Estimação em pequenos domínios com modelos que combinam informação seccional e cronológica

Luís de Noronha e Pereira

Escola Superior de Gestão, Hotelaria e Turismo, Universidade do Algarve

Pedro Simões Coelho

Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa

Resumo: O objectivo principal deste artigo consiste na proposta de um novo estimador para parâmetros de interesse em pequenos domínios com dados de nível área. O estimador combinado proposto é assistido por um modelo que se enquadra na classe do modelo linear geral misto. Esse modelo é uma extensão do modelo de Fay e Herriot (1979), mas que permite a utilização de informação, relativa à variável de interesse e à sua relação com as variáveis auxiliares, proveniente de vários domínios e de vários períodos de tempo em simultâneo. É também deduzida uma aproximação de segunda ordem do Erro Quadrático Médio (EQM) *model-based* do estimador combinado proposto. São ainda propostos novos métodos de estimação dos parâmetros de variância de modelos de nível área. Neste artigo são apresentados os resultados de um estudo empírico por simulação de Monte Carlo que teve como objectivo a avaliação dos méritos relativos do estimador combinado proposto.

Palavras–chave: Estimação em pequenos domínios, estimador combinado, estimação *model-assisted*, estimação das componentes de variância, estimação do EQM

Abstract: The main purpose of this paper is to propose a new estimator for the interest parameters in small domains with area level data. A model that is a particular case of the general mixed linear model class assists the proposed combined estimator. This model is an extension of a model due to Fay and Herriot (1979), but it allows the integration of information related to the interest variable and to its relation to the auxiliary variables, from several domains and several periods of time, simultaneously. It is also deducted a second order approach to the model-based Mean Squared Error (MSE) of the proposed combined estimator. In addition, new estimation methods of variance components of area level models are also proposed. In this paper, the results of an empirical study carried out through a Monte Carlo simulation are presented. The study aimed at appraising the relative merits of the proposed combined estimator.

Keywords: Small area estimation, combined estimator, model-assisted estimation, estimation of variance components, estimation of MSE

1 Introdução

Apesar da maioria da investigação efectuada sobre estimação em pequenos domínios, no âmbito dos modelos de nível área, estar centrada na utilização de dados seccionais, a partir do final século passado os modelos que combinam dados seccionais e cronológicos passaram a ser utilizados com maior frequência. Veja-se, por exemplo, os trabalhos realizados por Ghosh e Rao (1994), Rao (1999) e Pfeffermann (2002), para se ter uma ideia da investigação que tem sido feita neste domínio. Actualmente é possível encontrar na literatura várias aplicações de modelos que combinam dados de natureza seccional e cronológica, uma vez 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. Esses modelos podem ser classificados em dois grandes grupos: os modelos do tipo do modelo de Rao-Yu e os modelos do tipo *state space*. Contudo, estes dois tipos de 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 possível especificação de estruturas de covariância cronológica arbitrárias sobre os efeitos aleatórios do modelo e/ou sobre as variáveis residuais. A abordagem proposta neste artigo para estimar parâmetros em pequenos domínios, quando está disponível informação de natureza seccional e cronológica, tenta explorar toda essa flexibilidade.

2 Modelo de nível área com dados seccionais e cronológicos

2.1 Especificação do modelo

Na especificação do modelo supõe-se que existem observações disponíveis para a variável de interesse e para as variáveis explicativas referentes a T períodos de tempo e que não é possível fazer a sua ligação ao nível das unidades individuais. Desta forma, 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 áreas mais vastas, designadas por 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 $\widehat{\theta}_{adt}$ um estimador directo de um parâmetro da variável de interesse no

d-ésimo domínio pertencente à região a no período t, θ_{adt} (a=1,..., A; d=1,..., $D(a)$; t=1,..., T). Assume-se que $\widehat{\theta}_{adt}$ é um estimador não enviesado no desenho de θ_{adt} e está disponível sempre que $n_{adt} \geq 1$, isto é,

