Small Area Estimation using Linear Mixed Models with Heterogeneous Covariance Structures of Random Effects

Pereira, Luis N.
University of the Algarve – School of Management, Hospitality and Tourism, Department of Quantitative Methods
Campus da Penha
8005-139 Faro, Portugal
E-mail: Lmper@ualg.pt

Coelho, Pedro S.
New University of Lisbon - ISEGI
Campus de Campolide
1070-312 Lisboa, Portugal
E-mail: psc@isegi.unl.pt

1. Introduction

The problem of Small Area Estimation is about how to produce reliable estimates of domain characteristics when the sample sizes within the domains are very small or even zero. This makes it necessary to borrow strength from related areas through linking models based on auxiliary information, such as longitudinal survey data, leading to model-based indirect estimates. It is usually plausible to assume that observations taken over time on the same domain are correlated. However, it is rarely assumed that changes on variability across the data occur, that is, variability is heterogeneous. Ignoring or avoiding the heterogeneous covariance structure present in data may result in "poor" inferences.

This work focuses on the use of temporal area level models with heterogeneous covariance structures of random effects, as a vehicle for borrowing strength across the areas and over time. Heterogeneous compound symmetry (CSH) and heterogeneous first-order autoregressive [ARH(1)] covariance structures of random effects are used. These linking models are particular cases of the linear mixed models. Under the area level models proposed, formal expressions for BLUP and EBLUP of the small area parameter of interest have been explicitly obtained. We compare the performance of these models with the Rao-Yu model (Rao and Yu, 1994). The basic Rao-Yu model involves autocorrelated random effects with a homogeneous covariance structure (first-order autoregressive plus common covariance: AR(1)+J). The application of the proposed procedure is illustrated using the Portuguese prices of the habitation transaction and prices of bank evaluation of habitation series. A Monte Carlo simulation is carried out to perform empirical analysis.

In section 2, we consider the Rao-Yu model (Rao and Yu, 1994). The proposed chronological model, on the line of general linear mixed models, using heterogeneous covariance structures between random effects is presented in section 3. The BLUP and EBLUP of this model have been presented. Results of the Monte Carlo comparative study are given in section 4. Finally, section 5 contains concluding remarks.

2. Rao-Yu Model

The model proposed by Rao and Yu (1994) is the following three stage area specific model:

(1) $\hat{\theta}_{it} = \theta_{it} + e_{it}$,

(2) $\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + v_i + u_{it}$, $\quad u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, $\quad |\rho| < 1$,

where θ_{it} is the parameter of inferential interest for the i^{th} small-area at t^{th} time point (i=1, …, m; t=1, …, T) and $\hat{\theta}_{it}$ is its design-unbiased direct survey estimator, e_{it}'s are independent sampling errors normally distributed, given the θ_{it}'s, with mean 0 and known variance σ^2_{it}, $\mathbf{x}_{it}$ (p×1) is a column vector of an area-

by-time specific auxiliary variables and $\boldsymbol{\beta}$ ($p\times1$) is a column vector of regression parameters. Further, v_i's are random area specific effects with $v_i \overset{iid}{\sim} N(0,\sigma_v^2)$ and u_{it}'s are random area-by-time specific effects with $\varepsilon_{it} \overset{iid}{\sim} N(0,\sigma^2)$, following a common AR(1) process for each i. The errors $\{e_{it}\}$, $\{v_i\}$ and $\{u_{it}\}$ are assumed to be mutually independent. Combining the sampling error model (1) with the linking model (2), Rao and Yu (1994) obtained the following model:

$$\hat{\theta}_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_i + u_{it} + e_{it}, \qquad u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \qquad |\rho| < 1. \tag{3}$$

Rao and Yu (1994) applied a special form of the model (3) assuming $e_{it} \overset{iid}{\sim} N(0,1)$. They showed that the model (3) can be rewritten in matrix form as:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{u} + \mathbf{e}, \tag{4}$$

