Proceedings of Q2008
European Conference on Quality in Official Statistics

Small area estimation for the price of habitation transaction: comparison of different uncertainty measures of temporal EBLUP

Luís N. Pereira, ESGHT – Universidade do Algarve[1]
Pedro S. Coelho, ISEGI – Universidade Nova de Lisboa

1. Introduction

Local level planning requires reliable statistics for small areas, but normally due to cost or logistic constraints, sample surveys are often planned to provide reliable estimates only for large geographical regions and large subgroups of a population. As a result, many domains of interest can be unplanned at the design stage and the sample sizes in these domains are rarely large enough to provide adequate precision for direct domain estimators. For example, data from the Prices of the Habitation Transaction Survey (PHTS) in Portugal only allows the production of reliable direct estimates for the mean price of habitation transaction for the country and for large geographical regions, like NUTSII (NUTS stands for the "Nomenclature of Territorial Units for Statistics" which is defined by the regulation EC No 1059/2003 of the European Parliament). In fact, the PHTS sample sizes are very small or even zero in many other domains of interest (e. g. NUTSIII, municipalities). For these unplanned small areas with small sample sizes, traditional direct estimators are either not feasible or provide unacceptably large variation coefficients. This creates the need to employ indirect estimators that "borrow information" from related small areas and time periods through linking models, using recent census and current administrative data, in order to increase the effective sample size and thus precision. Several indirect small area estimators based on explicit linking models have been proposed in recent years. For a review of indirect small area estimators, we refer to the book by Rao (2003).

In this paper, it is assumed that the small area parameters of interest follow a Rao-Yu longitudinal model (Rao and Yu, 1994) in order to combine information from longitudinal surveys and related auxiliary variables. The Rao-Yu model is a cross-sectional and time-series stationary area level model involving autocorrelated random effects and sampling errors with an arbitrary covariance matrix over time, and using analysis of variance (ANOVA) estimates of variance components. In fact, this model is a special case of the general linear mixed model (LMM) involving autocorrelated random effects with a homogeneous covariance structure (first-order autoregressive plus common covariance: AR(1)+J). Under this model, the empirical

[1] Address for correspondence: Luís N. Pereira, Escola Superior de Gestão, Hotelaria e Turismo – Universidade do Algarve, Campus da Penha, 8005-139 Faro, Portugal (Lmper@ualg.pt).

best linear unbiased prediction (EBLUP) approach is used for the estimation of small area parameters of interest.

While EBLUP estimators are fairly easy to obtain under the Rao-Yu model, measuring its quality is a challenging problem due to difficulties on estimating the mean squared prediction error (MSPE) of such estimators. Nevertheless, estimation of the MSPE of the EBLUP is of significant practical interest. Therefore, research on estimation of MSPE of EBLUP in small area estimation problems has received a considerable attention in recent years. See Rao (2003) and Jiang and Lahiri (2006) for a review of MSPE estimation.

The aim of this paper is to compare different uncertainty measures of the small area estimator based on the Rao-Yu model, using real data from the PHTS conducted by the Portuguese Statistical Office.

The paper is organized as follows. Section 2 reviews the cross-sectional and time-series stationary area level model due to Rao and Yu (1994) and describes how the EBLUP is obtained. Different MSPE estimators of the temporal EBLUP are presented in section 3. In the framework of an application with real data, section 4 addresses both the estimates of mean price of the habitation transaction and a comparison of different uncertainty measures of these estimates. Finally, the paper ends with a conclusion in section 5.

2. The Rao-Yu model

In order to take advantage of the chronological nature of data, Rao and Yu (1994) proposed the following area specific model:

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

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

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 (based only on the sample from the i^{th} small-area at t^{th} time point), e_{it}'s are independent sampling errors normally distributed, given the θ_{it}'s, with mean 0 and known variance σ_{it}^2, $\mathbf{x}_{it} = (x_{it1},...,x_{itp})'$ (p×1) is a column vector of an area-by-time specific auxiliary variables and $\boldsymbol{\beta}$ (p×1) is a column vector of regression parameters. Further, v_i's are random area specific effects with $v_i \overset{iid}{\sim} N\left(0,\sigma_v^2\right)$ and u_{it}'s are random area-by-time specific effects following a common AR(1) process for each i, with $\varepsilon_{it} \overset{iid}{\sim} N\left(0,\sigma^2\right)$. The random effects, v_i and u_{it}, represent the area and the area-by-time characteristics not accounted by the auxiliary variables. The constant ρ is a measure of the level of temporal autocorrelation. 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 stationary small area model:

