NATARAJAN, R. (2000) Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine,* 19**,** 1793-1819.

MAITI, T. (2004) Applying jackknife method of mean squared prediction error estimation in Saipe. *Statistics in Transition* 6**,** 685-695.

MARKER, D. A. (1999) Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics,* 15**,** 1-24.

MCCULLOCH, C. E. & SEARLE, S. R. (2001) *Generalized, Linear, and Mixed Models,* New York, John Wiley & Sons.

MCLEAN, R. A. & SANDERS, W. L. (1988) Approximating Degrees of Freedom for Standard Errors in Mixed Linear Models. *Proceedings of the Statistical Computing Section.* New Orleans, American Statistical Association.

MCLEAN, R. A., SANDERS, W. L. & STROUP, W. W. (1991) A unified approach for mixed linear models. *The American Statistician,* 45**,** 54-64.

MILES, D. (1995) *Housing, financial markets and the wider economy. Series in Financial Economics and Quantitative Analysis,* New York, John Wiley & Sons.

MILITINO, A. F., UGARTE, M. D. & GOICOA, T. (2007*a*) A BLUP synthetic versus an EBLUP estimator: an empirical study of a small area estimation problem. *Journal of Applied Statistics,* 34**,** 153-165.

MILITINO, A. F., UGARTE, M. D. & GOICOA, T. (2007*b*) Combining sampling and model weights in agriculture small area estimation. *Environmetrics,* 18**,** 87-99.

MILLER, J. J. (1977) Asymptotic properties of maximum likelihood estimates in the Mixed Model of the Analysis of Variance. *The Annals of Statistics,* 5**,** 746-762.

MOLINA, I., SALVATI, N. & PRATESI, M. (2007) Bootstrap for estimating the mean squared error of the spatial EBLUP. *Working Paper 07-34.* Madrid, Departamento de Estadística, Universidad Carlos III de Madrid.

MOONEY, C. Z. (1997) Monte Carlo simulation. *Sage University Paper series on Quantitative Applications in the Social Sciences (series no. 07-116).* Thousand Oaks, Sage.

MORAN, P. A. F. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society: Series B,* 10**,** 243-251.

MORAN, P. A. F. (1950) Notes on continous stochastic phenomena. *Biometrika,* 37**,** 17-23.

MOURA, F. A. S. & HOLT, D. (1999) Small area estimation using multilevel models. *Survey Methodology,* 25**,** 73-80.

MUKHOPADHYAY, P. (1998) *Small area estimation in survey sampling,* New Delhi, Narosa.

MURTEIRA, B., RIBEIRO, C. S., SILVA, J. A. & PIMENTA, C. (2002) *Introdução à estatística,* Lisboa, McGraw-Hill.

NEYMAN, J. & SCOTT, E. L. (1948) Consistent estimates based on partially consistent observations. *Econometrica,* 16**,** 1-32.

ORD, K. (1975) Estimation methods for models of spatial interaction. *Journal Accoust Soc Amer,* 70**,** 120-126.

PANTULA, S. G. & POLLOCK, K. H. (1985) Nested analysis of variance with autocorrelated errors. *Biometrics,* 41**,** 909-920.

PATTERSON, H. D. & THOMPSON, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika,* 58**,** 545-554.

PATTERSON, H. D. & THOMPSON, R. (1974) Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference.* Biometric Society.

PETERSEN, K. B. & PEDERSEN, M. S. (2008) The Matrix Cookbook. WordPress [disponível em http://matrixcookbook.com, acedido em 22/03/2008].

PETRUCCI, A., PRATESI, M. & SALVATI, N. (2005) Geographic information in small area estimation: small area models and spatially correlated random area effects. *Statistics in Transition,* 7**,** 609-623.

PETRUCCI, A. & SALVATI, N. (2004*a*) Small area estimation considering spatially correlated errors. *Working Paper 2004/10.* Firenzi, Universitá degli Studi di Firenzi.

PETRUCCI, A. & SALVATI, N. (2004*b*) Small area estimation using spatial information: the Rathbun Lake Watershed case study. *Working Paper 2004/02.* Firenzi, Universitá degli Studi di Firenzi.

PETRUCCI, A. & SALVATI, N. (2006) Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics,* 11**,** 169-182.

PFEFFERMANN, D. (1984) On extensions for the Gauss-Markov Theorem to the Case of Stochastic Regression Coefficients. *Journal of the Royal Statistical Society: Series B,* 46**,** 139-148.

PFEFFERMANN, D. (2002) Small area estimation - New developments and directions. *International Statistical Review,* 70**,** 125-143.

PFEFFERMANN, D. & BARNARD, C. H. (1991) Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics,* 9**,** 73-84.

PFEFFERMANN, D. & BURCK, L. (1990) Robust small area estimation combining time series and cross-sectional data. *Survey Methodology,* 16**,** 217-237.

PFEFFERMANN, D., FEDER, M. & SIGNORELLI, D. (1998) Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics,* 16**,** 339-348.

PFEFFERMANN, D. & GLICKMAN, H. (2004) Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

PFEFFERMANN, D. & SVERCHKOV, M. (2005) Small Area Estimation under Informative Sampling. *Statistics in Transition,* 7**,** 675-684.

PFEFFERMANN, D. & TILLER, R. B. (2005) Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis,* 26**,** 893-816.

PFEFFERMANN, D. & TILLER, R. B. (2006) Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association,* 101**,** 1387-397.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C.-E. & SINGH, M. P. (1987) *Small area statistics: An international symposium,* New York, John Wiley & Sons.

PRASAD, N. G. N. & RAO, J. N. K. (1990) The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association,* 85**,** 163-171.

PRATESI, M. & SALVATI, N. (2004) Spatial EBLUP in agricultural surveys : an application based on Italian census data. *Report n. 256.* Pisa, University of Pisa.

PRATESI, M. & SALVATI, N. (2005) Small area estimation: the EBLUP estimator with autoregressive random area effects. *Report n. 261*. Pisa, University of Pisa.

PRATESI, M. & SALVATI, N. (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications,* 17**,** 113-141.

PURCELL, N. J. & KISH, L. (1979) Estimation for small domains. *Biometrics,* 35**,** 365-384.

PURCELL, N. J. & KISH, L. (1980) Postcensal Estimates for Local Areas (or Domains). *International Statistical Review,* 48**,** 3-18.

QUENOUILLE, M. (1949) Approximation tests of correlation in time series. *Journal of the Royal Statistical Society: Series B,* 11**,** 18-84.

RAO, C. R. (1970) Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association,* 65**,** 161-172.

RAO, C. R. (1971) Estimation of variance and covariance components - MINQUE theory. *Journal of Multivariate Analysis,* 1**,** 257-275.

RAO, C. R. (1972) Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association,* 67**,** 112-115.

RAO, J. N. K. (1999*a*) Some current trends in sample survey theory and methods. *Sankhyā: The Indian Journal of Statistics,* 61**,** 1-57.

