Small Area Estimation of the Mean Price of the Habitation Transaction in Portugal: Methodological and Practical Issues

Pereira, Luis N.
University of the Algarve - ESGHT, 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

The Portuguese National Statistical Institute intends to produce estimations for the mean price of the habitation transaction for the NUTS III and for Portugal municipalities, using data collected from a repeated survey. 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. This is a fertile ground to the use of auxiliary information and observations of the interest variable from related small areas or periods in time, in order to increase the "effective" sample size in the domain of interest.

The main purpose of this study is to propose new estimator for the mean price of the habitation transaction in small domains with area level data. An area level model with heterogeneous covariance structures of random effects, as a vehicle for borrowing strength across the areas and over time, assists the proposed combined estimator. The proposed model includes random area-by-time specific effects. Furthermore, the auxiliary variables are related to the parameters of inferential interest, θ_{it}, in the i^{th} small-area at t^{th} time point (i=1, …, m; t=1, …, T) through a linear model:

(1) $$\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t + u_{it},$$

where $\mathbf{x}_{it}$ (p×1) is a column vector of area-by-time specific auxiliary variables and $\boldsymbol{\beta}_t$ (p×1) 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 $E(u_{it}) = 0$, $\text{cov}(u_{it}, u_{i't'}) = \sigma_{u,tt'}$ for i=i' and 0 otherwise. Moreover, it assumes that direct estimators, $\hat{\theta}_{it}$, are available and design-unbiased:

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

where e_{it}'s are independent sampling errors normally distributed, given the θ_{it}'s, with mean 0 and known variance $\sigma^2_{e,it}$. Combining the sampling error model (1) with the linking model (2), we obtain the following model:

(3) $$\hat{\theta}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_t + u_{it} + e_{it}.$$

with $E(e_{it}u_{it}) = 0$. This model is an extension of a model by 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. A time series and cross-sectional area level model (3) may present any type of chronological covariance structure of the random effects. A heterogeneous first-order autoregressive structure of the random effects is used, $\sigma_{u,tt'} = \sigma_{u,t}\sigma_{u,t'}\rho^{|t-t'|}$, $|\rho| < 1$, t,t'=1, …, T, (Wolfinguer, 1996). We selected a first-order autoregressive structure because it should be

reasonable to assume that the random effects associated to a particular i^{th} small-area are correlated and correlation decays to zero as the time goes by. Further, we selected a heterogeneous covariance structure due to the observed increase of variability with time. Following Henderson's general results (Henderson, 1975) and assuming that $\boldsymbol{\psi} = (\sigma^2_{u,1},...,\sigma^2_{u,T},\rho)'$ is known, the Best Linear Unbiased Predictor (BLUP) of θ_{it} is given by:

$$\tilde{\theta}_{it}(\boldsymbol{\psi}) = \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_t + \sum_{t'=1}^{T} h_{itt'}\left(\theta_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_t\right), \tag{4}$$

where $\tilde{\boldsymbol{\beta}}_t$ is the generalized least squares of $\boldsymbol{\beta}_t$ and $h_{itt'}$ is a weight, which depends on the chronological covariance structure. This estimator can be classified as a combined estimator, since it can be decomposed into two components: a synthetic estimator, $\mathbf{x}'_{it}\boldsymbol{\beta}$, and a correction factor, $\sum_{t'=1}^{T} h_{itt'}\left(\theta_{it} - \mathbf{x}'_{it}\tilde{\boldsymbol{\beta}}_t\right)$, which is a function of the differences between the direct and synthetic estimates for the same domain. We can say that the weights, $h_{itt'}$, allow a correction of the synthetic part of the estimator (4) through the regression residuals from the domain that is the target of inference at t^{th} time point, but also from this domain at previous time points. These characteristics seem to be particularly interesting when estimating in small areas, where the available sample size is small, since it borrows information from outside the domain of study in order to assist the estimation. The Mean Squared Error (MSE) of BLUP can be decomposed in a sum of two components (Henderson, 1975). The first component of MSE of $\tilde{\theta}_{it}(\boldsymbol{\psi})$, due to estimating the random effects, is of order $o(1)$ and is given by:

$$g_{1it}(\boldsymbol{\psi}) = \sigma_{u,tt} - \mathbf{g}'_{it}\mathbf{V}_i^{-1}\mathbf{g}_{it}\,, \tag{5}$$

where $\mathbf{g}_{it} = col_{1\le t'\le T}(\sigma_{u,t't})$ is a column vector of order *T*, corresponding to the t^{th} column of covariance matrix of random effects and $\mathbf{V}_i$ is the covariance matrix of $\hat{\boldsymbol{\theta}}_i = col_{1\le t\le T}(\hat{\theta}_{it})$. The second component of MSE of $\tilde{\theta}_{it}(\boldsymbol{\psi})$, due to estimating the fixed effects of the model, is of order $o(m^{-1})$ for large *m* and is given by:

$$g_{2it}(\boldsymbol{\psi}) = \left(\mathbf{x}'_{it} - \mathbf{g}'_{it}\mathbf{V}_i^{-1}\mathbf{X}_i\right)\left(\sum_{i=1}^{m}\mathbf{X}'_i\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\mathbf{x}'_{it} - \mathbf{g}'_{it}\mathbf{V}_i^{-1}\mathbf{X}_i\right)', \tag{6}$$

where $\mathbf{X}_i = diag_{1\le t\le T}(\mathbf{x}'_{it})$.

