A estimação em pequenos domínios do preço médio de transacção da habitação

Luís N. Pereira[1]
Escola Superior de Gestão, Hotelaria e Turismo,
Universidade do Algarve

Pedro S. Coelho
Instituto Superior de Estatística e Gestão de Informação,
Universidade Nova de Lisboa

Resumo

O Instituto Nacional de Estatística de Portugal pretende produzir estimativas dos preços médios de transacção da habitação para os concelhos e para as regiões de Portugal classificadas ao nível III da Nomenclatura das Unidades Territoriais para Fins Estatísticos, com base na informação recolhida através do Inquérito aos Preços Médios de Transacção da Habitação. Contudo, para estes domínios de estimação, não é possível produzir estimativas directas com um grau de precisão aceitável devido às pequenas dimensões amostrais. Neste estudo propõe-se uma metodologia de estimação dos preços médios de transacção da habitação para pequenos domínios, que tira partido, por um lado, de informação auxiliar de natureza administrativa e, por outro, de informação amostral de natureza histórica. Recorre-se a um modelo que combina dados cronológicos e seccionais, envolvendo efeitos aleatórios com uma estrutura de covariância arbitrária ao longo do tempo e erros de sondagem heterocedásticos. O preço médio de avaliação bancária da habitação é utilizado como variável auxiliar no modelo. No âmbito do modelo proposto, é obtido o estimador em dois passos para pequenos domínios. O estimador proposto é assistido por uma classe de modelos que se enquadra no âmbito do modelo linear geral misto. A precisão relativa das estimativas produzidas por este estimador combinado é comparada com a precisão relativa das estimativas baseadas no desenho, produzidas pelo estimador directo pós-estratificado, utilizado como termo de comparação.

[1] Doutorando na Faculdade de Economia da Universidade do Algarve, no ramo de Métodos Quantitativos Aplicados à Economia e à Gestão, especialidade de Estatística.

Palavras-chave: Estimação em pequenos domínios, estimadores directos, estimadores combinados, estimação baseada no modelo, dados cronológicos.

Abstract

Title: ***Small Area Estimation***

The Portuguese National Statistical Institute intends to produce estimations for the mean price of the habitation transaction for Portugal councils and for other regions, using information collected through a Survey to the Habitation Transaction Prices. However, for these domains of estimation, it is not possible to provide direct estimations with an acceptable degree of precision because sample sizes in small areas are seldom large enough. In this study a methodology for the estimation of mean price of the habitation transaction of small areas is proposed. This methodology takes advantage, on the one hand of auxiliary administrative information, and on the other hand, of sample historical information. A combined cross-sectional and time-series model involving random effects with an arbitrary covariance structure over time and heterocedastic sampling errors is proposed. The mean price of the bank evaluation transaction is used in the model as auxiliary covariable. In the scope of this model the Empirical Best Linear Unbiased Predictor is obtained. The proposed estimator is assisted by a class of models that is a particular case of the general mixed linear model. The relative precision of the produced estimations by this combined estimator is compared with the relative precision of the design-based estimations obtained through the direct pos-stratificated estimator, which was used as comparing benchmark.

Key-words: Small area estimation, direct estimators, combined estimators, model-based estimation, chronological data.

1. Introdução

A expressão "pequeno domínio" é comummente utilizada como designação de áreas geográficas pequenas, como por exemplo, concelhos ou freguesias, ou como referência a subpopulações pequenas, como por exemplo as minorias étnicas ou os portadores de uma doença muito específica ou rara. A essa expressão está frequentemente associado o problema das pequenas dimensões amostrais retiradas dos domínios de interesse, que dificultam a obtenção de estimativas directas com precisão aceitável.

Nos últimos anos tem-se verificado por todo o mundo um grande crescimento da procura de estatísticas oficiais dignas de confiança e com um grande nível de desagregação. Como consequência, a maioria dos Institutos Nacionais de Estatística tem vindo a sentir a necessidade de publicação sistemática de indicadores estatísticos para pequenos domínios, ao mesmo tempo que se têm vindo a desenvolver diversas linhas de investigação sobre a utilização de estimadores *model-based*, como forma de resolução da limitação da escassez de amostra nesses pequenos domínios.

Em Portugal, o Instituto Nacional de Estatística (INE) está interessado na divulgação de indicadores estatísticos de preços de transacção da habitação para pequenos domínios (regiões de Portugal classificadas ao nível III da Nomenclatura das Unidades Territoriais para Fins Estatísticos – NUTS III, Concelhos e Natureza do Alojamento). A estimação de preços de transacção da habitação é actualmente suportada pelo Inquérito aos Preços Médios de Transacção da Habitação (IPTH), um inquérito repetido no tempo da responsabilidade do INE. No entanto, o método de amostragem utilizado neste inquérito não permite a produção de estimativas directas com um nível de precisão aceitável para níveis mais desagregados do que NUTS II, devido às pequenas dimensões amostrais. Este facto serve de motivação para a utilização de estimadores indirectos, que utilizem informação emprestada de outros domínios ou de outros períodos de tempo, tendo como objectivo a obtenção de estimativas mais precisas.

Os estimadores para pequenos domínios tradicionais, pedem informação emprestada tanto a partir de pequenos domínios semelhantes, como a partir do mesmo domínio ao longo do tempo, mas não de ambos. No entanto, nos últimos anos têm sido desenvolvidas abordagens, nas quais os estimadores utilizam simultaneamente informação emprestada de outros domínios e de outros períodos de tempo.

Os estimadores propostos nas abordagens de Choudhry e Rao (1989), Pfeffermann e Burk (1990), Rao e Yu (1992, 1994), Singh, Mantel e Thomas (1994), Ghosh e Nangia (1993), Ghosh, Nangia e Kim (1996), Datta, Lahiri e Maiti (2002) e de Coelho (2000), assim como o estimador proposto neste artigo, exploram simultaneamente as duas dimensões, produzindo estimativas para pequenos domínios com melhor precisão à que poderia ser obtida através de uma abordagem *design-based*.

Neste estudo propõe-se uma metodologia de estimação dos preços médios de transacção da habitação para pequenos domínios, que tira partido, por um lado, de informação auxiliar de natureza administrativa (avaliações bancárias de habitações) fornecida pelo Inquérito aos Preços de Avaliação Bancária da Habitação (IABH) e, por outro, de informação amostral de natureza histórica através do recurso a modelos baseados em séries cronológicas.

ıdo que não é possível estabelecer a ligação entre os dados amostrais e os dos administrativos a nível individual (isto é, ao nível de cada transacção habitação) a investigação foi circunscrita a modelos ao nível de área, que tabelecem a associação das várias fontes de dados ao nível de freguesia.

Utilizam-se estimadores baseados em variantes da estimação pela regresso, dando-se especial relevo à utilização de estimadores combinados que am partido de dados cronológicos. Os estimadores enquadram-se na classe ıs melhores previsores lineares não enviesados empíricos (*Empirical Best near Unbiased Predictors* – EBLUP) dos parâmetros de interesse.

A precisão relativa das estimativas produzidas por estes estimadores é comrada com a precisão relativa das estimativas *design-based* produzidas pelo timador directo pós-estratificado, que serviu como termo de comparação.

Na secção 2 apresenta-se o estimador directo pós-estratificado. Alguns odelos seccionais e cronológicos utilizados previamente são apresentados ı secção 3. Na secção 4 são propostos modelos que pretendem responder ao oblema de estimação que se coloca. O novo estimador em dois passos é mbém apresentado nesta secção. Na secção 5 é feita referência aos critérios selecção de modelos e de avaliação da qualidade da estimação. Os resuldos da aplicação dos estimadores propostos na estimação do preço médio transacção da habitação em Portugal são apresentados na secção 6. Nesta cção é também feita a comparação entre a precisão relativa das estimativas oduzidas por esses estimadores e a precisão relativa das estimativas produdas pelo estimador directo pós-estratificado. Finalmente, as principais conusões deste estudo dão corpo à secção 7.

. Estimadores directos

Nesta secção apresenta-se um dos estimadores para pequenos domínios ıais utilizado, o estimador directo pós-estratificado, que utiliza valores da ariável de interesse só para um período de tempo e para os elementos do omínio. Um vasto conjunto de estimadores directos para pequenos domínios ode ser encontrado nos trabalhos de Ghosh e Rao (1994), Singh *et al.* 1994), Coelho (1996) e Rao (2000).