$$\widehat{\theta}_{adt} = \theta_{adt} + \varepsilon_{adt} \tag{1}$$

onde os ε_{adt} são os erros da sondagem com $E_d(\varepsilon_{adt} \mid \theta_{adt})$. Os erros da sondagem associados a diferentes unidades são não correlacionados entre si, podendo ser heterocedásticos. Assume-se que os erros da sondagem são conhecidos. Assume-se também que existe um vector, $\mathbf{x}_{adt} = (x_{adt1}, ..., x_{adtp})'$, de p variáveis explicativas. Neste contexto, propõe-se o seguinte modelo que utiliza 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 aleatórios:

$$\theta_{adt} = \mathbf{x}'_{adt}\beta_{at} + u_{adt} \tag{2}$$

onde $\beta_{at} = (\beta_{at1}, ..., \beta_{atp})'$ é um vector de p efeitos fixos e u_{adt} são os efeitos aleatórios. Assume-se que os efeitos aleatórios têm média nula e que efeitos aleatórios associados a diferentes domínios são não correlacionados entre si. Contudo, supõe-se que os efeitos aleatórios associados a um determinado domínio poderão apresentar uma estrutura de covariância cronológica. Combinando (1) com o modelo (2), obtém-se o modelo seguinte:

$$\widehat{\theta}_{adt} = \mathbf{x}'_{adt}\beta_{at} + u_{adt} + \varepsilon_{adt} \tag{3}$$

onde $E_d(\varepsilon_{adt} \mid \theta_{adt}) = 0$, $V_d(\varepsilon_{adt} \mid \theta_{adt}) = \sigma^2_{\varepsilon,adt}$, $E(u_{adt}\varepsilon_{adt}) = 0$, $E_m(u_{adt}) = 0$ e $Cov_m(u_{adt}, u_{a'd't'}) = \begin{cases} \sigma_{u,att'} & , a = a',\ d = d' \\ 0 & , \textit{caso contrário} \end{cases}$.

O modelo (3) é uma extensão do modelo de Fay e Herriot (1979) para dados seccionais e cronológicos. Esse modelo é muito abrangente, pois pode apresentar qualquer tipo de estrutura de covariância cronológica dos efeitos aleatórios. Todavia, tendo em conta a parcimónia no número de parâmetros a estimar, e considerando admissível supor que a correlação cronológica associada aos dados amostrais de um determinado domínio decresce exponencialmente com o passar do tempo, propõe-se a utilização da estrutura de covariância cronológica auto-regressiva de primeira ordem heterogénea [ARH(1)]:

$$\sigma_{u,att'} = \begin{cases} \sigma^2_{u,at} & , t = t' \\ \sigma_{u,at}\sigma_{u,at'}\rho_a^{|t-t'|} & , t \neq t' \end{cases} \qquad |\rho_a| < 1. \tag{4}$$

No modelo (3), 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 cronológico totalmente regionalizado. Pode considerar-se um caso particular do modelo (3), denominado por modelo cronológico não regionalizado, na situação em que não

se supõe que os domínios possam ser agrupados em A regiões. Quando não se considera a existência de informação cronológica, podem considerar-se outros dois casos particulares do modelo (3), denominados por modelo contemporâneo regionalizado e por modelo contemporâneo não regionalizado (modelo proposto por Fay e Herriot (1979)).

2.2 Representação compacta do modelo

Sejam $\widehat{\theta}_{ad} = col_{1\leq t\leq T}\left(\widehat{\theta}_{adt}\right)$, $\widehat{\theta}_a = col_{1\leq d\leq D(a)}\left(\widehat{\theta}_{ad}\right)$ e $\widehat{\theta} = col_{1\leq a\leq A}\left(\widehat{\theta}_a\right)$ vectores coluna das estimativas de θ_{adt} de ordens T, $TD(a)$ e TD, respectivamente. O modelo (3) pode ser escrito na seguinte forma matricial:

$$\widehat{\theta} = \mathbf{X}\beta + \mathbf{u} + \varepsilon \tag{5}$$

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

Assume-se que $E_m(\mathbf{u}) = \mathbf{0}$ e que $V_m(\mathbf{u}) = \mathbf{G} = diag_{1\leq d\leq D(a);1\leq a\leq A}(\mathbf{G}_{ad})$, onde $\mathbf{G}_{ad} = \{\sigma_{u,att'}\}$ é uma matriz simétrica com elementos $\sigma_{u,att'}$, $t, t' = 1, ..., T$. Assume-se que $E_d(\varepsilon) = \mathbf{0}$ e que $V_m(\varepsilon) = \mathbf{R} = diag_{1\leq d\leq D(a);1\leq a\leq A}(\mathbf{R}_{ad})$, onde $\mathbf{R}_{ad} = diag_{1\leq t\leq T}\left(\sigma^2_{\varepsilon,adt}\right)$. Assume-se ainda que $E(\mathbf{u}\varepsilon') = \mathbf{0}$. A matriz de variâncias-covariâncias de $\widehat{\theta}$, de dimensão $TD \times TD$, é dada por $V_m\left(\widehat{\theta}\right) = \mathbf{V} = diag_{1\leq d\leq D(a);1\leq a\leq A}(\mathbf{V}_{ad})$, onde $\mathbf{V}_{ad} = \mathbf{G}_{ad} + \mathbf{R}_{ad}$. O modelo (3) é um caso particular do modelo linear geral misto com efeitos fixos e efeitos aleatórios, com $\mathbf{Z} = \mathbf{I}_{TD}$ e estrutura de variâncias-covariâncias diagonal por blocos.

2.3 O *Best Linear Unbiased Predictor*

Assumindo que as componentes de variância são conhecidas, então pode-se demonstrar que o *Best Linear Unbiased Predictor* (BLUP) de θ_{adt} é dado por:

$$\widetilde{\theta}_{adt} = \widetilde{\theta}^H_{adt}(\delta_a) = \mathbf{X}_{adt}\widetilde{\beta}_a + \mathbf{h}'_{adt}\left(\widehat{\theta}_{ad} - \mathbf{X}_{ad}\widetilde{\beta}_a\right) \tag{6}$$

onde $\mathbf{X}_{adt} = \left(\begin{array}{ccc} {}_1\mathbf{0}_{(t-1)p} & \mathbf{x}'_{adt} & {}_1\mathbf{0}_{(T-t)p} \end{array}\right)$ é um vector linha de ordem Tp, $\widetilde{\beta}_a = \widetilde{\beta}_a(\delta_a) = \left(\sum_{d=1}^{D(a)} \mathbf{X}'_{ad}\mathbf{V}^{-1}_{ad}\mathbf{X}_{ad}\right)^{-1} \times \left(\sum_{d=1}^{D(a)} \mathbf{X}'_{ad}\mathbf{V}^{-1}_{ad}\widehat{\theta}_{ad}\right)$ é o BLUE de β_a, $\mathbf{h}'_{adt} = \mathbf{h}'_{adt}(\delta_a)$ é um vector usado para ponderar os resíduos da regressão correspondente à t-ésima linha da matriz $\mathbf{H}_{ad} = \mathbf{G}_{ad}\mathbf{V}^{-1}_{ad} = col_{1\leq t\leq T}(\mathbf{h}'_{adt})$ de dimensão $T \times T$, sendo $\mathbf{h}_{adt} = col_{1\leq t'\leq T}(h_{adtt'})$, e onde $\delta_a = \left(\rho_a, \sigma^2_{u,a1}, ..., \sigma^2_{u,aT}\right)'$ é o vector das $(T+1)$ componentes de variância associadas à região a. O estimador (6) pode ser classificado como um estimador combinado, uma vez que pode ser decomposto em duas componentes: um estimador sintético, $\mathbf{X}_{adt}\widetilde{\beta}_a$, e

um factor de correcção, $\mathbf{h}'_{adt}\left(\widehat{\theta}_{ad}-\mathbf{X}_{ad}\widetilde{\beta}_a\right)$, que é função das diferenças entre as estimativas directas e as estimativas sintéticas do parâmetro de interesse. É possível observar em (6) 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 tirando partido da sua potencial autocorrelação cronológica, desde que existam observações em pelo menos uma das vagas anteriores. Esta é 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. A metodologia proposta pode ser enquadrada numa abordagem *model-assisted*, dado que o estimador (6) considera o plano de sondagem.