RAO, J. N. K. (1999*b*) Some recent advances in model-based small area estimation. *Survey Methodology,* 25**,** 175-186.

RAO, J. N. K. (2000) Statistical methodology for indirect estimations in small areas. EUSTAT-The Basque Statistics Institute [disponível em http://www.eustat.es/prodserv/seminarioviejo_i.html, acedido em 25/08/2008].

RAO, J. N. K. (2003) *Small area estimation,* New Jersey, John Wiley & Sons.

RAO, J. N. K. (2005*a*) Inferential Issues in Small Area Estimation: Some New Developments. *Statistics in Transition,* 7**,** 513-526.

RAO, J. N. K. (2005*b*) Interplay Between Sample Survey Theory and Practice: An Appraisal. *Survey Methodology,* 31**,** 117-138.

RAO, J. N. K. & WU, C. F. J. (1988) Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association,* 83**,** 231-241.

RAO, J. N. K. & YU, M. (1992) Small area estimation by combining time series and cross-sectional data. *Proceedings of the Section on Survey Research Methods.* American Statistical Association.

RAO, J. N. K. & YU, M. (1994) Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics,* 22**,** 511-528.

RICHARDSON, A. M. & WELSH, A. H. (1994) Asymptotic properties of restricted máximum likelihood (REML) estimates for hierarchical mixed linear models. *Australian Journal of Statistics,* 36**,** 31-43.

ROBINSON, D. L. (1987) Estimation and use of variance components. *The Statistician,* 36**,** 3-14.

ROBINSON, G. K. (1991) That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science,* 6**,** 15-51.

ROYSTON, P. (1982) An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics,* 31**,** 115-124.

SAEI, A. & CHAMBERS, R. (2003*a*) Small area estimation under linear and generalized linear mixed models with time and area effects. *Working Paper M03/15.* Southampton, Southampton Statistical Sciences Research Institute.

SAEI, A. & CHAMBERS, R. (2003*b*) Small area estimation: A review of methods based on the application of mixed models. *Working Paper M03/16.* Southampton, Southampton Statistical Sciences Research Institute.

SALVATI, N. (2004) Small area estimation by spatial models: The spatial empirical best linear unbiased prediction (Spatial EBLUP). *Working Paper 2004/03.* Firenzi, Universitá degli Studi di Firenzi.

SALVATI, N. (2005) Small area estimation: the EBLUP estimator using the CAR model. *Report n. 260.* Pisa, University of Pisa.

SÄRNDAL, C.-E. (1984) Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association,* 79**,** 624-631.

SÄRNDAL, C.-E. & HIDIROGLOU, M. A. (1989) Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association*, 84**,** 266-275.

SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. (1992) *Model assisted survey sampling,* New-York, Springer-Verlag.

SCHAIBLE, W. L. (1996) *Indirect Estimation in U.S. Federal Programs,* New York, Springer-Verlag.

SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. (1992) *Variance components,* New York, John Wiley & Sons.

SHAO, J. & TU, D. (1995) *The Jackknife and Bootstrap,* New York, Springer-Verlag.

SHAPIRO, S. S. & WILK, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika,* 52**,** 591-611.

SHAW, R. G. (1987) Maximum-likelihood approaches applied to quantitative genetics of natural populations. *Evolution,* 41**,** 812-826.

SINGH, A. C., MANTEL, H. J. & THOMAS, B. W. (1994) Time series EBLUPs for small areas using survey data. *Survey Methodology,* 20**,** 33-43.

SINGH, A. C., STUKEL, D. M. & PFEFFERMANN, D. (1998) Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society: Series B,* 60**,** 377-396.

SINGH, B. B. & SHUKLA, G. K. (1983) A test of autoregression in Gaussian spatial processes. *Biometrika,* 70**,** 523-527.

SINGH, B. B., SHUKLA, G. K. & KUNDU, D. (2005) Spatio-Temporal Models in Small Area Estimation. *Survey Methodology,* 31**,** 183-195.

SINGH, M. P. & TESSIER, R. (1976) Some estimators for domain totals. *Journal of the American Statistical Association,* 71**,** 322-325.

SITTER, R. R. (2001) Resampling methods in complex surveys: an overview in honour of J.N.K. Rao's retirement. *Proceedings of the Survey Methods Section.* SSC Annual Meeting.

SLUD, E. V. & MAITI, T. (2006) MSE estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society: Series B,* 68**,** 239-257.

SPE & ABE (2007) Glossário Estatístico Inglês-Português. Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística [disponível em http://www.spestatistica.pt/?q=artigo/glossario-estatistico-ingles---portugues].

SPEED, T. P. (1997) Restricted maximum likelihood (REML). IN KOTZ, S., READ, C. B. & BANKS, D. L. (Eds.) *Encyclopedia of Statistical Sciences.* New York, John Wiley & Sons.

STEPHENS, M. A. (1974) EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association,* 69**,** 730-737.

STUKEL, D. M. & RAO, J. N. K. (1999) On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference,* 78**,** 131-147.

THOMPSON, S. K. (1992) *Sampling,* New-York, John Wiley & Sons.

THOMPSON, W. A. (1962) The problem of negative estimates of variance

components. *The Annals of Mathematical Statistics,* 33**,** 273-289.

TUKEY, A. P. (1958) Bias and confidence in not quite large samples (abstract). *The Annals of Mathematical Statistics,* 29**,** 614.

UGARTE, M. D., MILITINO, A. F. & GOICOA, T. (2009) Benchmarked estimates in small areas using linear mixed models with restrictions. *Test***,** doi: 10.1007/s11749-008-0094-x (in press).

UPTON, G. J. G. & FINGLETON, B. (1985) *Spatial Data Analysis by Example,* New York, John Wiley & Sons.

VERBEKE, G. & MOLENBERGHS, G. (2000) *Linear mixed models for longitudinal data,* New-York, Springer-Verlag.

VERBYLA, A. P. (1990) A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics,* 32**,** 227-230.

WALL, M. M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference,* 121**,** 311-324.

WANG, J. & FULLER, W. A. (2003) The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association,* 98**,** 716-723.

WANG, J., FULLER, W. A. & QU, Y. (2008) Small area estimation under a restriction. *Survey Methodology,* 34**,** 29-36.

WHITTLE, P. (1954) On stationary process in the plane. *Biometrika,* 41**,** 434-449.

WIKIPÉDIA (s.d.) Unidades Territoriais Estatísticas de Portugal. Wikipédia [disponível http://pt.wikipedia.org/wiki/Unidades_Territoriais_Estat%C3%ADsticas_de_Portugal, acedido em 12/04/2009].

WOLFINGER, R. D. (1993) Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computing,* 22**,** 1079-1106.

WOLFINGER, R. D. (1996) Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological and Environmental Statistics,* 1**,** 205-230.