Seja θ_d a média populacional da variável de interesse no d-ésimo pequeno domínio, d=1, ..., D, e y_i a i-ésima observação da variável de interesse, Y, i = 1, ..., N. Nesta secção, onde se apresenta um estimador que tiliza apenas dados seccionais, para simplificação da notação ignora-se a ependência de θ_d e de y_i do período de tempo, t. Para um plano de sonlagem genérico, o estimador directo pós-estratificado da média no d-ésimo ıequeno domínio é dado por:

$$\hat{\theta}_d = \sum_{i \in s_d} \frac{y_i}{\pi_i} \Big/ \sum_{i \in s_d} \frac{1}{\pi_i} \tag{2.1}$$

onde π_i é a probabilidade da i-ésima unidade de amostragem pertencer à amostra e s_d representa o conjunto das n_d unidades amostrais pertencentes ao d-ésimo domínio.

No caso de uma sondagem aleatória estratificada de uma população de conglomerados, utilizada neste estudo, o estimador directo pós-estratificado da média no d-ésimo pequeno domínio é dado por:

$$\hat{\theta}_d = \sum_h \frac{M_h}{m_h} \sum_{g \in s_h} \tau_{gd} \Big/ \sum_h \frac{M_h}{m_h} \sum_{g \in s_h} N_{gd} \tag{2.2}$$

onde m_h e M_h representam, respectivamente, o número de conglomerados na amostra e na população do estrato h; y_{hgi} é a i-ésima observação da variável de interesse pertencente ao g-ésimo conglomerado do estrato h; τ_{gd} é o total da variável de interesse no d-ésimo domínio do conglomerado g; e N_{gd} representa o número de unidades secundárias do conglomerado g, que intersectam o domínio d (g = 1, ..., M; h = 1, ..., H; d = 1, ..., D). Este estimador não apresenta em geral níveis de precisão aceitáveis quando as dimensões amostrais de conglomerados em pequenos domínios são pequenas. Este problema pode ser colmatado através da utilização de estimadores que utilizem valores da variável de interesse de domínios e/ou de períodos de tempo relacionados, aumentando desta forma o tamanho "efectivo" da amostra, que por conseguinte aumentará a precisão do estimador. A resolução deste problema passa pela utilização de modelos que utilizem dados seccionais e cronológicos, os quais são apresentados na secção seguinte.

3. Modelos de nível área com dados cronológicos

O estimador directo pós-estratificado apresentado na secção anterior utiliza unicamente dados amostrais da variável de interesse só para um período de tempo e apenas para os elementos do domínio. 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. Nesta secção, apresentam-se algumas extensões do modelo de Fay-Herriot, a estimadores para pequenos domínios que utilizam dados seccionais e cronológicos.

Apesar da maioria da investigação efectuada sobre estimação em peque-domínios, no âmbito dos modelos de nível área, estar centrada na utili-ão de dados seccionais num período de tempo, a partir do final do século sado os modelos que combinam dados seccionais e cronológicos passa-ı a ser utilizados com maior frequência. Veja-se, por exemplo, os excelen-trabalhos de Ghosh e Rao (1994), Rao (1999*a*), Pfeffermann (2002) e Rao 03) para se ter uma ideia da investigação que tem sido feita neste domí-. Actualmente é possível encontrar na literatura várias extensões dos delos de nível área: modelo de Fay-Herriot multivariado, modelo com ɔs de sondagem correlacionados, modelos com dados seccionais e crono-icos e modelos espaciais que admitem a existência de correlações espa-is (dependentes da proximidade geográfica) entre os efeitos aleatórios de nínio. Dentro dos modelos seccionais e cronológicos, distinguem-se ainda nodelo de Rao-Yu e os modelos *state space*.

Na secção seguinte, apresenta-se uma breve revisão da literatura sobre delos que utilizam dados seccionais e cronológicos, e que se resumem a os especiais do modelo linear geral misto.

3.1. *Modelo de Rao-Yu*

Sejam θ_{dt} e $\hat{\theta}_{dt}$, respectivamente, um parâmetro populacional da variá-de interesse e a sua estimativa directa, e $x'_{dt} = (x_{dt1} \ldots x_{dtp})$ um vector p variáveis explicativas associadas ao pequeno domínio d no período $d = 1, \ldots, D$; $t = 1, \ldots, T$).

Rao e Yu (1992, 1994) propuseram uma extensão ao modelo de Fay-erriot, no âmbito da qual deduziram o estimador em dois passos através de ıa abordagem EBLUP. O seu modelo, conhecido por modelo de Rao-Yu, ısiste no modelo de erro de sondagem:

$$\hat{\theta}_{dt} = \theta_{dt} + \varepsilon_{dt} \tag{3.1}$$

de $\varepsilon_{dt}|\theta_{dt} \overset{ind}{\sim} N(0, \Sigma_d)$, com matriz de covariâncias, Σ_d, arbitrária mas nhecida; e no modelo de ligação:

$$\theta_{dt} = \mathbf{x}'_{dt}\boldsymbol{\beta} + \upsilon_d + u_{dt} \tag{3.2}$$

de $\boldsymbol{\beta}$ é um vector p-dimensional de parâmetros de regressão, os efeitos :atórios de domínio são $\upsilon_d \overset{iid}{\sim} N(0, \sigma_\upsilon^2)$ e se assume que os u_{dt}'s seguem ı processo auto-regressivo de primeira ordem [AR(1)] para cada d,

$$u_{dt} = \rho u_{d,t-1} + \xi_{dt}, \quad |\rho| < 1 \tag{3.3}$$

com $\xi_{dt} \overset{iid}{\sim} N(0, \sigma^2)$. Assume-se também que os ξ_{dt}, ε_{dt} e υ_d são independentes uns dos outros. O modelo combinado baseado em (3.1) e (3.2) é dado por:

$$\begin{aligned} \hat{\theta}_{dt} &= \mathbf{x}'_{dt}\boldsymbol{\beta} + \upsilon_d + u_{dt} + \varepsilon_{dt} \\ u_{dt} &= \rho u_{d,t-1} + \xi_{dt}, \quad |\rho| < 1 \end{aligned} \tag{3.4}$$

Rao e Yu (1992, 1994), procederam à ordenação das estimativas directas do parâmetro da variável de interesse, $\hat{\boldsymbol{\theta}} = col_{1\le d\le D}(\hat{\boldsymbol{\theta}}_d)$ e $\hat{\boldsymbol{\theta}}_d = col_{1\le t\le T}(\hat{\theta}_{dt})$, para que o modelo apresentado em (3.4) pudesse ser escrito na forma compacta como:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{u} + \boldsymbol{\varepsilon} \tag{3.5}$$

com $\mathrm{X} = col_{1\le d\le D}(\mathrm{X}_d)$, $\mathrm{X}_d = col_{1\le t\le T}(\mathrm{X}'_{dt})$, $\mathrm{Z} = \mathrm{I}_D \otimes 1_T$, $\mathrm{v} = col_{1\le d\le D}(\upsilon_d)$, $\mathbf{u} = col_{\substack{1\le t\le T \\ 1\le d\le D}}(u_{dt})$, $\boldsymbol{\varepsilon} = col_{\substack{1\le t\le T \\ 1\le d\le D}}(\varepsilon_{dt})$, onde $\mathbf{I}_D$, é uma matriz identidade de dimensão $D\times D$, $\mathbf{1}_T$ é um vector T-dimensional unitário, e $\otimes$ designa o produto directo. Estes autores assumiram que $\mathbf{v}$, $\mathbf{u}$ e $\boldsymbol{\varepsilon}$, são mutuamente independentes, têm média nula e que as suas matrizes de covariância são dadas por $V_m(\mathbf{v}) = \sigma_\upsilon^2\mathbf{I}_D$, $V_m(\mathbf{u}) = \sigma^2(\mathbf{I}_D \otimes \boldsymbol{\Gamma})$ e $V_m(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} = diag_{1\le d\le D}(\boldsymbol{\Sigma}_d)$, onde $\boldsymbol{\Gamma}$ é uma matriz de dimensão $T\times T$ com elementos $\gamma_{ij} = \rho^{|i-j|}/(1-\rho^2)$, i, j=1, ..., T. Com base nestes resultados e no modelo (3.5), Rao e Yu (1992, 1994) obtiveram $V_m(\hat{\boldsymbol{\theta}}) = \mathbf{V} = diag_{1\le d\le D}(\mathbf{V}_d)$, onde $\mathbf{V}_d = \boldsymbol{\Sigma}_d + \sigma^2\boldsymbol{\Gamma} + \sigma_\upsilon^2\mathbf{J}_T$ com $\mathbf{J}_T = \mathbf{1}_T\mathbf{1}'_T$.