O EQM de $\widetilde{\theta}_{adt}$, que depende de $(T+1)$ componentes de variância, δ_a, pode ser decomposto na soma de duas componentes (Henderson, 1975):

$$EQM\left[\widetilde{\theta}_{adt}\left(\delta_a\right)\right]=E\left[\widetilde{\theta}_{adt}\left(\delta_a\right)-\theta_{adt}\right]^2=g_{1adt}\left(\delta_a\right)+g_{2adt}\left(\delta_a\right) \tag{7}$$

com

$$g_{1adt}\left(\delta_a\right)=\sigma_{u,att}-\mathbf{g}'_{adt}\mathbf{V}_{ad}^{-1}\mathbf{g}_{adt} \tag{8}$$

e

$$g_{2adt}\left(\delta_a\right)=\mathbf{M}_{adt}\left(\sum_{d=1}^{D(a)}\mathbf{X}'_{ad}\mathbf{V}_{ad}^{-1}\mathbf{X}_{ad}\right)^{-1}\mathbf{M}'_{adt} \tag{9}$$

onde $\mathbf{M}_{adt}=\mathbf{X}_{adt}-\mathbf{g}'_{adt}\mathbf{V}_{ad}^{-1}\mathbf{g}_{adt}$, $\mathbf{g}_{adt}=col_{1\leq t'\leq T}\left(\sigma_{u,at't}\right)$, sendo $g_{1adt}\left(\delta_a\right)$ devida à estimação dos efeitos aleatórios do modelo e $g_{2adt}\left(\delta_a\right)$ devida à estimação dos efeitos fixos do modelo.

2.4 O *Empirical Best Linear Unbiased Predictor*

O *Empirical Best Linear Unbiased Predictor* (EBLUP), conhecido por BLUP empírico, obtém-se através da substituição dos parâmetros desconhecidos, δ_a, por estimadores assimptoticamente consistentes $\widehat{\delta}_a=\left(\widehat{\rho}_a,\widehat{\sigma}^2_{u,a1},...,\widehat{\sigma}^2_{u,aT}\right)'$ na expressão do BLUP:

$$\widehat{\widetilde{\theta}}_{adt}=\widetilde{\theta}^H_{adt}\left(\widehat{\delta}_a\right)=\mathbf{X}_{adt}\widehat{\beta}_a+\widehat{\mathbf{h}}'_{adt}\left(\widehat{\theta}_{ad}-\mathbf{X}_{ad}\widehat{\beta}_a\right). \tag{10}$$

No âmbito da classe de modelos apresentada, propõe-se que o parâmetro de correlação seja estimado através de um estimador simplista proposto por Rao e Yu (1994), e que os parâmetros de variância, dos efeitos aleatórios e dos erros da sondagem, sejam estimados através de um dos dois métodos dos momentos adaptados propostos por Pereira (2005).

2.5 Erro quadrático médio do *Empirical Best Linear Unbiased Predictor*

Sob condições de normalidade dos efeitos aleatórios do modelo, uma medida de variabilidade associada ao EBLUP, $\widehat{\widetilde{\theta}}_{adt}$, é dada por (Kackar e Harville, 1984):

$$EQM\left(\widehat{\widetilde{\theta}}_{adt}\right) = EQM\left[\widetilde{\theta}_{adt}\left(\delta_a\right)\right] + E\left[\widehat{\widetilde{\theta}}_{adt}\left(\widehat{\delta}_a\right) - \widetilde{\theta}_{adt}\left(\delta_a\right)\right]^2 \tag{11}$$

onde o último termo é obtido por aproximação porque é geralmente intratável. Supondo a normalidade de $\mathbf{u}$ e de ε e assumindo certas condições de regularidade, uma aproximação de segunda ordem do último termo de (11) pode ser dada por (Prasad e Rao, 1990; Datta e Lahiri, 2000):