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

Note that the well-known Fay-Herriot model (Fay and Herriot, 1979) may be obtained from model (3) setting $T = 1$, $\rho = 0$ and $\sigma^2 = 0$. 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 expressed 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*×1) is a column vector of 1's. Further, $\mathbf{e}$, $\mathbf{v}$ and $\mathbf{u}$ are mutually independent, with $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_m \otimes \boldsymbol{\Gamma})$, $\mathbf{v} \sim N(0, \sigma_v^2\mathbf{I}_m)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, where $\boldsymbol{\Gamma}$ (*T*×*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 LMM with block-diagonal homogeneous covariance structure, $Cov(\hat{\boldsymbol{\theta}}) = \mathbf{V} = block\ diag_{1\le i\le m}(\mathbf{V}_i)$ with $\mathbf{V}_i = \sigma^2\boldsymbol{\Gamma} + \sigma_v^2\mathbf{J}_T + \mathbf{I}_T$. 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 $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_t$ is the t^{th} row of $\boldsymbol{\Gamma}$. In practice, $\boldsymbol{\psi}$ is unknown and is estimated from the data. Let $\hat{\boldsymbol{\psi}}$ be a consistent estimator of $\boldsymbol{\psi}$. Then, an empirical BLUP of θ_{it}, say $\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})$, is obtained from $\tilde{\theta}_{it}(\boldsymbol{\psi})$ with $\boldsymbol{\psi}$ replaced by $\hat{\boldsymbol{\psi}}$.

Rao and Yu (1994) provided method of moments estimators of σ_v^2 and σ^2, through an extension of Henderson method 3 (Henderson, 1953), and proposed a naïve estimator of ρ. However, their main research focused on derivation of an EBLUP estimator of θ_{it} and an estimator of its MSPE, assuming known ρ. Thus, from this point forward we define the vector of variance components as $\boldsymbol{\psi} = [\sigma_v^2(\rho), \sigma^2(\rho)]'$.

3. Uncertainty measures of temporal EBLUP

While the temporal EBLUP is fairly easy to obtain, the estimation of its uncertainty is a challenging problem due to the variability caused by the estimation of the variance components. A naïve measure of uncertainty of the temporal EBLUP can be obtained from:

$$mspe_{it}^{N}\left[\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}). \quad (6)$$

where the terms $g_{1it}(\boldsymbol{\psi})$, which measures the uncertainty of the EBLUP due to the estimation of the random effects, and $g_{2it}(\boldsymbol{\psi})$ due to the estimation of the fixed effects, can be found in Rao and Yu (1994).

However, in practical applications this estimator could underestimate the true MSPE, especially in cases where $\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})$ varies with $\boldsymbol{\psi}$ to a significant extent and where the variability of $\hat{\boldsymbol{\psi}}$ is not small.

Fortunately, the difficult problem of estimating the MSPE of EBLUP estimators, taking the variability of the estimated variance components into account, has been faced in the small area literature by adopting different approaches. A well known general method is based on the Taylor series expansion of MSE under normality and valid for a general longitudinal linear mixed model (Prasad and Rao 1990; Datta and Lahiri 2000, Das *et al*. 2004). More recently, due to the advent of high-speed computers, resampling methods have been proposed under a general longitudinal LMM. Jiang *et al*. (2002) introduced a unified jackknife method, which is also valid for nonnormal and nonlinear mixed models and for M-estimators of model parameters, while Butar and Lahiri (2003) proposed a parametric bootstrap method based on the assumption of normality. Both resampling-based methods are applicable for various methods of estimating the variance components and are analytically less onerous than the Taylor series method. Furthermore, all these approaches are second order accurate.

Under normality assumptions of both the random effects and the random errors, an analytical estimator of the MSPE of temporal EBLUP under the Rao-Yu model is given by (Rao and Yu, 1994):

$$mspe_{it}^{RY}\left[\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) + 2g_{3it}(\hat{\boldsymbol{\psi}}). \quad (7)$$

where $g_{3it}(\boldsymbol{\psi})$, which measures the uncertainty due to the estimation of the variance components, can be found in Rao and Yu (1994).

Alternatively, Pereira and Coelho (2009) proposed resampling-based methods in order to measure the uncertainty of the temporal EBLUP under the Rao-Yu model, and using ANOVA estimates of variance components. They introduced a parametric bootstrap method based on the assumption of normality and a linearized weighted jackknife method.