Estes autores deduziram o melhor previsor linear não enviesado (*Best Linear Unbiased Predictor* – BLUP) de θ_{dt} a partir dos resultados gerais obtidos por Henderson (1975), uma vez que o modelo (3.4) é um caso especial do modelo linear geral misto. Assumindo que σ^2, σ_υ^2 e ρ são conhecidos, o BLUP de θ_{dt} é dado por:

$$\tilde{\theta}_{dt} = \mathbf{x}'_{dt}\tilde{\boldsymbol{\beta}} + \left(\sigma_\upsilon^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_t\right)'V_d^{-1}\left(\hat{\boldsymbol{\theta}}_d - \mathbf{X}_d\tilde{\boldsymbol{\beta}}\right) \tag{3.6}$$

$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ e γ_t é a t-ésima linha de Γ. O BLUP de θ_{dt} também ser escrito como um estimador combinado, sendo a soma rada do estimador directo, $\hat{\theta}_{dt}$, do estimador sintético, $\mathbf{x}'_{dt}\tilde{\boldsymbol{\beta}}$, e dos os $\left(\hat{\theta}_{dj} - \mathbf{x}'_{dj}\tilde{\boldsymbol{\beta}}\right)$, j=1, ..., T-1:

$$\tilde{\theta}_{dt} = w^*_{dt}\hat{\theta}_{dt} + \left(1 - w^*_{dt}\right)\mathbf{x}'_{dt}\tilde{\boldsymbol{\beta}} + \sum_{j=1}^{T-1} w^*_{dt}\left(\hat{\theta}_{dj} - \mathbf{x}'_{dt}\tilde{\boldsymbol{\beta}}\right) \qquad (3.7)$$

$\left(w^*_{d1} \ \ldots \ w^*_{dT}\right) = \left(\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_t\right)' V_d^{-1}$. Para o caso de um modelo AR(1) ɔ conhecido, Rao e Yu (1992, 1994) estimaram as componentes de cia, σ_v^2 e σ^2, através de uma extensão do método de Fuller e Battese). O EBLUP de θ_{dt} sob um modelo AR(1) com ρ conhecido, $\hat{\tilde{\theta}}_{dt}$, obtém-ɪavés da substituição em (3.6) de σ_v^2 e σ^2 por $\hat{\sigma}_v^2(\rho)$ e $\hat{\sigma}^2(\rho)$. Na ão em que ρ é desconhecido, Rao e Yu (1994) propuseram três métodos ʻeis para a sua estimação: um método baseado em conjecturas *a priori* o valor de ρ, um estimador pelo método dos momentos que ignora os de sondagem (inconsistente) e outro que não ignora os erros de sonda- consistente).

Em 1989, Choudhry e Rao tinham já proposto uma forma especial do lo de Rao-Yu para produzirem estimativas EBLUP para o desemprego ıl nas *census divisions*, utilizando dados do *Canadian Labour Force* y. Estes autores trataram os erros compósitos, $a_{dt} = u_{dt} + \varepsilon_{dt}$, como um sso AR(1) e assumiram que $\theta_{dt} = \mathbf{x}'_{dt}\boldsymbol{\beta} + \upsilon_d$. Existem outros trabalhos tilizaram o modelo de Rao-Yu modificado, mas que seguiram uma agem de estimação pelo método *Empirical Bayes* (EB) ou *Hierarchical* (HB). Segundo Rao (2003), os estimadores EB são idênticos aos ıdores EBLUP no caso em que se assume a normalidade no modelo misto.

Datta *et al.* (2002) e You (1999) obtiveram os estimadores EBLUP (EB) stimadores do erro quadrático médio (EQM) associados, corrigidos até ında ordem, para o modelo de Rao-Yu supondo um passeio aleatório os u_{dt}. Datta *et al.* (2002) usaram os métodos da máxima verosimi- a e da máxima verosimilhança restrita para estimarem as componentes riância, enquanto You (1999) utilizou o método dos momentos. Datta (2002) usaram o estimador EBLUP para estimarem o rendimento me- das famílias americanas com quatro pessoas, nos cinquenta estados canos e no distrito de Columbia, a partir dos dados do *Current Popu- Survey*.

3.2. Modelos *State Space*

Sejam θ_{dt} e $\hat{\theta}_{dt}$, respectivamente, um parâmetro populacional da variável de interesse e a sua estimativa directa, e $\mathbf{x}'_{dt} = \left(1 \ \ x_{dt1} \ \ \ldots \ \ x_{dtp}\right)$ um vector $(p+1)$-dimensional de variáveis explicativas associadas ao pequeno domínio d no período t (d=1, ..., D; t=1, ..., T).

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, $\boldsymbol{\beta}_{dt}$, obedecendo a um processo auto-regressivo. Pfeffermann e Burk (1990) propuseram um modelo geral com a seguinte forma:

$$\hat{\theta}_{dt} = \mathbf{x}'_{dt}\boldsymbol{\beta}_{dt} + \varepsilon_{dt} \qquad (3.8)$$

onde os coeficientes, $\boldsymbol{\beta}'_{dt} = \left(\beta_{dt0} \ \beta_{dt1} \ldots \beta_{dtp}\right)$, podem variar seccionalmente e cronologicamente, os erros da sondagem, ε_{dt}, são não correlacionados cronologicamente para cada d, e têm $E_m(\varepsilon_{dt}) = 0$ e $V_m(\varepsilon_{dt}) = \sigma_d^2$. A variação de $\boldsymbol{\beta}_{dt}$ ao longo do tempo é especificada pela seguinte equação de transição:

$$\begin{bmatrix} \beta_{dtj} \\ \beta_{dj} \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \beta_{d,t-1,j} \\ \beta_{dj} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{dtj}, \qquad j=0,\ 1,\ \ldots,\ p. \qquad (3.9)$$

Os $\boldsymbol{\beta}_{dj}$ são coeficientes fixos, $\mathbf{T}_j$ define uma matriz conhecida de dimensão 2×2 com (0,1) na segunda linha, e os erros do modelo, η_{dtj}, satisfazem $E_m(\eta_{dtj}) = 0$ e $E_m(\eta_{djt}\eta_{dlt}) = \sigma_{\eta,jl}$; 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 Burk (1990) consideraram ainda a possibilidade de existência de correlação contemporânea de um parâmetro entre dois domínios, formulada como $E_m(\eta_{djt}\eta_{d'jt}) = \sigma_{\eta,jl}\rho_j$; $d \neq d'$, j=0, 1, ..., p. 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. Pfeffermann e Burk (1990), apresentaram o modelo definido por (3.8) e (3.9) de forma compacta obedecendo à formulação clássica do modelo *state space*:

$$\hat{\boldsymbol{\theta}}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \varepsilon_t \qquad (3.10)$$

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t \qquad (3.11)$$

om $\hat{\boldsymbol{\theta}}_t = col_{1\le d\le D}(\hat{\theta}_{dt})$, $\mathbf{Z}_t = diag_{1\le d\le D}(\mathbf{Z}_{dt})$, $\mathbf{Z}_{dt} = (1\ 0\ x_{dt1}\ 0\ \ldots\ x_{dtp}\ 0)$, $\boldsymbol{\alpha}_t = col_{1\le d\le D}(\boldsymbol{\alpha}_{dt})$, $\boldsymbol{\alpha}'_{dt} = (\beta_{dt0}\ \beta_{d0}\ \beta_{dt1}\ \beta_{d1}\ \ldots\ \beta_{dtp}\ \beta_{dp})$, $\boldsymbol{\varepsilon}_t = col_{1\le d\le D}(\varepsilon_{dt})$, $\boldsymbol{\eta}_t = col_{1\le d\le D}(\boldsymbol{\eta}_{dt})$, $\boldsymbol{\eta}_{dt} = (\eta_{dt0}\ \eta_{dt1}\ \ldots\ \eta_{dtp})$, $\mathbf{T} = \mathbf{I}_D \otimes \tilde{\mathbf{T}}$, $\tilde{\mathbf{T}} = block\ diag_{1\le j\le p}(\mathbf{T}_j)$, $\mathbf{G} = I_D \otimes \tilde{\mathbf{G}}$ e $\tilde{\mathbf{G}} = \mathbf{I}_{p+1} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Assume-se que $\boldsymbol{\varepsilon}_t$ e $\boldsymbol{\eta}_t$ são não correlacionados ontemporaneamente e cronologicamente, têm média nula e as suas matrizes e covariância são dadas por $E_m(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t) = \boldsymbol{\Sigma}_t = diag_{1\le d\le D}(\sigma_d^2)$ e $E_m(\boldsymbol{\eta}_t\boldsymbol{\eta}'_t) = \boldsymbol{\Lambda}$ nde $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_{dd'}]$, d, d' = 1, ..., D, com $\boldsymbol{\Lambda}_{dd'} = \begin{cases} E_m(\boldsymbol{\eta}_{dt}\boldsymbol{\eta}'_{dt}) & ,d = d' \\ diag_{1\le j\le p}(\sigma_{\eta,jj}\rho_j) & ,d \ne d' \end{cases}$.