WOLTER, K. M. (2007) *Introduction to Variance Estimation,* New York, Springer.

WOODRUFF, R. S. (1971) A Simple Method for Approximating the Variance of a

Complicated Estimate. *Journal of the American Statistical Association,* 66**,** 411 - 414.

WU, C. F. J. (1986) Jacknife, Bootstrap and other Resampling Methods in Regression Analysis (with discussion). *The Annals of Statistics,* 14**,** 1261-1350.

YOU, Y. & CHAPMAN, B. (2006) Small Area Estimation Using Area Level Models and Estimated Sampling Variances. *Survey Methodology,* 32**,** 97-103.

YOU, Y., RAO, J. N. K. & GAMBINO, J. (2001) Model-based unemployment rate estimation for the canadian labour force survey: A hierarchical bayes approach. *Technical Report.* Ottawa, Household Survey Methods Division, Statistics Canada.

YU, M. (1993) Nested-Error Regression Models and Small Area Estimation Combining Cross-Sectional and Time Series Data. *Department of Mathematics and Statistics.* PhD Thesis, Ottawa, Carleton University.

ANEXOS

ANEXO 1

Inquérito aos Preços Médios de Transacção na Habitação

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

Indicadores de Preços na Habitação

IPH - Mediação

(•) 1.1 - Inquérito aos Preços Médios de Transacção na Habitação

(Registado no INE, sob o Nº. 9200)

( )

1.3 - Password:

Continuar | Corrigir | Sair

Este Instrumento de Notação do Sistema Estatístico Nacional, conforme Lei Nº 6/89, de 15 de Abril, É DE RESPOSTA OBRIGATÓRIA.

A resposta aos presentes inquéritos deve ser entregue no prazo máximo de 15 dias, após o período a que respeitam.

Validade: 2006/12/31

2.1 - Password: ***

2.2 - Nome:

2.3 - Morada:

2.4 - Código Postal: 2.5

2.6 - Concelho:

2.7 - Telefone: 2.8 - Fax:

2.9 - Email:

2.10 - Nº de Contribuinte:

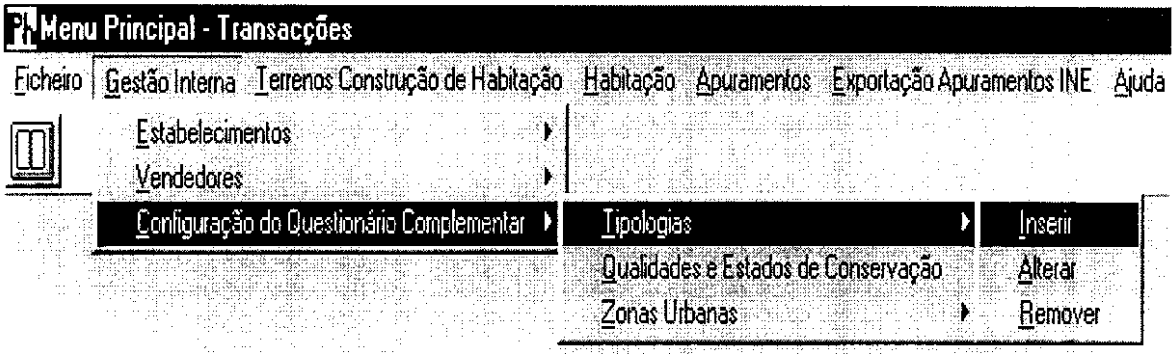
CONTACTO

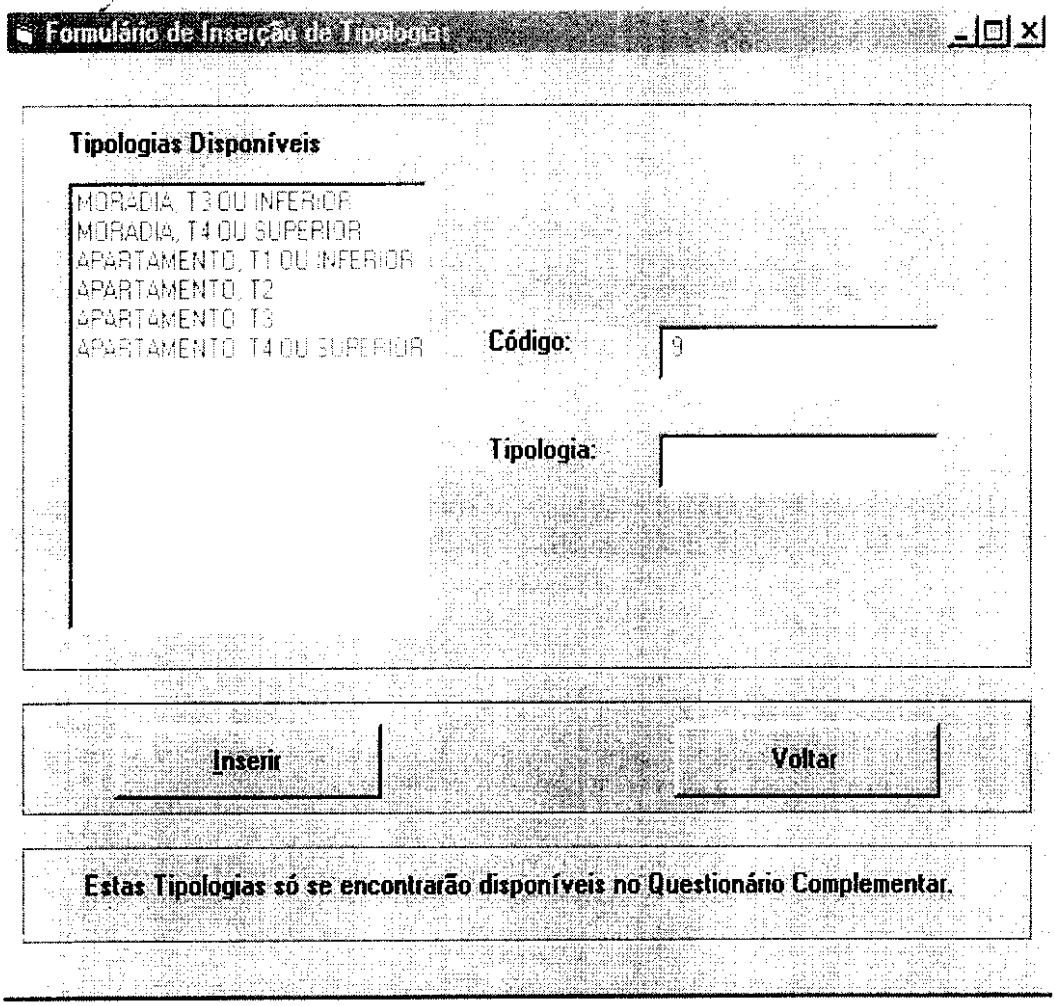
2.11 - Nome:

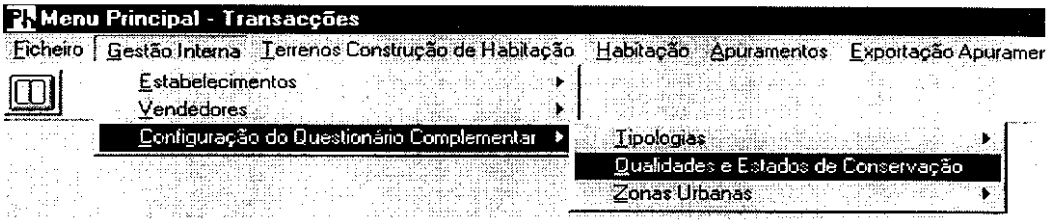
2.12 - Departamento:

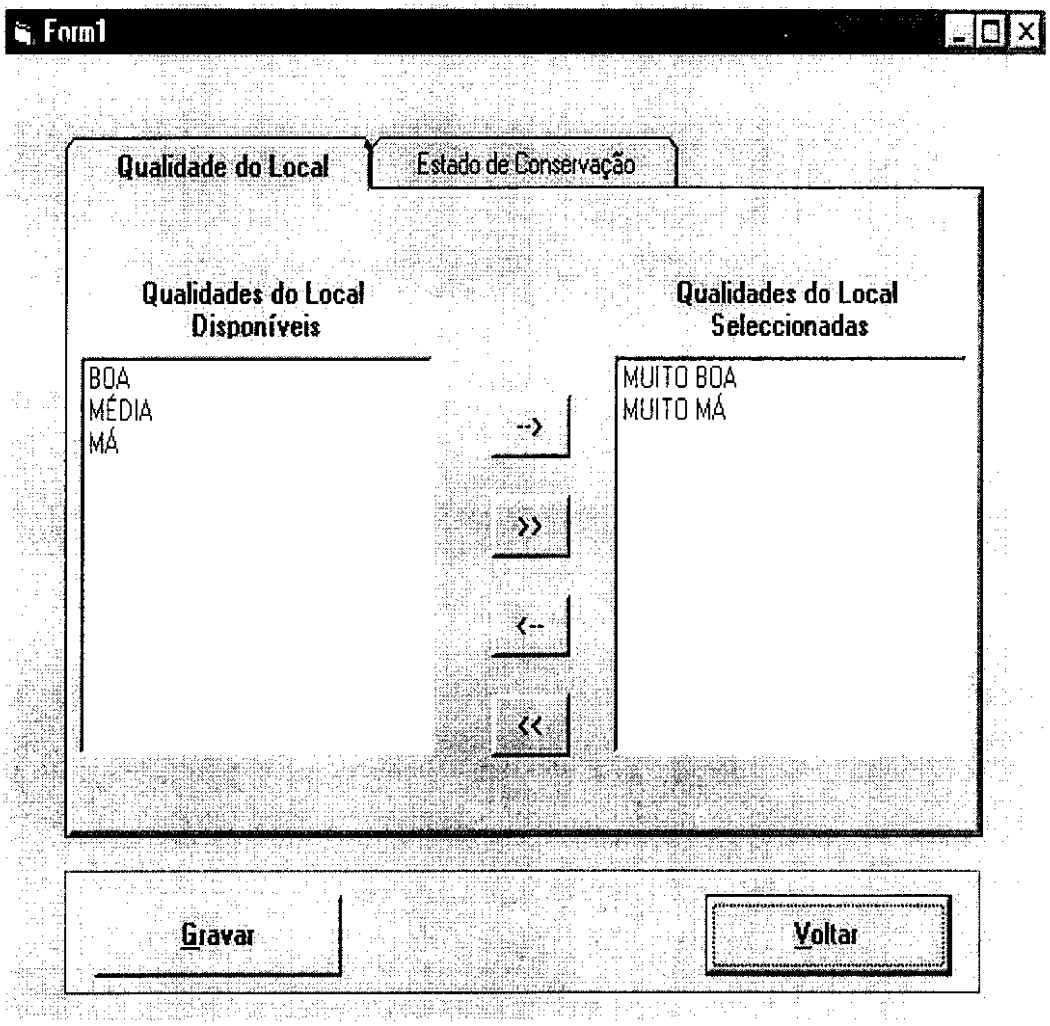
Gravar Voltar

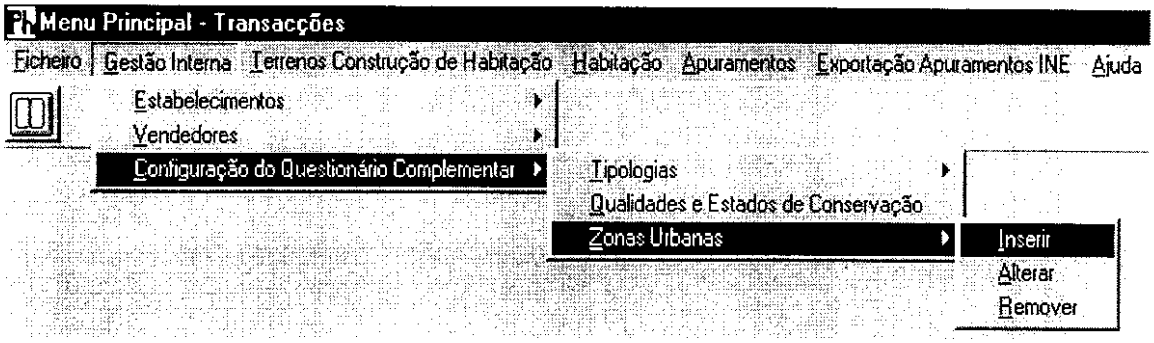
Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.