In general, the parametric bootstrap method consists of generating parametrically a large number of area bootstrap samples (*B*) from the model fitted to the original data, re-estimating the model parameters for each bootstrap sample and then estimating the separate components of the MSPE. The bootstrap MSPE estimator of the temporal EBLUP under the Rao-Yu model is calculated with the following approximation:

$$mspe_{it}^{B}\left[\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = 2\left[g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}})\right] - B^{-1}\sum_{b=1}^{B}\left[g_{1it}(\hat{\boldsymbol{\psi}}^{*(b)}) + g_{2it}(\hat{\boldsymbol{\psi}}^{*(b)})\right] + g_{3it}^{*}. \quad (8)$$

where $\hat{\boldsymbol{\psi}}^{*}$ is the same as $\hat{\boldsymbol{\psi}}$ but it is calculated from the bootstrap data instead of the full data, and the terms $g_{1it}\left(\boldsymbol{\psi}^{*(b)}\right)$, $g_{2it}\left(\boldsymbol{\psi}^{*(b)}\right)$ and g_{3it}^{*} can be found in Pereira and Coelho (2009).

In general, the jackknife method consists of dropping out the j^{th} small area data set, j=1, ..., m, estimating the model parameters using the remaining data for each small area and then estimating the separate components of the MSPE. The linearized weighted jackknife MSPE estimator of the temporal EBLUP under the Rao-Yu model is given by:

$$mspe_{it}^{J}\left[\breve{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}}) - \mathbf{c}'_{WJ,t}(\hat{\boldsymbol{\psi}})\nabla g_{1t}(\hat{\boldsymbol{\psi}}) + tr\left[\mathbf{A}_t \boldsymbol{\upsilon}_{WJ,t}\right] + \\ + tr\left\{\mathbf{L}_t(\hat{\boldsymbol{\psi}})\left[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]\left[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})\right]' \mathbf{L}'_t(\hat{\boldsymbol{\psi}})\boldsymbol{\upsilon}_{WJ,t}\right\}, \quad (9)$$

where $\nabla g_{1t}(\boldsymbol{\psi}) = \left(\frac{\partial g_{1it}}{\partial \sigma^2}, \frac{\partial g_{1it}}{\partial \sigma_v^2}\right)'$, $\mathbf{L}_t(\boldsymbol{\psi}) = \left(\frac{\partial \mathbf{b}_t}{\partial \sigma^2}, \frac{\partial \mathbf{b}_t}{\partial \sigma_v^2}\right)'$, $\mathbf{c}_{WJ,t} = \sum_{j=1}^{m} w_{jt}\left(\hat{\boldsymbol{\psi}}_{-j} - \hat{\boldsymbol{\psi}}\right)$ is a weighted jackknife estimator of the bias of $\hat{\boldsymbol{\psi}}$, $\boldsymbol{\upsilon}_{WJ,t} = \sum_{j=1}^{m} w_{jt}\left(\hat{\boldsymbol{\psi}}_{-j} - \hat{\boldsymbol{\psi}}\right)\left(\hat{\boldsymbol{\psi}}_{-j} - \hat{\boldsymbol{\psi}}\right)'$ is a weighted jackknife estimator of the covariance matrix of $\hat{\boldsymbol{\psi}}$. Here, $\hat{\boldsymbol{\psi}}_{-j}$ is the estimator of $\boldsymbol{\psi}$ after deleting the j^{th} small-area data and w_{jt} are the weights: $w_{jt} = 1 - \mathbf{x}'_{jt}\left(\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}'_{it}\right)^{-1}\mathbf{x}_{jt}$. For more details about these specific resampling methods see Pereira and Coelho (2009).

4. Application

The Prices of the Habitation Transaction Survey, PHTS, was a longitudinal survey conducted by the Portuguese Statistical Office with the goal to collect data about the prices of the habitation transaction. The target population of the PHTS survey consisted of all habitation transactions made within each reference period. The primary sampling units (PSU) were companies of real estate mediation. The survey was based on a stratified cluster sampling without replacement, in which strata were formed by grouping the PSUs (companies of real estate mediation) according to geographic regions (NUTSIII, municipalities) and gross sales. The companies were selected through random simple sampling with probability proportional to size and all the secondary units (habitation transactions) within each selected PSU were observed. The average sample size of PSUs was 458 companies of real estate mediation per wave.

The main goal was the estimation of the mean price of the habitation transaction per square meter by NUTSII level. However, due to new demands, estimates at NUTSIII and municipality levels were also required. For these unplanned small areas with

small sample sizes, traditional direct estimators are either not feasible or provide unacceptably large variation coefficients.