Pfeffermann e Burk (1990) fizeram a estimação dos coeficientes de egressão do modelo *state space* (3.11) e (3.12), através de um filtro de Kalman. No caso em que os coeficientes de regressão seguem um passeio leatório, estes autores deduziram que o BLUP de θ_{dt}, $\tilde{\theta}_{dt} = \tilde{\theta}_{dt}^H$, pode ser scrito como um estimador combinado, sendo uma soma ponderada do stimador directo, $\hat{\theta}_{dt}$, do estimador sintético, $\mathbf{x}'_{dt}\tilde{\boldsymbol{\beta}}_{dt|t-1}$, e dos "factores de justamento" $(\hat{\theta}_{d't} - \mathbf{x}'_{d't}\tilde{\boldsymbol{\beta}}_{d't|t-1})$, $d \ne d'$, onde $\tilde{\boldsymbol{\beta}}_{dt|t-1}$ é o estimador de $\boldsymbol{\beta}_{dt}$ com ase em toda a informação disponível até ao período $t-1$. Quando os parâmetros desconhecidos das matrizes de variâncias-covariâncias, $\boldsymbol{\delta}$, são substituídos pelos seus estimadores, $\hat{\boldsymbol{\delta}}$, obtém-se o EBLUP, $\hat{\tilde{\theta}}_{dt} = \hat{\tilde{\theta}}_{dt}^H$.

Na literatura sobre esta matéria, encontram-se outros trabalhos onde são tilizados modelos do tipo *state space*. Ghosh e Nangia (1993) e Ghosh *et l.* (1996) também propuseram modelos do tipo *state space* para estimarem, egundo uma abordagem bayesiana, o rendimento mediano das famílias mericanas com quatro pessoas, nos cinquenta estados americanos e no distrito de Columbia. Pfeffermann, Feder e Signorelli (1998) aplicaram um nodelo deste tipo aos dados da força de trabalho na Austrália.

. Modelo proposto de nível área com dados cronológicos

O quadro do modelo linear geral misto apresenta-se bastante adequado ara representar realidades que possam ser descritas através de dados de ıatureza mista (seccional e cronológica). Neste contexto, encontram-se o nodelo de Rao-Yu e os modelos *state space*. No primeiro especifica-se que s efeitos aleatórios, u_{dt}, seguem um processo auto-regressivo de primeira rdem ou um passeio aleatório, enquanto nos segundos é necessário recorrer

a uma equação de transição para especificar completamente o modelo. Contudo, 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 e cronológica. Deste modo, parece potencialmente interessante explorar toda a flexibilidade oferecida pelo modelo linear geral misto para representar essas realidades, não só através da inclusão de efeitos fixos e de efeitos aleatórios no modelo, mas também através da especificação de estruturas de covariância arbitrárias sobre os efeitos aleatórios do modelo e/ou sobre as variáveis residuais. A abordagem proposta em seguida para estimar parâmetros em pequenos domínios tenta explorar toda essa flexibilidade.

Esta forma 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 (Coelho, 2000).

4.1. *Especificação do modelo*

Na especificação do modelo supõe-se que as observações disponíveis resultam de um inquérito longitudinal (com ou sem rotação) realizado ao longo de T momentos. Supõe-se que as unidades estatísticas de análise podem ser agrupadas em D domínios, nível a que se pretende fazer inferência. Por sua vez, os domínios podem ser agrupados em A regiões e às quais estão referenciados os efeitos fixos do modelo. O número de domínios contidos numa região a representa-se por $D(a)$, sendo o número total de domínios igual a $D = \sum_{a=1}^{A} D(a)$.

Seja $\hat{\theta}_{adt}$ o estimador directo de um parâmetro da variável de interesse no d-ésimo pequeno domínio pertencente à região a no período t, designado por θ_{adt} (a=1, ..., A; d=1, ..., $D(a)$; t=1, ..., T), que se assume estar disponível sempre que a dimensão amostral no pequeno domínio é não nula, $n_{adt} \ge 1$. Assume-se que $\hat{\theta}_{adt}$ é um estimador não enviesado no desenho de θ_{adt}, isto é,

$$\hat{\theta}_{adt} = \theta_{adt} + \varepsilon_{adt} \qquad (4.1)$$

onde os ε_{adt} são os erros da sondagem associados ao domínio ad, no período t, com média nula, dado θ_{adt}. Os erros da sondagem associados a diferentes unidades são não correlacionados entre si, podendo, no entanto, ser heteroce-

ásticos. Assume-se que os erros da sondagem são conhecidos. Assume-se ambém que existe um vector de p variáveis explicativas, $\mathbf{x}_{adt} = (x_{adt1}, x_{adt2}, ..., x_{adtp})'$ para cada um dos domínios pertencente à egião a no período t. Neste contexto, propõe-se o seguinte modelo que ıtiliza informação de natureza seccional e cronológica, o qual assume que θ_{adt} está relacionado com $\mathbf{x}'_{adt}$ através de um modelo linear com efeitos leatórios:

$$\theta_{adt} = \mathbf{x}'_{adt}\boldsymbol{\beta}_{at} + u_{adt} \tag{4.2}$$

nde $\boldsymbol{\beta}_{at} = (\beta_{at1} \ \ldots \ \beta_{atp})'$ é um vector de p efeitos fixos referenciados à egião a no período t e os u_{adt} são os efeitos aleatórios associados ao d-ésimo lomínio pertencente à região a, no período t. Assume-se que os efeitos leatórios têm média nula e que efeitos aleatórios associados a diferentes lomínios são não correlacionados entre si. Contudo, supõe-se que os efeitos leatórios associados a um determinado domínio poderão apresentar uma :strutura de covariância cronológica. Algumas estruturas passíveis de serem ıplicadas neste contexto são apresentadas na secção 4.2.

Combinando (4.1) com o modelo (4.2), obtém-se uma generalização do nodelo de Fay-Herriot (1979) para dados seccionais e cronológicos:

$$\hat{\theta}_{adt} = \mathbf{x}'_{adt}\boldsymbol{\beta}_{at} + u_{adt} + \varepsilon_{adt} \tag{4.3}$$

nde $E_m(u_{adt}) = 0$, $Cov_m(u_{adt}, u_{a'd't'}) = \begin{cases} \sigma_{u,att'} & , \ a = a', d = d' \\ 0 & , \ caso\ contrário \end{cases}$,

$E_d(\varepsilon_{adt} | \theta_{adt}) = 0$, $V_d(\varepsilon_{adt} | \theta_{adt}) = \sigma^2_{\varepsilon,adt}$ e $E_m(u_{adt}\varepsilon_{adt}) = 0$.

Este modelo com dados seccionais e cronológicos constitui uma exten- são ao modelo de nível área básico. Neste modelo, quer o vector dos efeitos fixos, quer a covariância cronológica dos efeitos aleatórios, poderão variar com a região em que se encontra cada domínio. Como tal, convenciona-se denominá-lo por modelo totalmente regionalizado. Pode considerar-se um caso particular deste modelo, designado por modelo só com efeitos fixos regionalizados, no qual a covariância cronológica dos efeitos aleatórios não varia com a região, dependendo apenas do afastamento entre os momentos t e t'.

O modelo (4.3) pode ser escrito na forma matricial como um caso particular do modelo linear geral misto. Para tal, é necessário ordenar as

estimativas directas, $\hat{\theta}_{adt}$: $\hat{\boldsymbol{\theta}} = col_{1\le a\le A}(\hat{\boldsymbol{\theta}}_a)$, $\hat{\boldsymbol{\theta}}_a = col_{1\le d\le D(a)}(\hat{\boldsymbol{\theta}}_{ad})$, $\hat{\boldsymbol{\theta}}_{ad} = col_{1\le t\le T}(\hat{\theta}_{adt})$. O modelo pode agora ser escrito da seguinte forma (Pereira, 2005):

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}, \tag{4.4}$$

onde $\mathbf{X} = diag_{1\le a\le A}(\mathbf{X}_a)$, $\mathbf{X}_a = col_{1\le d\le D(A)}(\mathbf{X}_{ad})$, $\mathbf{X}_{ad} = diag_{1\le t\le T}(\mathbf{x}'_{adt})$, $\boldsymbol{\beta} = col_{1\le a\le A}(\boldsymbol{\beta}_a)$, $\boldsymbol{\beta}_a = col_{1\le t\le T}(\boldsymbol{\beta}_{at})$, $\mathbf{u} = col_{1\le a\le A}(\mathbf{u}_a)$, $\mathbf{u}_a = col_{1\le d\le D(a)}(\mathbf{u}_{ad})$, $\mathbf{u}_{ad} = col_{1\le t\le T}(\mathbf{u}_{adt})$, $\boldsymbol{\varepsilon} = col_{1\le a\le A}(\boldsymbol{\varepsilon}_a)$, $\boldsymbol{\varepsilon}_a = col_{1\le d\le D(a)}(\boldsymbol{\varepsilon}_{ad})$ e $\boldsymbol{\varepsilon}_{ad} = col_{1\le t\le T}(\varepsilon_{adt})$.