where $\hat{\boldsymbol{\theta}} = col_{1\le i\le m}(\hat{\boldsymbol{\theta}}_i)$, $\hat{\boldsymbol{\theta}}_i = col_{1\le t\le T}(\hat{\theta}_{it})$, $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$, $\mathbf{X}_i = col_{1\le t\le T}(\mathbf{x}_{it}')$, $\mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_T$, $\mathbf{v} = col_{1\le i\le m}(v_i)$, $\mathbf{u} = col_{1\le i\le m}(\mathbf{u}_i)$, $\mathbf{u}_i = col_{1\le t\le T}(u_{it})$, $\mathbf{e} = col_{1\le i\le m}(\mathbf{e}_i)$, $\mathbf{e}_i = col_{1\le t\le T}(e_{it})$, $\mathbf{I}_m$ is the identity matrix of order m and $\mathbf{1}_T$ ($T\times1$) is a column vector of 1's. Further, $\mathbf{e}$, $\mathbf{v}$ and $\mathbf{u}$ are mutually independent, with $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma})$, $\mathbf{v} \sim N(\mathbf{0}, \sigma_v^2\mathbf{I}_m)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, where $\boldsymbol{\Gamma}$ ($T\times T$) is the matrix with elements $\rho^{|i-j|}/(1-\rho^2)$ and $\mathbf{R} = diag_{1\le i\le m;1\le t\le T}(\sigma_{it}^2)$. Assuming that $e_{it} \overset{iid}{\sim} N(0,1)$, we can now see that the model (4) is a special case of the general linear mixed model with block diagonal homogeneous covariance structure, $Cov(\hat{\boldsymbol{\theta}}) = \mathbf{V} = block\ diag_{1\le i\le m}(\mathbf{V}_i) = block\ diag_{1\le i\le m}(\sigma^2\boldsymbol{\Gamma} + \sigma_v^2\mathbf{J}_T + \mathbf{I}_T)$. Following Rao and Yu (1994), assuming $\boldsymbol{\psi} = (\sigma_v^2, \sigma^2, \rho)'$ is known, the BLUP estimator of θ_{it} is given by:

$$\tilde{\theta}_{it}(\boldsymbol{\psi}) = \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}} + (\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_t)'\mathbf{V}_i^{-1}(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}), \tag{5}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\gamma}_t$ is the t^{th} row of $\boldsymbol{\Gamma}$. Rao and Yu (1994) estimated σ_v^2 and σ^2 by an extension of the simple transformation method of Fuller and Battese (1973) and proposed a naïve estimator of ρ, in order to obtain the EBLUP estimator of θ_{it}, $\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})$.

3. Proposed Model

Let $\hat{\theta}_{it}$ be the design-unbiased direct survey estimator for the characteristic under study, θ_{it}, in the i^{th} small-area at t^{th} time point (i=1, …, m; t=1, …, T). The proposed model is the following two stage area specific model:

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \tag{6}$$

$$\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_t + u_{it}, \tag{7}$$

where e_{it}'s are independent sampling errors normally distributed, given the θ_{it}'s, with mean 0 and known variance σ_{it}^2, $\mathbf{x}_{it}$ ($p\times1$) is a column vector of an area-by-time specific auxiliary variables and $\boldsymbol{\beta}_t$ ($p\times1$) is a column vector of regression parameters for the t^{th} time point. Further, u_{it}'s are random area-by-time specific effects normally distributed with $E(u_{it}) = 0$, $\text{cov}(u_{it}, u_{it'}) = \sigma_{u,tt'}$ for i=i' and 0 otherwise, and $E(e_{it}u_{it}) = 0$. Combining the sampling error model (6) with the linking model (7), we obtain the following model:

$$\hat{\theta}_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_t + u_{it} + e_{it}. \tag{8}$$

Arranging the data, the model (8) can be rewritten in matrix form as:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \tag{9}$$

where $\hat{\boldsymbol{\theta}} = col_{1\le i\le m}(\hat{\boldsymbol{\theta}}_i)$, $\hat{\boldsymbol{\theta}}_i = col_{1\le t\le T}(\hat{\theta}_{it})$, $\mathbf{X} = col_{1\le i\le m}(\mathbf{X}_i)$, $\mathbf{X}_i = diag_{1\le t\le T}(\mathbf{x}_{it}')$, $\boldsymbol{\beta} = col_{1\le t\le T}(\boldsymbol{\beta}_t)$, $\mathbf{u} = col_{1\le i\le m}(\mathbf{u}_i)$, $\mathbf{u}_i = col_{1\le t\le T}(u_{it})$, $\mathbf{e} = col_{1\le i\le m}(\mathbf{e}_i)$, $\mathbf{e}_i = col_{1\le t\le T}(e_{it})$. Further, $\mathbf{e}$ and $\mathbf{u}$ are mutually independent, with $E(\mathbf{e}) = \mathbf{0}$, $V(\mathbf{e}) = diag_{1\le i\le m}(\mathbf{R}_i)$, $E(\mathbf{u}) = \mathbf{0}$ and $V(\mathbf{u}) = \mathbf{I}_m \otimes \boldsymbol{\Sigma}$, where $\mathbf{R}_i = diag_{1\le t\le T}(\sigma_{it}^2)$ and $\boldsymbol{\Sigma} = \{\sigma_{u,tt'}\}$ is a symmetric ($T\times T$) matrix with elements $\sigma_{u,tt'}$, t,t'=1, ..., T, that define a chronological heterogeneous covariance structure of random effects within the i^{th} small-area. In practice, we assume that random effects associated to a particular i^{th} small-area may present a chronological heterogeneous compound symmetry (CSH) or a first-order autoregressive [ARH(1)] covariance structures (Wolfinguer, 1996). In spite of the correlations remaining constant for CSH, $\sigma_{u,tt'} = \sigma_{u,t}\sigma_{u,t'}\rho$ for $t\neq t'$ and $\sigma_{u,tt'} = \sigma_{u,tt}^2$ otherwise, and exponentially declining for ARH(1), $\sigma_{u,tt'} = \sigma_{u,t}\sigma_{u,t'}\rho^{|t-t'|}$, $|\rho| < 1$, both allow distinct variances along the main