Therefore, indirect estimators are needed in order to “borrow information” from related small areas and time periods through linking models, using recent census and current administrative data. In this application, we use the temporal EBLUP assisted by the Rao-Yu model. As the Portuguese Statistical Office also conducted the Prices of Bank Evaluation in the Habitation Survey (PBEHS), we decided to use the bank evaluation of the habitations as auxiliary variable. Both the target variable and the auxiliary variable are measured in euros by square meter. Furthermore, the data are available on a quarter basis from seven time points, *t*=1, …, 7.

Table 1. Small areas, sample sizes, EBLUP estimates of mean price of the habitation transaction in the 7th wave and their naïve, analytical, bootstrap and jackknife MSPE estimates

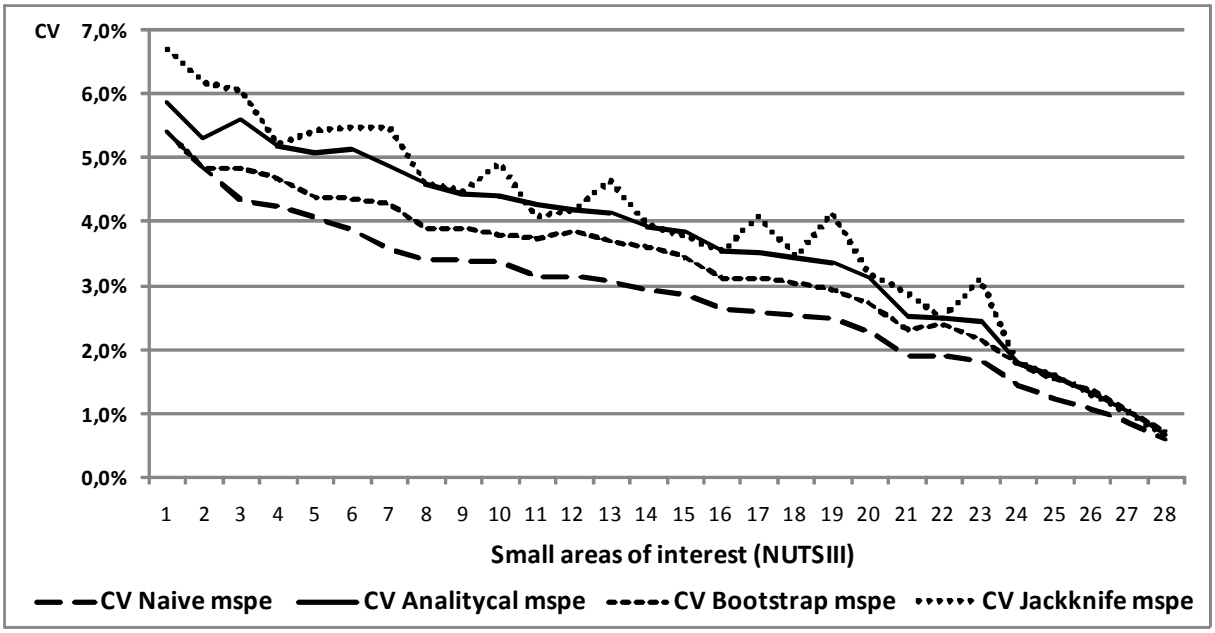
NUTSIII	n_{i7}	$\tilde{\mu}_{i7}^{EBLUP}$	$mspe^{N}$	$mspe^{RY}$	$mspe^{B}$	$mspe^{J}$
1	1	646	1220	1437	1859	1220
2	1	714	1201	1433	1934	1201
3	7	661	824	1375	1596	1027
4	6	718	933	1395	1389	1126
5	31	763	956	1511	1708	1120
6	18	704	745	1306	1485	948
7	19	670	575	1064	1349	818
8	56	756	666	1216	1201	868
9	12	769	676	1166	1177	902
10	23	820	766	1303	1609	969
11	19	804	643	1182	1074	903
12	17	666	439	786	771	653
13	22	710	471	866	1086	682
14	27	658	374	667	675	568
15	39	735	437	798	773	646
16	34	867	523	948	931	732
17	17	814	441	825	1103	651
18	40	876	487	909	910	704
19	12	960	559	1045	1559	791
20	24	974	493	921	967	704
21	77	937	320	561	731	473
22	49	866	272	470	459	426
23	26	1128	418	754	1217	593
24	90	956	192	295	291	287
25	89	1172	205	337	348	319
26	405	1041	126	185	182	201
27	263	1073	85	115	113	123
28	488	1321	60	75	74	92

In this application, the target parameter of interest is the mean price of the habitation transaction per square meter by NUTSIII level (28 small areas, *i*=1, ..., 28). Estimatives of this parameter are given by the temporal EBLUP after fitting the Rao-Yu model to PHTS and PBEHS data. Afterwards, we analyse different measures of uncertainty of this EBLUP: the naive (6), the analytical (7), the bootstrap (8) and the linearized weighted jackknife (9) MSPE estimators.