Assume-se que $E_m(\mathbf{u}) = 0$, $V_m(\mathbf{u}) = \mathbf{G} = diag_{\substack{1\le d\le D(a)\\ 1\le a\le A}}[\mathbf{G}_{ad}]$, onde $\mathbf{G}_{ad}$ é uma matriz simétrica com elementos $\sigma_{u,att'}$, t,t'=1, ..., T, de dimensão $T \times T$, que define a estrutura de variâncias-covariâncias cronológicas dos efeitos aleatórios associados ao d-ésimo domínio pertencente à região a. Assume-se também que $E_m(\boldsymbol{\varepsilon}) = \mathbf{0}$ e que $V_m(\boldsymbol{\varepsilon}) = \mathbf{R} = diag_{\substack{1\le d\le D(a)\\ 1\le a\le A}}(\mathbf{R}_{ad})$, onde $\mathbf{R}_{ad} = diag_{1\le t\le T}(\sigma^2_{\varepsilon,adt})$ é uma matriz de dimensão $T \times T$ das variâncias dos erros da sondagem. Assume-se ainda que $E_m(\mathbf{u}\boldsymbol{\varepsilon}') = \mathbf{0}$. A matriz de variâncias-covariâncias de $\hat{\boldsymbol{\theta}}$, de dimensão $TD \times TD$, é dada por $V_m(\hat{\boldsymbol{\theta}}) = \mathbf{V} = diag_{\substack{1\le d\le D(a)\\ 1\le a\le A}}(\mathbf{V}_{ad})$, onde $\mathbf{V}_{ad} = \mathbf{G}_{ad} + \mathbf{R}_{ad}$. Tal como foi definido acima, o modelo (4.4) é um caso particular do modelo linear geral misto com efeitos fixos e com efeitos aleatórios, com $\mathbf{Z} = \mathbf{I}_{TD}$ e estrutura de variâncias--covariâncias diagonal por blocos.

4.2. *Especificação das estruturas de covariância*

A flexibilidade da classe de modelos proposta recomenda que em cada caso concreto se proceda à selecção de um modelo específico, isto é, à escolha das variáveis explicativas e de adequadas estruturas de covariância. Deste modo, uma das etapas fundamentais da modelação consiste precisamente na especificação das estruturas de covariância.

Basta rever, por exemplo, Wolfinguer (1996) para se ter uma ideia da grande quantidade de estruturas de covariância cronológica existentes. Todas essas estruturas podem ser utilizadas na classe de modelos proposta, incluindo covariâncias arbitrárias não estruturadas, dada a abrangência do modelo especificado em (4.3).

As estruturas de covariância mais extremas em termos do número de âmetros são, por um lado, a não estruturada, e por outro, as estruturas *npound symmetry* e *first-order autoregressive*. Enquanto a primeira estru-a, a não estruturada, requer a estimação de um grande número de âmetros, $(T+1)T/2$, e consequentemente necessita de amostras de gran-; dimensões, as duas últimas estruturas requerem apenas a estimação de s parâmetros. Por outro lado, para além das estruturas de covariância *npound symmetry* e *first-order autoregressive* assumirem variâncias cons-tes, a estrutura *compound symmetry* assume também covariâncias constan-. Visto que, em geral, as dimensões amostrais são pequenas e não se espera e as duas estruturas mais parcimoniosas sejam apropriadas para os dados ıgitudinais, especialmente devido à hipótese da homogeneidade das variân-ıs, devem considerar-se outras estruturas de covariância que não sejam tão ıgentes. Deste modo, no caso em estudo, propõe-se a utilização de uma das guintes estruturas de covariância cronológica heterocedástica: *heteroge-ous first-order autoregressive* [ARH(1)]

$$\sigma_{u,att'} = \begin{cases} \sigma^2_{u,at} & ,t = t' \\ \sigma_{u,at}\sigma_{u,at'}\rho^{|t-t'|} & ,t \neq t' \end{cases}$$

heterogeneous compound symmetry (CSH),

$$\sigma_{u,att'} = \begin{cases} \sigma^2_{u,at} & ,t = t' \\ \sigma_{u,at}\sigma_{u,at'}\rho & ,t \neq t'' \end{cases}$$

nbas requerendo a estimação de $A(T+1)$ parâmetros.

4.3. *O melhor previsor linear não enviesado*

O valor do parâmetro da variável de interesse num pequeno domínio *ad* o período *t*, $\theta_{adt} = \mathbf{x}'_{adt}\boldsymbol{\beta}_{at} + u_{adt}$, é um caso especial da combinação linear $= \mathbf{k}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u}$, onde $\mathbf{k}' = \left({}_1\mathbf{0}_{[(a-1)T+(t-1)]p} \quad \mathbf{x}'_{adt} \quad {}_1\mathbf{0}_{[(A-a)T+(T-t)]p}\right)$, sendo ${}_i\mathbf{0}_j$ uma ıatriz nula de dimensão $i \times j$ e $\mathbf{m}' = (0 \ldots 0\ 1\ 0 \ldots 0)$ um vector DT-dimen-ional com 1 na (adt)-ésima posição e zeros nas outras posições. Note-se que modelo (4.3) escrito na forma compacta é um caso especial do modelo near geral misto, donde o BLUP de $\tau = \theta_{adt}$ pode ser obtido a partir dos esultados gerais deduzidos por Henderson (1975).

Assumindo que as componentes de variância são conhecidas, ou seja, que as matrizes **G** e **R** são conhecidas, então o BLUP de τ é dado por:

$$\tilde{\tau} = \mathbf{k}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\tilde{\mathbf{u}} \tag{4.5}$$

onde $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ é o melhor estimador linear não enviesado (BLUE) de $\boldsymbol{\beta}$, e $\tilde{\mathbf{u}} = \mathbf{G}\mathbf{V}^{-1}\left(\hat{\boldsymbol{\theta}} - X\tilde{\boldsymbol{\beta}}\right)$ é o melhor previsor linear não enviesado de **u**. Utilizando as estruturas $\mathbf{k}'$, $\mathbf{m}'$, **G** e **V**, demonstra-se que (4.5) se reduz a (Pereira, 2005):

$$\tilde{\theta}_{adt} = \mathbf{X}_{adt}\tilde{\boldsymbol{\beta}}_a + \mathbf{w}'_{adt}\left(\hat{\boldsymbol{\theta}}_{ad} - \mathbf{X}_{ad}\tilde{\boldsymbol{\beta}}_a\right), \tag{4.6}$$

onde $\mathbf{X}_{adt} = \left({}_1\mathbf{0}_{(t-1)p} \quad \mathbf{x}'_{adt} \quad {}_1\mathbf{0}_{(T-t)p}\right)$, $\tilde{\boldsymbol{\beta}}_a = \left[\sum_{d=1}^{D(a)}\mathbf{X}'_{ad}\mathbf{V}^{-1}_{ad}\mathbf{X}_{ad}\right]^{-1}\left[\sum_{d=1}^{D(a)}\mathbf{X}'_{ad}\mathbf{V}^{-1}_{ad}\hat{\boldsymbol{\theta}}_{ad}\right]$, e $\mathbf{w}'_{adt}$ é um vector de dimensão T, correspondente à t-ésima linha da matriz $\mathbf{G}_{ad}\mathbf{V}^{-1}_{ad}$.

O estimador obtido pode ser classificado como um estimador combinado, uma vez que pode ser decomposto em duas componentes: um estimador sintético, $\mathbf{X}_{adt}\tilde{\boldsymbol{\beta}}_a$, e um factor de correcção, $\mathbf{w}'_{adt}\left(\hat{\boldsymbol{\theta}}_{ad} - \mathbf{X}_{ad}\tilde{\boldsymbol{\beta}}_a\right)$, que é uma função das diferenças entre as estimativas directas e as estimativas sintéticas do parâmetro de interesse. Pode-se afirmar que os pesos reflectidos em $\mathbf{w}'_{adt}$ permitem que o estimador sintético seja corrigido pelos erros de previsão do domínio que é alvo de inferência no período *t*. A partir da expressão anterior é possível observar que quando um determinado domínio não está representado na amostra da *t*-ésima vaga, continua a ser possível fazer previsões para o factor de correcção associado a esse domínio, tirando partido da sua potencial autocorrelação cronológica, desde que existam observações em pelo menos uma das vagas anteriores. Esta é sem dúvida uma característica muito apelativa do estimador proposto: é possível evitar que o estimador proposto se reduza a um estimador sintético "puro", mesmo quando a dimensão amostral observada no período *t* no domínio que é alvo de inferência é nula.