diagonal. We selected heterogeneous covariance structures due to the observed increase of variability with time. In the class of heterogeneous covariance structures, CSH and ARH(1) were selected because they are parsimonious and usually fit the data well. In that way, the model (9) is a special case of the general linear mixed model with block diagonal heterogeneous covariance structure, $Cov(\hat{\boldsymbol{\theta}}) = \mathbf{V} = block\ diag_{1\le i\le m}(\mathbf{V}_i) = block\ diag_{1\le i\le m}(\boldsymbol{\Sigma} + \mathbf{R}_i)$. Following Henderson's general results (Henderson, 1975) and assuming that $\psi = (\sigma^2_{u,1},...,\sigma^2_{u,T},\rho)'$ is known, the BLUP estimator of θ_{it} is given by:

$$(10) \qquad \tilde{\theta}_{it}(\psi) = \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_t + \sum_{t'=1}^{T} h_{itt'}\left(\theta_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_t\right),$$

where $\tilde{\boldsymbol{\beta}} = col_{1\le t\le T}(\boldsymbol{\beta}_t) = \left(\sum_{i=1}^{m}\mathbf{X}'_i\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{X}'_i\mathbf{V}_i^{-1}\hat{\boldsymbol{\theta}}_i\right)$ and $\mathbf{h}'_{it} = [col_{1\le t\le T}(h_{itt'})]'$ is the t^{th} row of $\mathbf{H}_i = \boldsymbol{\Sigma}\mathbf{V}_i^{-1}$. We estimated the variance components by an extension of the simple transformation method of Fuller and Battese (1973) due to Pereira (2005), and ρ by a Rao and Yu (1994) naïve estimator, in order to obtain the EBLUP estimator of θ_{it}, $\hat{\theta}_{it}(\hat{\psi})$. A particular case of the model (8) can be considered, assuming that regression parameters doesn't vary over time.

4. Monte Carlo Study

The Rao-Yu model and the proposed model were compared empirically by means of a Monte Carlo simulation from a real time series obtained from the Prices of the Habitation Transaction Survey (PHTS) and the Prices of Bank Evaluation in the Habitation Survey (PBEHS). The data are available on a quarter basis from seven time points, t=1, ..., 7. In this study, 28 regions (NUTSIII) were used as domains of interest, i=1, ..., 28. A total of 1,000 Monte Carlo simulations were conducted from a pseudo-population of 4655 Portuguese companies of real estate mediation. For each simulation, samples were drawn independently for each time point using a stratified cluster sampling. These simulations aimed at an evaluation of the design-based properties of the EBLUP estimators used to estimate the mean price of the habitation transaction. Various linear mixed models used for finding out improved estimates of PHTS were used. They are summarized in table 1.

Table 1: Linear Mixed Models

Models	Structures	Fixed and random effects models	Estimators
1	AR(1)+J	$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t + v_i + u_{it} + e_{it},\ u_{it} = \rho u_{i,t-1} + \varepsilon_{it},\ \lvert\rho\rvert < 1$	$\hat{\mu}_{1,it}$
2	CSH	$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it} + e_{it}$	$\hat{\mu}_{2,it}$
3	ARH(1)	$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it} + e_{it}$	$\hat{\mu}_{3,it}$
4	CSH	$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t + u_{it} + e_{it}$	$\hat{\mu}_{4,it}$
5	ARH(1)	$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t + u_{it} + e_{it}$	$\hat{\mu}_{5,it}$

In what concerns to goodness-of-fit comparison between models with the same fixed effects but different covariance structures (models 1-3), the indices Akaike's Information Criterion and Schwartz's Bayesian Criterion show that AR(1)+J is the best choice of covariance structure.