Table 1 shows by small areas (NUTSIII), the sample sizes in the 7th wave (n_{i7}), EBLUP estimates of mean price of the habitation transaction in the same wave ($\breve{\mu}_{i7}$) and the corresponding uncertainty measures over small areas.

The coefficients of variation (vc) of temporal EBLUP estimates of mean price of the habitation transaction are plotted in figure 1. The coefficient of variation is defined as $\sqrt{mspe_{it}\left[\breve{\theta}_{it}(\hat{\psi})\right]}\Big/\breve{\theta}_{it}(\hat{\psi})$. From this figure we can observe a similar behavior for all uncertainty measures of the temporal EBLUP. The Jackknife estimator is the one exhibiting a more specific pattern as it does not show a totally monotonic decrease of the variation coefficients along with areas 1 to 28. We can also see that the bootstrap-based coefficients of variation are lower than the analytical and jackknife-based ones. The naïve estimator is the one producing the smaller variation coefficients, as was expected according to the theory, since this estimator is known to underestimate the true MSPE. The results confirm a good precision of the EBLUP estimates even when the sample sizes are very small, since all coefficients of variation are below 7%. Obviously the precision of the EBLUP estimates for small areas 24-28 are also very good, since they have large sample sizes.

Figure 1. Coefficients of variation for the EBLUP estimates of mean price of the habitation transaction in the 7th wave



5. Concluding remarks

Statistical offices are currently required to provide small area estimates, but as sample surveys are usually designed to produce estimates for large planned domains, model-based methods can be used to provide reliable indirect estimators at small area level.

In this paper, a temporal EBLUP estimator was used to provide estimates of mean price of the habitation transaction at NUTSIII level. The coefficients of variation of these estimates were assessed by the naive, analytical, bootstrap and weighted jackknife MSPE estimators.

The results reveal that all MSPE estimators present similar behavior over the small areas, although the naive MSPE estimator tends to underestimate the MSE, as was said previously.

Results confirm that the use of the temporal EBLUP offers good precision when estimating the price of habitation transaction even for domains with very small sample size.

Furthermore, the results suggest that the linearized weighted jackknife MSPE estimator could be the best resampling-based estimator to assess the uncertainty of the mean price of the habitation transaction EBLUP estimates. This is due to its similarity to the analytical MSPE estimator and its ability to produce results less smooth than the ones achieved by the other estimators.

Acknowledgements

The authors acknowledge the Portuguese Statistical Office for the availability of the data used in the research. The first author's research was supported in part by the Portuguese Foundation for Science and Technology (fellowship SFRH/BD/36764/2007).

References

Butar, F.B., and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76.

Das, K., Jiang, J., and Rao, J.N.K. (2004), "Mean squared error of empirical predictor", *The Annals of Statistics*, 32, pp. 818-840.

Datta, G.S., and Lahiri, P. (2000), "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems", *Statistica Sinica*, 10, pp. 613-627.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of income for small places: An

application of James-Stein procedures to census data", *Journal of the American Statistical Association*, 74, pp. 269-277.

Henderson, C.R. (1953), "Estimation of variance and covariance components", *Biometrics*, 9, pp. 226-252.

Jiang, J., and Lahiri, P. (2006), "Mixed Model Prediction and Small Area Estimation", *Test*, *15(1)*, pp. 1-96.

Jiang, J., Lahiri, P. and Wan, S.-M. (2002), "A unified jackknife theory for empirical best prediction with M-estimation", *The Annals of Statistics,* 30, pp. 1782-1810.

Pereira, L.N., and Coelho, P.S. (2009), "Assessing different uncertainty measures of EBLUP: a resampling-based approach**",** *Journal of Statistical Computation and Simulation*, forthcoming.

Prasad, N.G.N., and Rao, J.N.K. (1999), "On robust small area estimation using a simple random effects model", *Survey Methodology*, 25(1), pp. 67-72.

Rao, J.N.K. (2003), *Small area estimation*, New Jersey: John Wiley & Sons.

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), pp. 511-528.