$$E\left[\widehat{\widetilde{\theta}}_{adt}\left(\widehat{\delta}_a\right) - \widetilde{\theta}_{adt}\left(\delta_a\right)\right]^2 \doteq tr\left[\left(\nabla \mathbf{b}'_{adt}\right)\mathbf{V}\left(\nabla \mathbf{b}'_{adt}\right)'\overline{\mathbf{V}}\left(\widehat{\delta}_a\right)\right] = g_{3adt}\left(\delta_a\right) \tag{12}$$

onde $\nabla \mathbf{b}'_{adt} = \left(\frac{\partial \mathbf{b}'_{adt}}{\partial \delta_a}\right)$ e $\overline{\mathbf{V}}\left(\widehat{\delta}_a\right)$ é a matriz de variâncias-covariâncias assimptótica dos estimadores, $\widehat{\delta}_a$, de δ_a. Sendo necessário nas aplicações práticas uma medida da variabilidade associada a $\widehat{\widetilde{\theta}}_{adt}\left(\widehat{\delta}_a\right)$, e tendo em conta a dificuldade de cálculo de uma aproximação da variabilidade associada a $\widehat{\delta}_a$, decidiu-se seguir uma abordagem simplista que consiste em fazer a aproximação do $EQM\left[\widehat{\widetilde{\theta}}_{adt}\left(\widehat{\delta}_a\right)\right]$ pelo $EQM\left[\widetilde{\theta}_{adt}\left(\delta_a\right)\right]$. O estimador simplista do EQM do EBLUP é dado por:

$$eqm_S\left[\widehat{\widetilde{\theta}}_{adt}\left(\widehat{\delta}_a\right)\right] = g_{1adt}\left(\widehat{\delta}_a\right) + g_{2adt}\left(\widehat{\delta}_a\right). \tag{13}$$

3 Estudo empírico por simulação

3.1 Simulação de Monte Carlo numa pseudo-população

Foi realizado um estudo empírico por simulação de Monte Carlo, sobre uma pseudo-população de empresas de mediação imobiliária, com o objectivo de se identificarem os méritos relativos dos estimadores combinados propostos relativamente a diversos outros estimadores directos e indirectos que são habitualmente utilizados na estimação do parâmetro de interesse em pequenos domínios. Este estudo foi baseado em séries cronológicas multivariadas resultantes do Inquérito aos Preços Médios de Transacção na Habitação (IPTH) e do Inquérito aos Preços de Avaliação Bancária na Habitação. A partir de uma pseudo-população de 4,655 empresas de mediação imobiliária sediadas em Portugal Continental, foram extraídas 1,000 amostras independentes longitudinais de 229

empresas da pseudo-população, utilizando uma sondagem aleatória estratificada por conglomerados. Foram avaliadas as propriedades *design-based* de um conjunto de estimadores utilizados para estimar o preço médio de transacção da habitação por metro quadrado, ao nível das regiões de Portugal classificadas ao nível III da Nomenclatura das Unidades Territoriais para Fins Estatísticos (NUTS III). Os domínios, d=1, ..., 28, correspondem ao nível de agregação de NUTS III em Portugal Continental, as regiões, a=1, 2, 3, correspondem a um nível de agregação mais abrangente definido como um grupo de domínios e os momentos de tempo, t=1, ..., 7, correspondem aos sete trimestres em que foram realizadas as vagas do IPTH.