Note-se que o parâmetro populacional θ_{adt} referenciado ao nível do domínio *ad* no período *t*, expresso em função de efeitos fixos e de efeitos aleatórios com dados ao nível do subdomínio *adc* no período *t*,

$$\theta_{adt} = \mathbf{x}'_{adt}\boldsymbol{\beta}^c_{at} + \sum_{c=1}^{C(ad)} p_{adc}u^c_{adct} \tag{4.7}$$

ontinua a ser um caso especial da combinação linear $\tau = \mathbf{k}'^{*}\boldsymbol{\beta}^{c} + \mathbf{m}'^{*}\mathbf{u}^{c}$.)s vectores $\boldsymbol{\beta}^{c}$ e $\mathbf{u}^{c}$ são, respectivamente, os vectores dos efeitos fixos e dos feitos aleatórios estimados com um modelo referenciado ao nível de ubdomínio, $\mathbf{k}'^{*} = \left({}_{1}\mathbf{0}_{[(a-1)T+(t-1)]p} \quad \mathbf{x}'_{adt} \quad {}_{1}\mathbf{0}_{[(A-a)T+(T-t)]p}\right)$ é um vector Tap-dimen-

ional, $\mathbf{m}'^{*} = \left({}_{1}\mathbf{0}_{\left[\sum_{a=1}^{a-1} C(a) + \sum_{d=1}^{d-1} C(ad)\right]T + C(ad)(t-1)} \quad \ldots \quad \mathbf{m}'^{c} \quad \ldots \quad {}_{1}\mathbf{0}_{\left[\sum_{a=a+1}^{A} C(a) + \sum_{d=d+1}^{D(a)} C(ad)\right]T + C(ad)(T-t)} \right)$

um vector CT-dimensional e $\mathbf{m}^{c} = col_{1 \le c \le C(ad)}(p_{adc})$ é um vector $C(ad)$-dimensional com o peso de cada subdomínio dentro do respectivo domínio, $_{adc}$. O número de subdomínios contidos no d-ésimo domínio da região a epresenta-se por $C(ad)$, sendo o número total de subdomínios da região a gual a $C(a) = \sum_{d=1}^{D(a)} C(ad)$ e o número total de subdomínios igual a $C = \sum_{a=1}^{A}\sum_{d=1}^{D(a)} C(ad)$.

Neste caso também é possível deduzir o BLUP de $\tau = \theta_{adt}$ a partir dos esultados gerais deduzidos por Henderson (1975), e de forma em tudo idêntica à seguida para o caso do modelo com efeitos aleatórios referenciados ao ível de domínio, dado que o modelo (4.8) escrito na forma compacta continua a ser um caso especial do modelo linear geral misto.

4.4. *O melhor previsor linear não enviesado empírico*

O BLUP apresentado em (4.6), $\tilde{\theta}_{adt}$, depende de $(T+1)$ parâmetros $\boldsymbol{\delta}_a = \left(\rho_a, \sigma^2_{u,a1}, \ldots, \sigma^2_{u,aT}\right)'$, que são geralmente desconhecidos nas aplicações práticas. O previsor em dois passos (BLUP empírico ou EBLUP) obtém-se através da substituição dos parâmetros desconhecidos, $\boldsymbol{\delta}_a$, por estimadores assimptoticamente consistentes $\hat{\boldsymbol{\delta}}_a = \left(\hat{\rho}_a, \hat{\sigma}^2_{u,a1}, \ldots, \hat{\sigma}^2_{u,aT}\right)'$ na expressão do BLUP:

$$\hat{\tilde{\theta}}_{adt} = \hat{\tilde{\theta}}^{H}_{adt}\left(\hat{\boldsymbol{\delta}}_a\right) = \mathbf{X}_{adt}\hat{\boldsymbol{\beta}}_a + \hat{\mathbf{h}}'_{adt}\left(\hat{\boldsymbol{\theta}}_{ad} - \mathbf{X}_{ad}\hat{\boldsymbol{\beta}}_a\right). \tag{4.8}$$

No EBLUP (4.8), $\hat{\boldsymbol{\beta}}_a$ e $\hat{\mathbf{h}}'_{adt}$ correspondem aos valores de $\tilde{\boldsymbol{\beta}}_a(\boldsymbol{\delta}_a)$ e de $\mathbf{h}'_{adt}(\boldsymbol{\delta}_a)$, quando $\boldsymbol{\delta}_a$ é substituído por $\hat{\boldsymbol{\delta}}_a$.

Existem diversos métodos que podem ser usados para estimar de forma onsistente as componentes de variância num modelo linear geral misto.)s métodos mais conhecidos são os métodos de verosimilhança que se

baseiam nos pressupostos da normalidade de $\mathbf{u}$ e $\boldsymbol{\varepsilon}$, o método dos momentos (Fuller & Battese, 1973) ou a estimação MIVQUE (Rao, 1970, 1972; Wolfinguer, Tobias e Sall, 1994). Algumas referências aos métodos de verosimilhança podem ser encontradas em Amemiya (1971), Hartley e Rao (1967), Harville (1977), Laird e Ware (1982) e Jennrich e Schluchter (1986). Neste estudo, propõe-se estimar o parâmetro ρ_a através de simulação e as componentes de variância, $\sigma^2_{\varepsilon,adt}$ e $\sigma^2_{u,at}$, através do método proposto por Prasad e Rao (1999). Todos estes parâmetros foram estimados externamente ao modelo, dado que representam as variâncias que resultam do desenho amostral adoptado e que não dependem do modelo postulado.

5. Selecção de modelos e avaliação da qualidade da estimação

O diagnóstico e a selecção de modelos, no que se refere à comparação de estruturas de covariância alternativas, é efectuada com base no critério de informação de Akaike (AIC). Segundo Bozdogan (1987), este procedimento de selecção de modelos, que consiste em minimizar o critério de informação de máxima verosimilhança restrita, penaliza os modelos com um grande número de parâmetros de covariância. Nos casos em que os parâmetros de covariância são estimados externamente ao modelo, o valor do AIC não depende do número de parâmetros de covariância.

Segundo Coelho (2000), para estimadores do tipo *model-based*, a avaliação da qualidade da estimação pode ser realizada com base em estimativas do EQM *model-based* dos estimadores propostos. Segundo este autor, no quadro dos estimadores combinados, as estimativas do EQM *model-based* apresentam-se como medidas realistas da precisão do estimador, porque o enviesamento do estimador estará parcial ou totalmente representado nos erros de previsão dos efeitos aleatórios. A dedução do EQM *model-based* dos estimadores propostos é suportada pela teoria do modelo linear geral misto. No espaço de inferência restrito, e desprezando a incerteza resultante da estimação das componentes de variância, deduz-se que o EQM *model-based* do erro de previsão de $E(\theta_{adt} \mid \mathbf{u}) = \mathbf{x}'_{adt}\boldsymbol{\beta}_{at} + u_{adt}$ é dado por (Coelho, 2000):

$$\begin{aligned} & E\left\{\left[\tilde{E}(\theta_{adt} \mid \mathbf{u}) - E(\theta_{adt} \mid \mathbf{u})\right]\left[\tilde{E}(\theta_{adt} \mid \mathbf{u}) - E(\theta_{adt} \mid \mathbf{u})\right]'\right\} = \\ & = \mathbf{m}'\left(\mathbf{R}^{-1} + \mathbf{G}^{-1}\right)^{-1}\mathbf{m} + \left(\mathbf{m}'\mathbf{G}\mathbf{V}^{-1}\mathbf{X} - \mathbf{k}'\right)\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{m}'\mathbf{G}\mathbf{V}^{-1}\mathbf{X} - \mathbf{k}'\right)' \end{aligned} \tag{5.1}$$

Estudo empírico

A precisão relativa das estimativas produzidas pelo EBLUP proposto foi mparada com a precisão relativa das estimativas produzidas pelo estimador recto pós-estratificado, para um conjunto de dados relativos a cinco trimes-s (Janeiro de 2002 a Março de 2003). Os valores da variável explicada do odelo (4.3) correspondem às estimativas directas do preço médio de nsacção da habitação (apartamentos ou moradias) por metro quadrado de ea útil, obtidas a partir dos dados recolhidos pelo IPTH. Este inquérito mestral conduzido pelo INE, é aplicado ao universo de empresas de me-ação imobiliária sedeadas em Portugal continental que transaccionam pré-os urbanos. A base de sondagem existente é formada por 4671 empresas de ediação imobiliária. Neste inquérito é utilizada uma sondagem aleatória tratificada de uma população de conglomerados. A população de conglo-erados, formada pelas empresas de mediação imobiliária, foi estratificada elo cruzamento de três variáveis: valor de volume de negócios; NUTS III concelho (apenas para as áreas Metropolitanas de Lisboa e do Porto). Todas transacções realizadas por um conglomerado seleccionado para a amostra io observadas.

A única variável auxiliar utilizada na regressão foi o preço médio de valiação bancária da habitação, obtida de forma administrativa através do ABH.