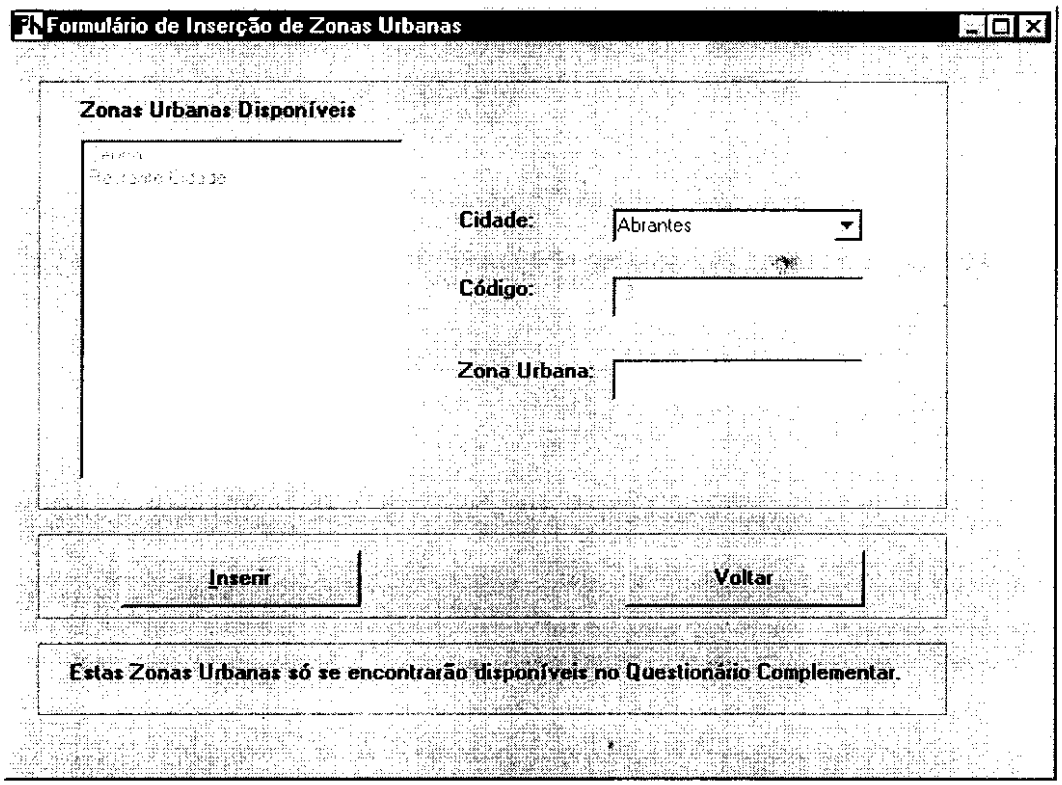

Menu Principal - Transacções
Ficheiro
Gestão Interna
Terrenos Construção de Habitação
Habitação
Apuramentos
Exportação Apuramentos INE
Ajuda
Estabelecimentos
Vendedores
Configuração do Questionário Complementar
Tipologias
Qualidades e Estados de Conservação
Zonas Urbanas
Inserir
Alterar
Remover


Formulário de Inserção de Tipologias
Tipologias Disponíveis
MORADIA, T3 OU INFERIOR
MORADIA, T4 OU SUPERIOR
APARTAMENTO, T1 OU INFERIOR
APARTAMENTO, T2
APARTAMENTO, T3
APARTAMENTO, T4 OU SUPERIOR
Código:
9
Tipologia:
Inserir
Voltar
Estas Tipologias só se encontrarão disponíveis no Questionário Complementar.


Menu Principal - Transacções
Ficheiro
Gestão Interna
Terrenos Construção de Habitação
Habitação
Apuramentos
Exportação Apuramer
Estabelecimentos
Vendedores
Configuração do Questionário Complementar
Tipologias
Qualidades e Estados de Conservação
Zonas Urbanas


Form1
Qualidade do Local
Estado de Conservação
Qualidades do Local Disponíveis
BOA
MÉDIA
MÁ
-->
>>
<--
<<
Qualidades do Local Seleccionadas
MUITO BOA
MUITO MÁ
Gravar
Voltar


Menu Principal - Transacções
Ficheiro
Gestão Interna
Terrenos Construção de Habitação
Habitação
Apuramentos
Exportação Apuramentos INE
Ajuda
Estabelecimentos
Vendedores
Configuração do Questionário Complementar
Tipologias
Qualidades e Estados de Conservação
Zonas Urbanas
Inserir
Alterar
Remover


Formulário de Inserção de Zonas Urbanas
Zonas Urbanas Disponíveis
Cidade:
Abrantes
Código:
Zona Urbana:
Inserir
Voltar
Estas Zonas Urbanas só se encontrarão disponíveis no Questionário Complementar.

Formulário de Inserção de Registos de Transacções de Terrenos

Ref.: 9.3 - Ano: 9.4 - Mês:

9.5 - O Terreno tem Alvará de Loteamento ou Estudo de Viabilidade?

9.6 - O registo diz respeito a um Contrato Promessa?

9.7 - Se respondeu Não no campo anterior, indique se já houve registo da Transacção em causa.

9.8 - Área Total de Construção Autorizada m2

9.9 - Da Área referida, 50% ou mais é destinada a Habitação? Se a Área em causa fôr nula, responda Não.

Número de Fogos para: 9.10 - Moradias

9.11 - Apartamentos

9.12 - Concelho: Porto 9.13 - Freguesia: Campanhã

Valor da Transacção: 9.14 Cts.

9.15 Euros

Gravar Voltar Fechar

Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.

Formulário de Inserção de Registos de Transacções de Habitações

Ref.: 11.3 - Ano: 11.4 - Mês:

11.5 - O registo diz respeito a um Contrato Promessa?

11.6 - Se respondeu Não no campo anterior, indique se já existia anteriormente registo da Transacção em causa.

11.7 - Ano de Construção / Reconstrução

11.8 - Área Útil: m2

11.9 - Concelho: 11.10 - Freguesia:

Valor da Transacção: 11.11 Cts.