3.2 Estimadores

Os estimadores directos e sintéticos para a média em pequenos domínios analisados no estudo empírico são os seguintes: estimador directo pós-estratificado, $\widehat{\mu}_{dt,POS}$; estimador sintético pelo quociente, $\widehat{\mu}_{dt,SQ}$; e estimador sintético pela regressão, $\widehat{\mu}_{dt,SR}$. A forma e propriedades destes três estimadores podem ser encontradas em Coelho (1996). Os estimadores combinados para a média em pequenos domínios baseados na classe de modelos proposta são os seguintes: estimador baseado num modelo contemporâneo não regionalizado com componentes de variância estimadas pelo método ANOVA adaptado, $\widehat{\widetilde{\mu}}_{1,dt,AN}$; estimador baseado num modelo contemporâneo não regionalizado com componentes de variância estimadas pelo método de transformação de Fuller-Battese adaptado, $\widehat{\widetilde{\mu}}_{2,dt,FB}$; estimador baseado num modelo contemporâneo regionalizado com componentes de variância estimadas pelo método ANOVA adaptado, $\widehat{\widetilde{\mu}}_{3,adt,AN}$; estimador baseado num modelo contemporâneo regionalizado com componentes de variância estimadas pelo método de transformação de Fuller-Battese adaptado, $\widehat{\widetilde{\mu}}_{4,adt,FB}$; estimador baseado num modelo cronológico não regionalizado com componentes de variância estimadas pelo método ANOVA adaptado, $\widehat{\widetilde{\mu}}_{5,dt,AN}$; estimador baseado num modelo cronológico não regionalizado com componentes de variância estimadas pelo método de transformação de Fuller-Battese adaptado, $\widehat{\widetilde{\mu}}_{6,dt,FB}$; estimador baseado num modelo cronológico regionalizado com componentes de variância estimadas pelo método ANOVA adaptado, $\widehat{\widetilde{\mu}}_{7,adt,AN}$; estimador baseado num modelo cronológico regionalizado com componentes de variância estimadas pelo método de transformação de Fuller-Battese adaptado, $\widehat{\widetilde{\mu}}_{8,adt,FB}$. Os oito estimadores combinados apresentados acima correspondem a variantes do estimador (6), utilizando diferentes métodos de estimação de componentes de variância. Os estimadores baseados em modelos contemporâneos diferenciam-se dos estimadores baseados em modelos cronológicos pelo facto dos primeiros só utilizarem informação relativa a um período de tempo. Os estimadores baseados em modelos regionalizados distinguem-se dos estimadores baseados em modelos não regionalizados pelo facto dos primeiros considerarem que os domínios são agrupados em três regiões, com diferentes efeitos fixos e covariância cronológica dos efeitos aleatórios.

3.3 Medidas de Precisão e de Enviesamento

Os vinte e oito domínios foram divididos em seis grupos mutuamente exclusivos em função da sua dimensão média (número de transacções), pois a avaliação dos méritos relativos dos estimadores propostos deve ser feita em função das dimensões amostrais dos domínios. As dimensões médias amostrais dos seis grupos de domínios são as seguintes: grupo 1 - 1 transacção; grupo 2 - 2 a 5 transacções; grupo 3 - 6 a 9 transacções; grupo 4 - 10 a 19 transacções; grupo 5 - 20 a 29 transacções; grupo 6 - pelo menos 30 transacções.

A qualidade dos estimadores do parâmetro de interesse foi avaliada através do conjunto seguinte de medidas de precisão e de enviesamento *design-based*: enviesamento absoluto, enviesamento relativo, EQM, variância, rácio de enviesamento absoluto, taxa de cobertura do intervalo de confiança *design-based*, taxa de cobertura do intervalo de confiança *model-based*, precisão relativa, erro-padrão relativo e coeficiente de variação. Para a análise dos resultados decidiu-se calcular a média de todas as medidas de precisão e de enviesamento expostas acima para cada grupo de domínios.

3.4 Resultados

A partir das análises comparativas sobre os méritos relativos dos estimadores combinados propostos, concluiu-se que o estimador combinado que apresenta melhores propriedades, em termos de precisão e de enviesamento para a generalidade dos domínios, é $\widehat{\widetilde{\mu}}_{8,adt,FB}$. Com base nos resultados obtidos, foi possível observar que $\widehat{\mu}_{dt,SQ}$ é o estimador que apresenta a pior precisão e o pior comportamento em termos de enviesamento em todos os grupos de domínios, enquanto $\widehat{\mu}_{dt,SR}$ é o estimador que apresenta o pior comportamento em termos de rácio de enviesamento. Por outro lado, verificou-se globalmente que todos os estimadores combinados propostos apresentam um comportamento melhor em termos de precisão do que $\widehat{\mu}_{dt,POS}$. Estes ganhos são numericamente mais expressivos para os estimadores combinados baseados em modelos cronológicos regionalizados ($\widehat{\widetilde{\mu}}_{7,adt,AN}$ e $\widehat{\widetilde{\mu}}_{8,adt,FB}$). Relativamente ao enviesamento, observou-se para a generalidade dos domínios que só os estimadores combinados $\widehat{\widetilde{\mu}}_{7,adt,AN}$ e $\widehat{\widetilde{\mu}}_{8,adt,FB}$ apresentam menores enviesamentos do que $\widehat{\mu}_{dt,POS}$. Os resultados do estudo empírico também mostraram que todos os estimadores combinados propostos são globalmente mais precisos e apresentam menores enviesamentos absolutos, e principalmente menores rácios de enviesamento, do que $\widehat{\mu}_{dt,SR}$.