Com este modelo pretende-se estimar os preços médios de transacção a habitação (incluindo apartamentos e moradias), dos apartamentos e das noradias ao nível de desagregação NUTS III e Concelho. Foram produzidas stimativas desses preços médios utilizando dados ao nível de desagregação e NUTS III e de concelho, e estruturas de covariância CSH e ARH(1). Para ada uma das estimações anteriores foram utilizados os seguintes modelos: nodelo totalmente regionalizado (Modelo I), modelo só com efeitos fixos egionalizados (Modelo II) e modelo não regionalizado (Modelo III).

Antes da estimação, foram excluídos todos os outliers severos, separa-amente para cada NUTS II, dos preços de transacção da habitação e dos reços de avaliação bancária da habitação por metro quadrado de área útil. oi também utilizado o método de classificação hierárquica ascendente, ba-eado no critério de Ward, que permitiu a partição das vinte e oito NUTS III m quatro classes (regiões) homogéneas. Com este agrupamento pretende bter-se ganhos de precisão na estimação, através da introdução de efeitos ixos no modelo que variem não só temporalmente, mas também seccional-nente ao nível de região, e através da especificação de estruturas de cova-iância diferentes para cada uma das regiões. Se existirem regiões relativa-nente homogéneas, de tal forma que os pequenos domínios pertencentes a

uma mesma classe sejam semelhantes e as regiões bem separadas, então é de admitir que os efeitos aleatórios de cada região possam apresentar a sua própria estrutura de covariância.

Neste caso, $A = 4$; $d(1) = 16$, $d(2) = 6$, $d(3) = 4$, $d(4) = 2$ e $T = 5$; $\hat{\theta}_{adt}$ é a estimativa do preço médio de transacção da habitação (apartamento ou moradia) por metro quadrado referente ao domínio *ad* (NUTS III), no trimestre t (a=1, 2, 3, 4; d=1,..., $D(a)$; t=1, ..., 5); e $x_{ad,t-1}$ é o preço médio de avaliação bancária da habitação (respectivamente, apartamento ou moradia) por metro quadrado referente ao domínio *ad* (NUTS III), no trimestre t-1. O desfasamento temporal existente na variável auxiliar deve-se ao facto das avaliações bancárias das habitações se realizarem, na sua grande maioria, dois a quatro meses antes da sua transacção. Este desfasamento temporal na variável auxiliar permite produzir, no trimestre t, previsões do preço médio de transacção da habitação para o trimestre t+1.

Os três modelos utilizados na estimação são os seguintes:

Modelo I: $$\hat{\theta}_{adt} = \beta_{at1} + x_{ad,t-1}\beta_{at2} + u_{adt} + \varepsilon_{adt}. \qquad (6.1)$$

onde $E_m(u_{adt}) = 0$, $Cov_m(u_{adt}, u_{a'd't'}) = \begin{cases} \sigma_{u,att'} & , \ a = a', d = d' \\ 0 & , \ \textit{caso contrário} \end{cases}$,

$E_d(\varepsilon_{adt}|\theta_{adt}) = 0$, e $V_d(\varepsilon_{adt}|\theta_{adt}) = \sigma^2_{\varepsilon,adt}$, $E_m(u_{adt}\varepsilon_{adt}) = 0$;

Modelo II: $$\hat{\theta}_{adt} = \beta_{at1} + x_{ad,t-1}\beta_{at2} + u_{adt} + \varepsilon_{adt}. \qquad (6.2)$$

onde $E_m(u_{adt}) = 0$, $Cov_m(u_{adt}, u_{a'd't'}) = \begin{cases} \sigma_{u,tt'} & , \ a = a', d = d' \\ 0 & , \ \textit{caso contrário} \end{cases}$,

$E_d(\varepsilon_{adt}|\theta_{adt}) = 0$, e $V_d(\varepsilon_{adt}|\theta_{adt}) = \sigma^2_{\varepsilon,adt}$, $E_m(u_{adt}\varepsilon_{adt}) = 0$;

Modelo III: $$\hat{\theta}_{dt} = \beta_{t1} + x_{d,t-1}\beta_{t2} + u_{dt} + \varepsilon_{dt}. \qquad (6.3)$$

onde $E_m(u_{dt}) = 0$, $Cov_m(u_{dt}, u_{d't'}) = \begin{cases} \sigma_{u,tt'} & , \ d = d' \\ 0 & , \ \textit{caso contrário} \end{cases}$,

$E_d(\varepsilon_{dt}|\theta_{dt}) = 0$, $V_d(\varepsilon_{dt}|\theta_{dt}) = \sigma^2_{\varepsilon,dt}$ e $E_m(u_{dt}\varepsilon_{dt}) = 0$.

A estimação do parâmetro ρ foi feita através de simulação. Em cada um dos modelos, fez-se variar este parâmetro entre 0,00 e 0,99, de 0,01 em 0,01, e escolheu-se o valor de ρ que minimiza o valor do AIC. Verificou-se em todas as situações que o valor do AIC era mínimo quando $\rho = 0{,}99$, sendo este o valor utilizado.

As estimativas do preço médio de transacção da habitação para cada NUTS III (domínio de estudo), produzidas a partir do modelo totalmente regionalizado com dados de concelho são obtidas da seguinte forma:

$$\hat{\tilde{\theta}}_{adt} = \tilde{\beta}_{at1} + x_{ad,t-1}\tilde{\beta}_{at2} + \sum_{c=1}^{C(ad)} \frac{n_{aval,adct}}{n_{aval,adt}} \tilde{u}_{adct}, \qquad (6.4)$$

onde $\tilde{\beta}_{at1}$ e $\tilde{\beta}_{at2}$ são as estimativas dos efeitos fixos associados à região a, no trimestre t; $n_{aval,adct}$ é o número de avaliações bancárias da habitação observadas no c-ésimo concelho pertencente ao domínio ad (NUTS III), no trimestre t; $n_{aval,adt}$ é o número de avaliações bancárias da habitação observadas no domínio ad (NUTS III), no trimestre t; e $\tilde{u}_{adct}$ é a estimativa do efeito aleatório associado ao c-ésimo concelho pertencente ao domínio ad (NUTS III), no trimestre t.

TABELA 1
Valor do AIC em cada um dos modelos

	Estrutura de Covariância	
	ARH(1)	CSH
Dados de NUTS III		
Modelo I	1342,8	1342,5
Modelo II	1349,0	1349,0
Modelo III	1551,8	1551,7
Dados de Concelho		
Modelo I	8713,1	8707,6
Modelo II	8783,7	8784,8
Modelo III	8967,8	8972,0

O diagnóstico e a selecção de modelos, no que se refere à comparação de estruturas de covariância alternativas, foi efectuada com base no critério de informação de Akaike. Apesar dos valores do AIC não apresentarem diferenças significativas entre os modelos com estrutura de covariância ARH(1) e os modelos com estrutura de covariância CSH, verificou-se na maior parte dos casos que os últimos modelos apresentam valores menores no critério de

informação de máxima verosimilhança restrita. Desta forma, decidiu-se utilizar uma estrutura de covariância do tipo CSH.

A avaliação da qualidade da estimação foi suportada pelas estimativas do EQM *model-based* do erro de previsão de $E(\theta_{adt} \mid \mathbf{u}) = \mathbf{x}'_{adt}\boldsymbol{\beta}_{at} + u_{adt}$. Com base nessas estimativas do EQM, foram calculadas estimativas da precisão relativa (PR) das estimativas do preço médio de transacção da habitação, dos apartamentos e das moradias nos domínios de estudo, para um nível de confiança de 95%.

A selecção do tipo de modelo (não regionalizado, só com efeitos fixos regionalizados, totalmente regionalizado) e do nível de desagregação dos dados (NUTS III, Concelho), foi efectuada com base no AIC e na PR média das estimativas.

TABELA 2
Valores do AIC e da PR média (entre parêntesis) em cada um dos modelos

	Modelo I	Modelo II	Modelo III
Habitação			
Dados de NUTS III	1342,5 (6,4%)	1349,0 (11,3%)	1551,7 (9,7%)
Dados de Concelho	8707,6 (5,6%)	8784,8 (8,2%)	8972,0 (6,9%)
Apartamentos			
Dados de NUTS III	1277,3 (6,1%)	1311,7 (12,0%)	1511,3 (10,0%)
Dados de Concelho	7618,2 (5,4%)	7679,5 (8,5%)	7860,4 (7,1%)
Moradias			
Dados de NUTS III	1363,4 (24,4%)	1362,8 (30,1%)	1612,1 (23,6%)
Dados de Concelho	6405,6 (18,4%)	6443,7 (18,9%)	6659,3 (14,8%)

A análise da tabela 2 permite concluir, em primeiro lugar, que os estimadores combinados propostos não permitem estimar com precisão aceitável o preço médio de transacção das moradias, ao nível de NUTS III. Apesar destas estimativas apresentarem uma melhor precisão relativa média do que as estimativas produzidas pelo estimador directo pós-estratificado, elas apresentam valores de precisão relativa compreendidos entre os 9% e os 53%. Estes níveis inaceitáveis de precisão relativa das estimativas são justificados pelas pequenas dimensões amostrais (ou nulas, em muitos casos) de transacções de moradias observadas dentro de cada domínio de estudo.