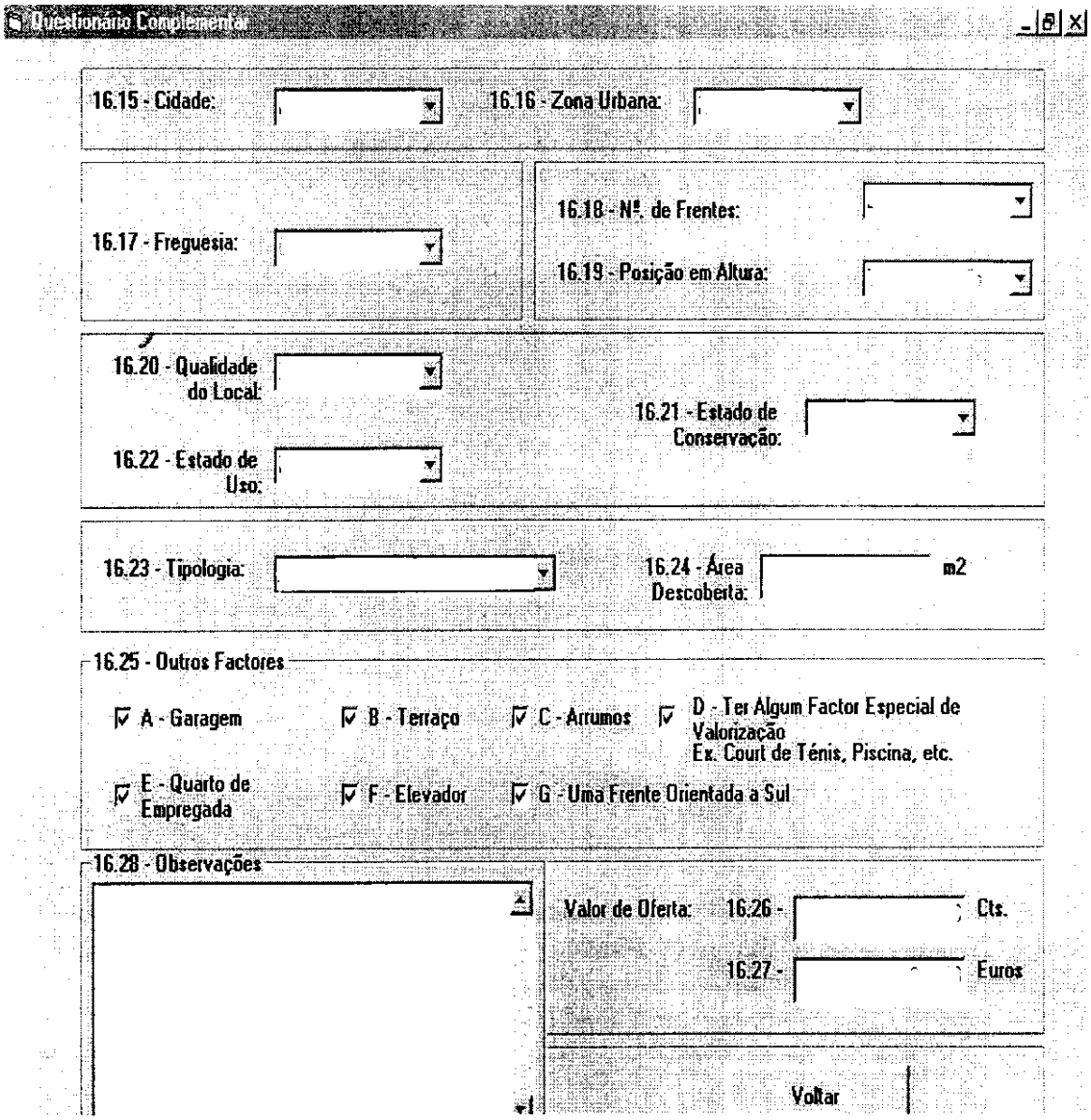
11.12 Euros

Tipologia do Alojamento: 11.13

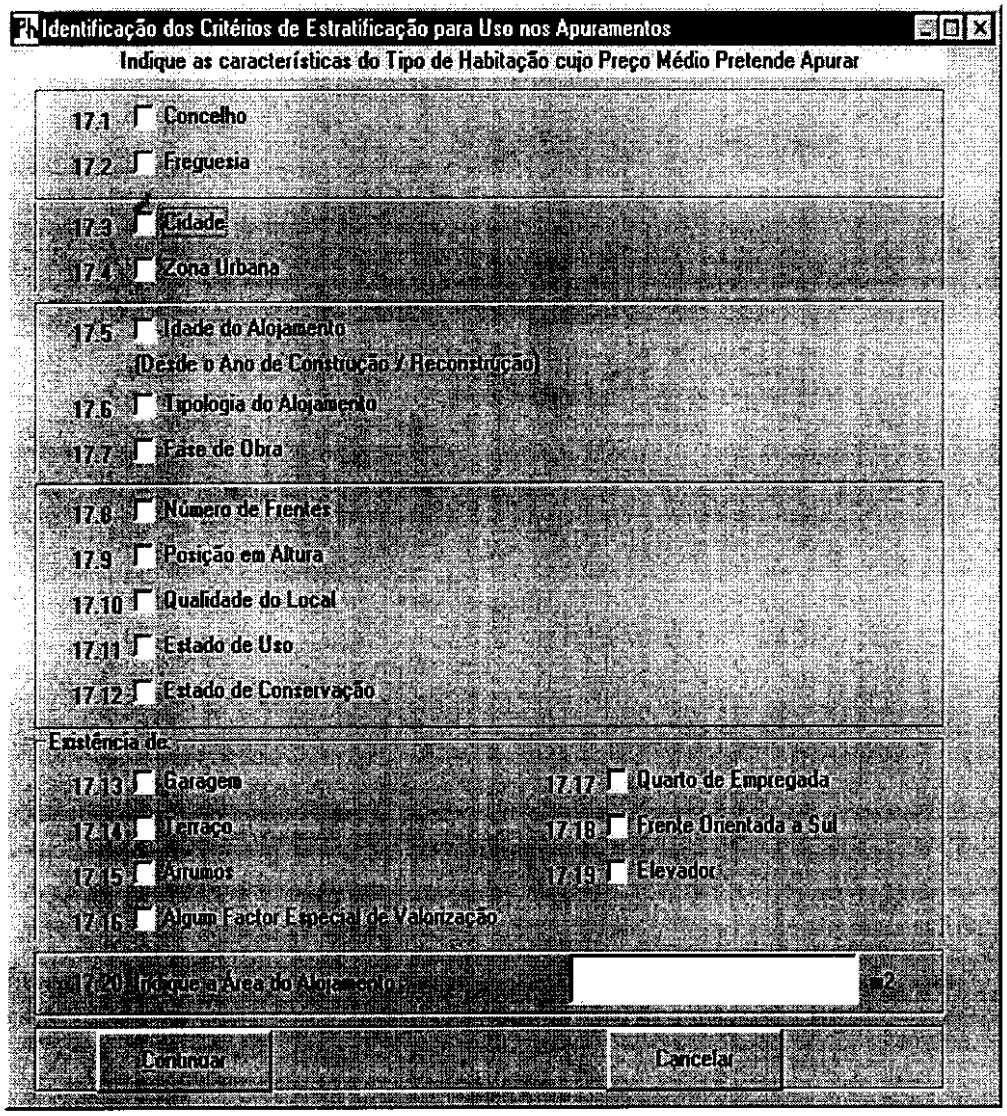
Fase da Obra: 11.14

Gravar | Voltar | Fechar | Questionário Complementar

Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.


Questionário Complementar
16.15 - Cidade:
16.16 - Zona Urbana:
16.17 - Freguesia:
16.18 - Nº. de Frentes:
16.19 - Posição em Altura:
16.20 - Qualidade do Local:
16.21 - Estado de Conservação:
16.22 - Estado de Uso:
16.23 - Tipologia:
16.24 - Área Descoberta:
m2
16.25 - Outros Factores
A - Garagem
B - Terraço
C - Arrumos
D - Ter Algum Factor Especial de Valorização
Ex. Court de Ténis, Piscina, etc.
E - Quarto de Empregada
F - Elevador
G - Uma Frente Orientada a Sul
16.28 - Observações
Valor de Oferta:
16.26 -
Cts.
16.27 -
Euros
Voltar


acções
arenos Construção de Habitação
Habitação
Apuramentos
Exportação Apuramentos INE
Ajuda
INE
Apuramentos Administrativos
Apuramentos de Performance
Preços Médios de Alojamentos