Perante os comentários anteriores, parece ser possível obter estimadores combinados baseados no modelo proposto, que apresentam globalmente, para domínios de pequena dimensão amostral, grandes ganhos de precisão relativamente ao estimador directo e aos estimadores sintéticos, permitindo simultaneamente ganhos em termos de enviesamento. Em particular, realçam-se os casos dos estimadores $\widehat{\widetilde{\mu}}_{7,adt,AN}$ e $\widehat{\widetilde{\mu}}_{8,adt,FB}$.

No que se refere à análise da qualidade das medidas de precisão *model-based*,

Tabela 1: Rácios entre medidas de precisão e de enviesamento de estimadores combinados (regionalizados *versus* não regionalizados)

Grupo	$\widehat{\widetilde{\mu}}_{3,adt,AN}$	$\widehat{\widetilde{\mu}}_{4,adt,FB}$	$\widehat{\widetilde{\mu}}_{7,adt,AN}$	$\widehat{\widetilde{\mu}}_{8,adt,FB}$
Média dos Enviesamentos Absolutos				
1	1.26	1.16	1.32	1.26
2	1.07	1.12	0.78	0.79
3	0.98	1.04	0.85	0.85
4	0.95	1.00	0.80	0.80
5	0.95	0.99	0.83	0.84
6	0.94	1.00	0.82	0.84
Média dos Erros Quadráticos Médios				
1	1.56	1.33	1.64	1.49
2	1.16	1.38	0.64	0.65
3	0.94	1.07	0.70	0.69
4	0.90	0.93	0.69	0.63
5	0.90	0.99	0.69	0.69
6	0.88	1.00	0.66	0.71
Média das Variâncias				
1	1.40	1.75	0.75	0.58
2	0.70	0.78	0.49	0.46
3	0.90	0.95	0.61	0.59
4	0.83	0.79	0.60	0.48
5	0.84	0.93	0.56	0.54
6	0.89	1.03	0.61	0.64

Tabela 2: Número de NUTS III em cada intervalo de precisão relativa média, para cada um dos estimadores

Precisão Relativa	$\widehat{\mu}_{dt,POS}$	$\widehat{\mu}_{dt,SQ}$	$\widehat{\mu}_{dt,SR}$	$\widehat{\widetilde{\mu}}_{8,adt,FB}$
$< 8\%$	1	0	0	1
$[\ 8\% ; 12\% [$	3	0	3	7
$[\ 12\% ; 16\% [$	8	2	11	10
$[\ 16\% ; 20\% [$	6	4	3	6
$\geq 20\%$	10	22	11	4
Total	28	28	28	28

esta foi efectuada através da comparação das taxas de cobertura dos intervalos de confiança *design-based* com as taxas de cobertura dos intervalos de confiança *model-based*, sob amostragem repetida. Os resultados da simulação vieram revelar que os intervalos de confiança *model-based* podem constituir uma alternativa

Tabela 3: Rácios entre medidas de precisão e de enviesamento de estimadores combinados (cronológicos *versus* contemporâneos)