No que se refere à estimação do preço médio de transacção da habitação e dos apartamentos, ao nível de NUTS III, verifica-se que os modelos totalmente regionalizados são aqueles que apresentam os menores valores no AIC e os melhores níveis de precisão relativa das estimativas produzidas. A este nível de desagregação, verifica-se também que a qualidade das estimativas produzidas através de estimadores combinados que utilizam dados de conce-

lho é melhor. Perante estas observações, escolhe-se o estimador combinado deduzido a partir do modelo totalmente regionalizado com dados de concelho para estimar o preço médio de transacção das habitações e dos apartamentos.

Tabela 3
Precisão Relativa das estimativas do preço médio de transacção da habitação

	Estimador Combinado			Estimador Directo		
	Mínima	Média	Máxima	Mínima	Média	Máxima
Nível de NUTS III						
Habitação	2,3%	5,6%	11,5%	2,7%	19,8%	186,8%
Apartamentos	2,2%	5,4%	12,2%	2,6%	20,3%	195,4%
Nível de Concelho						
Habitação	4,5%	14,9%	65,7%	3,9%	37,5%	338,4%

Ao nível de desagregação NUTS III, e ao contrário do que acontece com o estimador directo, o estimador combinado proposto permite estimar, com precisão aceitável, o preço médio de transacção da habitação e dos apartamentos, verificando-se em ambos os casos que as PR máximas estão na ordem dos 12%.

Contudo, ao nível de desagregação concelho, o estimador combinado proposto não permite estimar, com precisão aceitável, o preço médio de transacção da habitação, dos apartamentos e das moradias. Na situação em que as dimensões amostrais nos domínios são as mais elevadas, ou seja, na estimação do preço médio de transacção das habitações (que incluem os apartamentos e as moradias), as PR das estimativas situam-se entre os 4,5% e os 65,7%, sendo a PR média igual a 14,9%. De qualquer forma, também a este nível de desagregação, o estimador proposto apresenta ganhos de precisão significativos quando comparado com o estimador directo pós-estratificado. Ainda a este nível de desagregação, foi possível produzir estimativas com precisão aceitável para alguns concelhos pertencentes a NUTS III com maiores dimensões amostrais, sendo de realçar, que para essas NUTS III, foi também possível obter estimativas *design-based* aceitáveis, com base no estimador directo pós-estratificado.

Finalmente, observaram-se algumas indicações de que os estimadores combinados propostos produzem "boas" estimativas: por um lado, as estimativas produzidas através desses estimadores estão contidas, na sua maioria, nos intervalos de confiança das estimativas directas, e por outro lado, as estimativas produzidas pelos estimadores combinados e as suas precisões relativas apresentam maior estabilidade ao longo do tempo do que as estimativas directas.

7. Conclusão

A partir dos resultados obtidos é possível concluir que os estimadores combinados propostos, que tiram partido de dados cronológicos, apresentam, para qualquer nível de desagregação, ganhos significativos de precisão, quando comparados com os estimadores *design-based*. Conclui-se também que os modelos totalmente regionalizados são aqueles que apresentam os menores valores no AIC e os melhores níveis de precisão das estimativas produzidas. Estes estimadores permitem produzir estimativas do preço médio de transacção da habitação e dos apartamentos (por metro quadrado de área útil) com precisão adequada para a generalidade das NUTS III a nível nacional. No entanto, considera-se globalmente inaceitável a precisão das estimativas produzidas por estes estimadores para as moradias ao nível de desagregação NUTS III, e para a habitação, apartamentos e moradias ao nível de desagregação Concelho. Foi ainda possível produzir estimativas com precisão aceitável para alguns concelhos pertencentes a NUTS III com maiores dimensões amostrais, sendo de realçar, que para estas NUTS III, é também possível obter estimativas *design-based* aceitáveis, com base no estimador directo pós-estratificado.

Bibliografia

Amemiya, T. (1971), The Estimation of the Variances in a Variance-Components Model, *International Economic Review*, 12, 1-13.

Bozdogan, H. (1987), Model selection and Akaike's Information Criterio (AIC): the general Theory and its analytical extensions, *Psichometrika*, 52, 345-370.

Choudhry, G. H. e J. N. K. Rao (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*, 67-74.

Coelho, P. S. (1996), Estimadores Combinados para Pequenos Domínios, *Revista de Estatística*, 2, 23-43.

Coelho, P. S. (2000), *Estimação em Domínios sob o Modelo Linear Geral Misto com Informação Cronológica e Espacial*, Tese de Doutoramento, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisboa.

Datta, G. S., P. Lahiri e T. Maiti (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.

Fay, R. E. e R. A. Herriot (1979), Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association*, 74 (366), 269-277.

Fuller, W. A. e G. E. Battese (1973), Transformations for Estimation of Linear Models with Nested-Error Structure, *Journal of the American Statistical Association*, 68, 626-632.

Ghosh, M. e N. Nangia (1993), *Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach* (Technical Report), University of Florida: Department of Statistics, Gainsville.

Ghosh, M. e J. N. K. Rao (1994), Small Area Estimation: An Appraisal, *Statistical Science*, 9 (1), 55-93.

Ghosh, M., N. Nangia e D. Kim (1996), Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach, *Journal of the American Statistical Association*, 91, 1423-1431.

Hartley, H. O. e J. N. K. Rao (1967), Maximum Likelihood Estimation for the Mixed Analysis of Variance Model, *Biometrika*, 54, 93-108.

Harville, D. A. (1977), Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, 72, 322-340.

Henderson, C. R. (1975), Best Linear Unbiased Estimation and Prediction under a Selection Model, *Biometrics*, 31, 423-447.

Jennrich, R. e M. Schluchter (1986), Unbalanced Repeated-measures models with structured covariance matrices, *Biometrics*, 42, 805-820.

Laird, N. M. e J. H. Ware (1982), Random-Effects Models for Longitudinal Data, *Biometrics*, 38, 963-974.

Pfeffermann, D. (2002), Small Area Estimation – New Developments and Directions, *International Statistical Review*, 70(1), 125-143.

Pfeffermann, D. e L. Burck (1990), Robust Small Area Estimation Combining Time Series and Cross-Sectional Data, *Survey Methodology*, 16 (2), 217-237.

Pfeffermann, D., M. Feder e D. Signorelli (1998), Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas, *Journal of Business and Economic Statistics*, 16, 339-348.

Pereira, L. (2005), Estimação em Pequenos Domínios utilizando Informação Seccional e Cronológica – O caso da Estimação do Preço Médio de

Transacção da Habitação, Tese de Mestrado, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisboa.

Prasad, N. G. N. e J. N. K. Rao (1999), On Robust Small Area Estimation Using a Simple Random Effects Model, *Survey Methodology*, 25 (1), 67-72.

Rao, C. (1970), Estimation of Heteroscedastic Variances in Linear Models, *Journal of the American Statistical Association*, 65, 161-172.

Rao, C. (1972), Estimation of variance and covariance components in linear models, *Journal of the American Statistical Association*, 67, 112-115.

Rao, J. N. K. (1999), Some Recent Advances in Model-Based Small Area Estimation, *Survey Methodology*, 25 (2), 175-186.

Rao, J. N. K. (2000), Statistical Methodology for Indirect Estimations in Small Areas. *Seminario Internacional de Estadística en Euskadi. EUSTAT – The Basque Statistics Institute.* [Versão Electrónica]. Acedido em 30 de Outubro, 2006, a partir de http://www.eustat.es/prodserv/datos/vol0039.pdf.

Rao, J. N. K. (2003), *Small Area Estimation*. John Wiley & Sons.

Rao, J. N. K. e M. Yu (1992), Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

Rao, J. N. K. e M. Yu (1994), Small Area Estimation by Combining Time Series and Cross-Sectional Data, *Canadian Journal of Statistics*, 22, 511-528.

Singh, A. C., H. J. Mantel e B. W. Thomas (1994), Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20 (1), 33-43.

Wolfinguer, R. D. (1996), Heterogeneous variance-covariance structures for repeated measures, *Journal of Agricultural, Biological & Environmental Statistics*, 1, 205-230.

Wolfinguer, R. D., R. Tobias e J. Sall (1994), Computing Gaussian Likelihoods and their Derivatives for General Linear Mixed Models, *SIAM Journal of Scientific Computing*, 15(6), 1294-1310.

You, Y. (1999), *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*, Unpublished Ph.D. Thesis. Carleton University, Ottawa, Canada.