Identificação dos Critérios de Estratificação para Uso nos Apuramentos
Indique as características do Tipo de Habitação cujo Preço Médio Pretende Apurar
17.1 Concelho
17.2 Freguesia
17.3 Cidade
17.4 Zona Urbana
17.5 Idade do Alojamento
(Desde o Ano de Construção / Reconstrução)
17.6 Tipologia do Alojamento
17.7 Fase de Obra
17.8 Número de Frentes
17.9 Posição em Altura
17.10 Qualidade do Local
17.11 Estado de Uso
17.12 Estado de Conservação
Existência de:
17.13 Garagem
17.14 Terraço
17.15 Arrumos
17.16 Algum Factor Especial de Valorização
17.17 Quarto de Empregada
17.18 Frente Orientada a Sul
17.19 Elevador
17.20 Indique a Área do Alojamento
m2
Continuar
Cancelar

ANEXO 2

Inquérito aos Preços de Avaliação Bancária na Habitação

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

IPH Indicadores de Preços na Habitação

IPH - Mediação

○ 1.2 - Inquérito aos Preços de Avaliação Imobiliária na Habitação

(Registado no INE, sob o Nº. 9201)

1.3 - Password:

Continuar | Corrigir | Sair

Este Instrumento de Notação do Sistema Estatístico Nacional, conforme Lei Nº 6/89, de 15 de Abril, É DE RESPOSTA OBRIGATÓRIA.

A resposta aos presentes inquéritos deve ser entregue no prazo máximo de 15 dias, após o período a que respeitam.

Validade: 2006/12/31

2.1 - Password: xxx

2.2 - Nome:

2.3 - Morada:

2.4 - Código Postal: 2.5

2.6 - Concelho:

2.7 - Telefone: 2.8 - Fax:

2.9 - Email:

2.10 - Nº de Contribuinte:

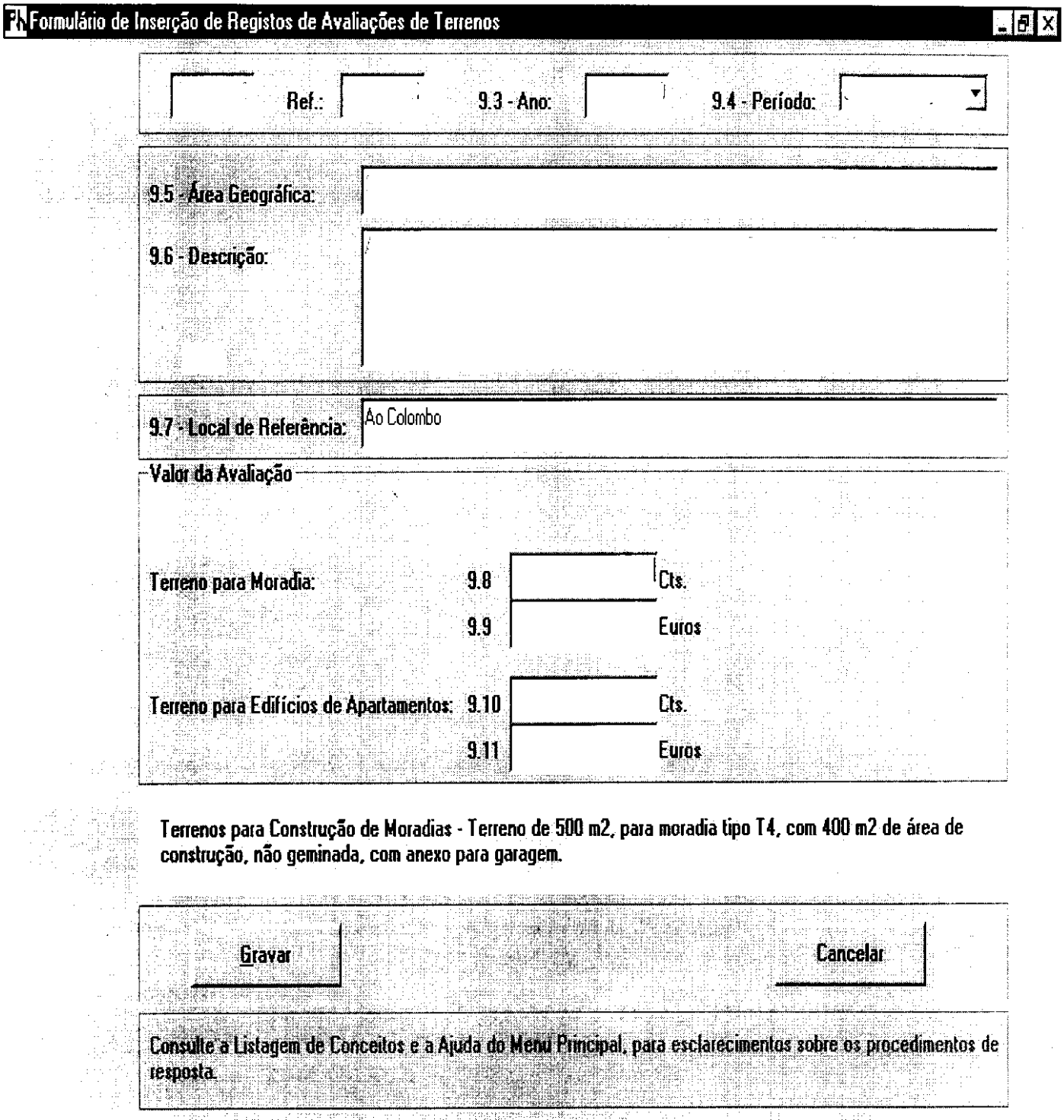
CONTACTO

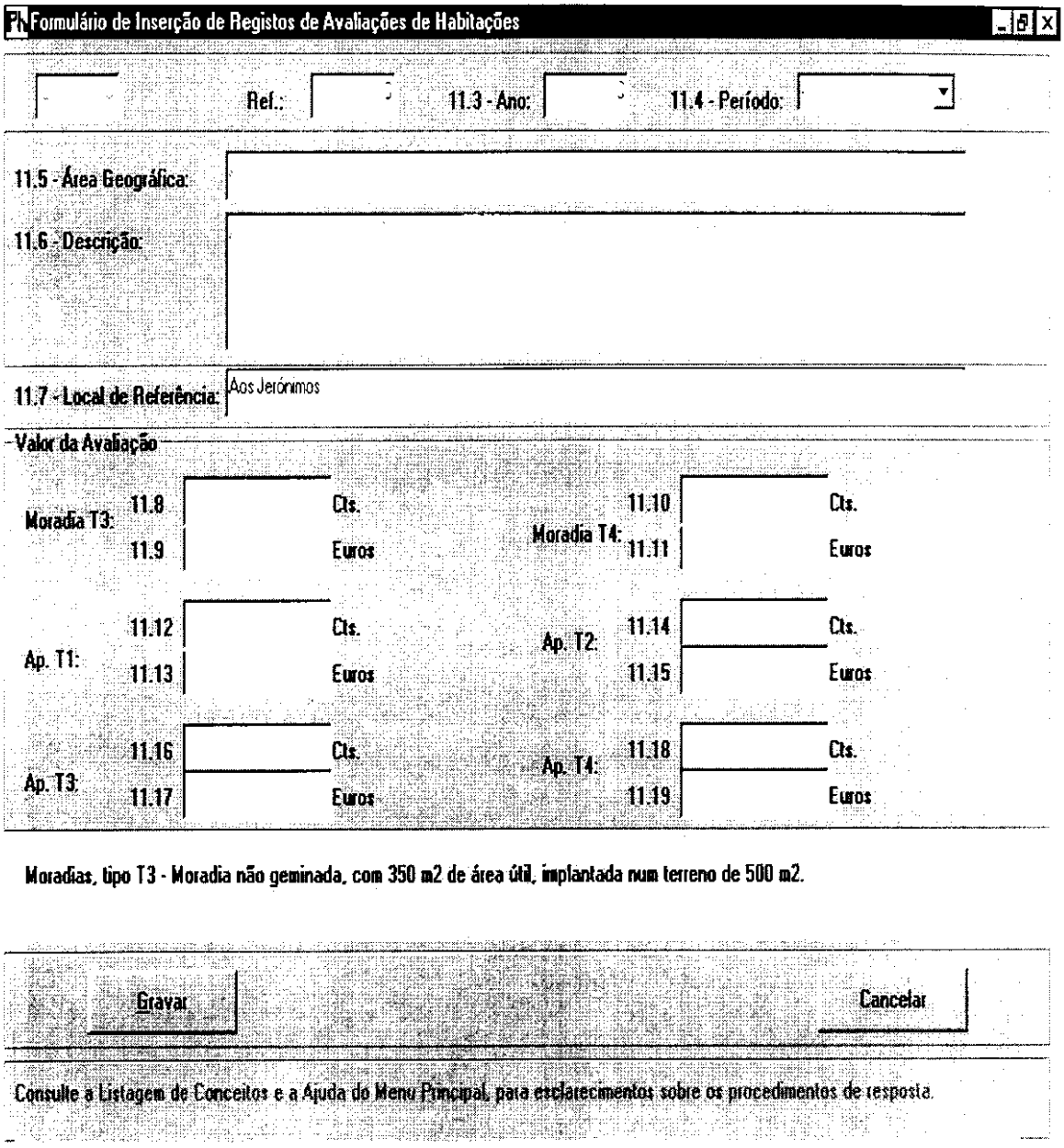
2.11 - Nome:

2.12 - Departamento:

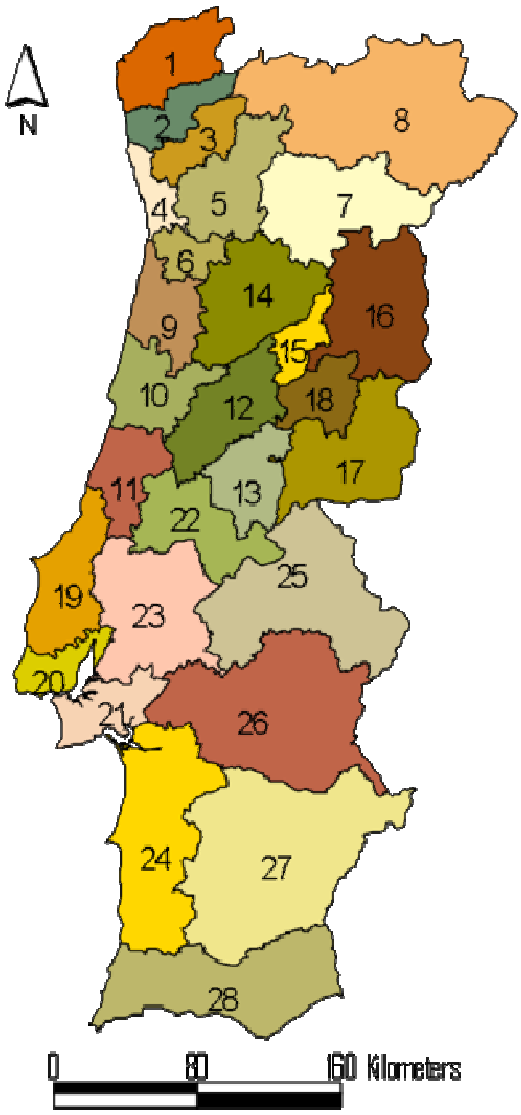
Gravar Voltar

Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.


Formulário de Inserção de Registos de Avaliações de Terrenos
Ref.:
9.3 - Ano:
9.4 - Período:
9.5 - Área Geográfica:
9.6 - Descrição:
9.7 - Local de Referência:
Ao Colombo
Valor da Avaliação
Terreno para Moradia:
9.8
Cts.
9.9
Euros
Terreno para Edifícios de Apartamentos:
9.10
Cts.
9.11
Euros
Terrenos para Construção de Moradias - Terreno de 500 m2, para moradia tipo T4, com 400 m2 de área de construção, não geminada, com anexo para garagem.
Gravar
Cancelar
Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.


Formulário de Inserção de Registos de Avaliações de Habitações
Ref.:
11.3 - Ano:
11.4 - Período:
11.5 - Área Geográfica:
11.6 - Descrição:
11.7 - Local de Referência:
Aos Jerónimos
Valor da Avaliação
Moradia T3:
11.8
Cts.
11.9
Euros
Moradia T4:
11.10
Cts.
11.11
Euros
Ap. T1:
11.12
Cts.
11.13
Euros
Ap. T2:
11.14
Cts.
11.15
Euros
Ap. T3:
11.16
Cts.
11.17
Euros
Ap. T4:
11.18
Cts.
11.19
Euros
Moradias, tipo T3 - Moradia não geminada, com 350 m2 de área útil, implantada num terreno de 500 m2.
Gravar
Cancelar
Consulte a Listagem de Conceitos e a Ajuda do Menu Principal, para esclarecimentos sobre os procedimentos de resposta.

ANEXO 3 – Mapa das NUTSIII de Portugal Continental



Legenda:

1-Minho-Lima; 2-Cávado; 3-Ave; 4-Grande Porto; 5-Tâmega; 6-Entre Douro e Vouga; 7-Douro; 8-Alto Trás-os-Montes; 9-Baixo Vouga; 10-Baixo Mondego; 11-Pinhal Litoral; 12-Pinhal Interior Norte; 13-Pinhal Interior Sul; 14-Dão-Lafões; 15 - Serra da Estrela; 16-Beira Interior Norte; 17-Beira Interior Sul; 18-Cova da Beira; 19-Oeste; 20-Grande Lisboa; 21-Península de Setúbal; 22-Médio Tejo; 23-Lezíria do Tejo; 24-Alentejo Litoral; 25-Alto Alentejo; 26-Alentejo Central; 27-Baixo Alentejo; 28-Algarve.

Fonte: Wikipédia (sd)