Grupo	$\widehat{\widetilde{\mu}}_{5,dt,AN}$	$\widehat{\widetilde{\mu}}_{6,dt,FB}$	$\widehat{\widetilde{\mu}}_{7,adt,AN}$	$\widehat{\widetilde{\mu}}_{8,adt,FB}$
Média dos Enviesamentos Absolutos				
1	0.92	0.90	0.96	0.97
2	1.00	0.94	0.72	0.67
3	1.00	0.99	0.86	0.81
4	1.02	1.01	0.86	0.81
5	1.01	0.99	0.88	0.84
6	1.00	1.00	0.87	0.84
Média dos Erros Quadráticos Médios				
1	0.87	0.84	0.92	0.94
2	1.03	1.02	0.56	0.48
3	1.00	1.01	0.74	0.65
4	1.06	1.04	0.81	0.70
5	1.01	1.00	0.77	0.70
6	1.00	0.99	0.76	0.70
Média das Variâncias				
1	1.49	2.00	0.81	0.66
2	1.26	1.27	0.87	0.75
3	1.03	0.95	0.70	0.59
4	1.07	1.03	0.77	0.63
5	1.01	0.98	0.67	0.57
6	1.00	1.00	0.68	0.62

viável aos intervalos de confiança *design-based*, mesmo sob uma perspectiva de amostragem repetida e baseados em estimativas simplistas do EQM *model-based* dos estimadores dos parâmetros de interesse.

4 Conclusão

Os resultados do estudo empírico por simulação vieram mostrar que o estimador combinado proposto (baseado num modelo cronológico e regionalizado) é o que apresenta melhores propriedades em termos de enviesamento e de precisão, quando comparado com outros estimadores directos e indirectos habitualmente utilizados na estimação do parâmetro de interesse em pequenos domínios. Por outro lado, os resultados do estudo empírico também vieram revelar que os intervalos de confiança *model-based* podem constituir uma alternativa viável aos intervalos de confiança *design-based*, mesmo sob uma perspectiva de amostragem repetida e baseados em estimativas simplistas do EQM *model-based* dos estimadores dos parâmetros de interesse.

Tabela 4: Rácios entre medidas de precisão e de enviesamento de estimadores combinados (Fuller-Battese adaptado *versus* ANOVA adaptado)

Grupo	$\widehat{\widetilde{\mu}}_{2,dt,FB}$	$\widehat{\widetilde{\mu}}_{4,adt,FB}$	$\widehat{\widetilde{\mu}}_{6,dt,FB}$	$\widehat{\widetilde{\mu}}_{8,adt,FB}$
Média dos Enviesamentos Absolutos				
1	1.12	1.03	1.09	1.05
2	0.99	1.03	0.94	0.96
3	0.93	0.99	0.93	0.94
4	0.92	0.96	0.91	0.91
5	0.95	0.99	0.94	0.95
6	0.94	1.00	0.94	0.96
Média dos Erros Quadráticos Médios				
1	1.25	1.07	1.20	1.09
2	0.90	1.06	0.89	0.91
3	0.86	0.98	0.87	0.86
4	0.85	0.88	0.83	0.77
5	0.89	0.98	0.88	0.89
6	0.88	1.00	0.87	0.93
Média das Variâncias				
1	0.79	0.99	1.06	0.82
2	0.77	0.85	0.78	0.73
3	0.87	0.93	0.80	0.78
4	0.84	0.81	0.82	0.66
5	0.87	0.96	0.84	0.81
6	0.86	0.99	0.86	0.90

Referências

[1] Coelho, P.S. (1996). *Estimação em pequenos domínios* (Working Paper n.º 48). Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisboa.

[2] Datta, G.S. e Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, Vol.10, p. 613-627.

[3] Fay, R.E. 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*, Vol.74, p. 269-277.

[4] Ghosh, M. e Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, Vol.9, p. 55-93.

[5] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, p. 423-447.

[6] Kackar, R.N. e 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*, Vol.79, p. 853-862.

[7] Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, Vol.70, p. 125-143.

[8] Pereira, L.N. (2005). *Estimação em Pequenos Domínios Utilizando Informação Seccional e Cronológica - O caso da estimação do preço médio de transacção da habitação.* Dissertação de Mestrado, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.

[9] Prasad, N.G.N. e Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, Vol.85, p. 163-171.

[10] Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, Vol.25, p. 175-186.

[11] Rao, J.N.K. e Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, Vol.22, p. 511-528.