The quality of the estimators of the parameter of interest was evaluated through the following precision and bias measures: absolute bias, variance, MSE and absolute bias ratio. To asses the relative gains of the proposed estimators, the ratio among absolute bias and precision measures of proposed estimators and Rao-Yu estimator, $\hat{\mu}_{1,it}$, was calculated. The 28 domains were divided in 6 groups mutually exclusive in function of its expected sample size, in order to facilitate the relative analysis in the average behaviour of the several estimators considered. Group 1 is the smallest ($\bar{n}_1 = 1$) and group 6 is the bigger ($\bar{n}_6 > 30$). The results of the evaluation measures averaged over small areas are reported in table 2. There is a clear pattern in the behaviour of various measures across different estimators. We have the following results from table 2:

a) All the estimators proposed performs overall better than the Rao-Yu estimator, except on group1 with respect to bias;

b) The estimators 3 and 5 (based on models with ARH(1) structures) perform slightly better than the estimators 2 and 4 (based on models with CSH structures), respectively;

c) The estimator 4 is quite similar to estimator 2, except on group 1 in which it is better;

d) The estimator 5 is quite similar to estimator 3, except on group 1 in which it is better.

The results show a considerable improvement of the statistical properties of the small area estimators based on models with heterogeneous covariance structures, except with respect to bias in very small domains. In these domains, estimators based on models with correlations decreasing for higher time lags perform better than the others.

Table 2: Average bias and precision measures of the EBLUP for groups of domains

Group	$\hat{\mu}_{1,it}$	$\hat{\mu}_{2,it}$	$\hat{\mu}_{3,it}$	$\hat{\mu}_{4,it}$	$\hat{\mu}_{5,it}$	$\hat{\mu}_{1,it}$	$\hat{\mu}_{2,it}$	$\hat{\mu}_{3,it}$	$\hat{\mu}_{4,it}$	$\hat{\mu}_{5,it}$
	Average Absolute Bias					Average Variance				
1	1.000	1.377	1.273	1.337	1.237	1.00	0.173	0.104	0.192	0.126
2	1.000	0.361	0.348	0.376	0.356	1.00	0.236	0.201	0.240	0.206
3	1.000	0.532	0.527	0.532	0.530	1.00	0.483	0.478	0.495	0.493
4	1.000	0.532	0.530	0.534	0.529	1.00	0.689	0.684	0.702	0.695
5	1.000	0.552	0.548	0.556	0.550	1.00	0.401	0.388	0.402	0.384
6	1.000	0.684	0.683	0.703	0.686	1.00	0.776	0.776	0.798	0.790
	Average MSE					Average Absolute Bias Ratio				
1	1.000	1.360	1.199	1.290	1.142	1.297	7.280	4.821	6.124	4.335
2	1.000	0.142	0.128	0.152	0.134	1.311	1.170	1.078	1.173	1.106
3	1.000	0.304	0.303	0.312	0.312	1.225	1.374	1.107	1.235	1.035
4	1.000	0.336	0.334	0.338	0.333	1.300	0.937	0.942	0.935	0.927
5	1.000	0.309	0.302	0.311	0.302	1.138	1.020	1.000	1.025	1.003
6	1.000	0.484	0.484	0.522	0.491	1.293	1.033	1.033	1.062	1.029

5. Concluding Remarks

We proposed a temporal area level model involving autocorrelated random effects with heterogeneous covariance structures. Under this general model, we obtained a BLUP and EBLUP estimators of a parameter of inferential interest. Our simulation results have shown that the estimators deducted from models involving autocorrelated random effects with heterogeneous covariance structures can lead to substantial gains on bias and on precision over the Rao-Yu estimator, especially when sample size is greater than one.

REFERENCES

Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

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

Pereira, L.N. (2005). *Estimação em Pequenos Domínios Utilizando Informação Seccional e Cronológica – O caso da estimação do preço médio de transacção da habitação*. MSc Thesis, Lisbon: ISEGI – New University of Lisbon.

Rao, J.N.K. and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, 22(4), 511-528.

Wolfinger, R.D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological and Environmental Statistics*, 1(2), 205-230.

ABSTRACT

This work intends to compare the performance of the temporal area level models with heterogeneous covariance structures of random effects with the Rao-Yu model (Rao and Yu, 1994). The properties of the EBLUP estimators are discussed from the design-based point of view.

RÉSUMÉ

Ce travail a comme but l'étude comparative de la performance des modèles temporels de niveau de secteur aux structures hétérogènes de covariance des effets aléatoires avec le modèle Rao - Yu (Rao and Yu, 1994). Les propriétés des estimateurs d'EBLUP sont discutées du point de vue design-based.