In practice, the variance components, $\boldsymbol{\psi}$, are usually unknown and they have to be estimated in a consistent way. There are several methods that can be used to produce consistent estimates of the variance components in a Linear Mixed Model. Nevertheless, none of these methods is suited to estimate the variance components of the proposed model, since they were developed for particular cases of cross-sectional models or time-series and cross-sectional models. In the context of time-series and cross-sectional models, Pantula and Pollock (1985) estimated the variance components in the nested error-regression model with autocorrelated random effects, while Rao and Yu (1994) estimated the variance components in the model with both autocorrelated random effects and sampling errors, involving a homogeneous covariance structure. In both cases, they extended a simple transformation method by Fuller and Battese (1973), for the special case of independent errors and known autocorrelation coefficient, ρ.

In this way, it is convenient to extend the transformation method by Fuller and Battese (1973) to the context of the proposed model, involving random effects with a general covariance structure with the form $\sigma_{u,tt'} = f(\sigma_{u,t},\sigma_{u,t'},\rho)$, and independent sampling errors. We also propose an extension of the Analysis of Variance (ANOVA) method to estimate the variance components, used by Prasad and Rao (1999) in the scope of the pseudo-empirical BLUP estimation, to the context of the proposed model. Furthermore, these two methods take into account the sampling design in the estimation of the sampling errors variance. This is a way of introducing some information about the sampling design in the estimation of variance components. On the contrary of the likelihood methods, none of the proposed methods to estimate the variance components requires the normality neither of the random effects nor of the sampling errors.

Regarding the results provided by Searle, Casella and McCulloch (1992), from the extension of the ANOVA method the unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2_{AN} = \frac{\sum_{i=1}^{m}\sum_{j\in s_{ad}}\left(y_{ij}-\bar{y}_i\right)^2}{\sum_{i=1}^{m} n_i - m}, \tag{7}$$

where $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ is the small area sample mean, and the ANOVA unbiased estimator of σ_u^2, truncated to zero, is given by:

$$\hat{\sigma}^2_{u,AN} = \max(\tilde{\sigma}^2_{u,AN}; 0), \tag{8}$$

where $\tilde{\sigma}^2_{u,AN} = \left(n^*\right)^{-1}\left\{Q_b - [m-1]\hat{\sigma}^2_{AN}\right\}$ with $n^* = \sum_{i=1}^{m} n_i - \frac{\sum_{i=1}^{m} n_i^2}{\sum_{i=1}^{m} n_i}$, $Q_b = \sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2$ and $\bar{y} = \frac{\sum_{i=1}^{m} n_i \bar{y}_i}{\sum_{i=1}^{m} n_i}$. The variance estimator of the sampling errors, $\sigma^2_{\varepsilon,i}$, that take into account the sampling design through $w_{ij} = \pi_{ij}^{-1} \Big/ \sum_{j\in s_i} \pi_{ij}^{-1}$, is given by:

$$\hat{\sigma}^2_{e,i,AN} = \hat{\sigma}^2_{AN} \sum_{j\in s_i} w_{ij}^2 . \tag{9}$$

The extension of the Fuller-Battese transformation method involves performing two ordinary least squares (OLS) regressions and then using the method of moments to get unbiased estimators of σ^2 and σ_u^2. As a result we have the following unbiased and consistent estimator of σ^2:

$$\hat{\sigma}^2_{FB} = \frac{SSE_1}{m-p}, \tag{10}$$

where SSE_1 is the residual sum of squares of the first OLS regression. An unbiased estimator of σ_u^2 is given by:

$$\tilde{\sigma}^2_{u,FB} = \left(\sum_{i=1}^{m} f_i^{-2} - \eta^*\right)^{-1}\left\{SSE_2 - [m-p]\hat{\sigma}^2_{FB}\right\}, \tag{11}$$

where $f_i = \sum_{i=1}^{n_i} w_{ij}^2$, $w_{ij} = \pi_{ij}^{-1} \Big/ \sum_{j\in s_i} \pi_{ij}^{-1}$, η^* is a constant and SSE_2 is the residual sum of squares of the second OLS regression. Since $\tilde{\sigma}^2_{u,FB}$ can take on negative values, we truncate $\tilde{\sigma}^2_{u,FB}$ to zero whenever it is negative: $\hat{\sigma}^2_{u,FB} = \max\left(\tilde{\sigma}^2_{u,FB}; 0\right)$. The truncated estimator $\hat{\sigma}^2_{u,FB}$ is no longer unbiased. An estimator of the variance, $\sigma^2_{e,i}$, that take into account the sampling design, is given by:

$$\hat{\sigma}^2_{e,i,FB} = \hat{\sigma}^2_{FB} \sum_{j\in s_i} w_{ij}^2 . \tag{12}$$

Further, in this work we propose the estimation of the autocorrelation coefficient, ρ, through a naive estimator proposed by Pantula and Pollock (1985). All variance components are estimated externally for each time point.

The Empirical Best Linear Unbiased Predictor (EBLUP) is obtained by replacing the parameter vector, $\boldsymbol{\psi}$, by an asymptotically consistent estimator $\hat{\boldsymbol{\psi}} = (\hat{\sigma}^2_{u,1},...,\hat{\sigma}^2_{u,T},\hat{\rho})'$ in equation (4). The resulting

EBLUP is given by $\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})$. While 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. In that case, this problem has an additional difficulty due to the number of variance components needed to be estimated. A naïve measure of uncertainty of the EBLUP, $\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})$, can be obtained from the MSE of BLUP, by replacing the parameter vector $\boldsymbol{\psi}$ by its estimator $\hat{\boldsymbol{\psi}}$: $mse_N\left[\hat{\theta}_{it}(\hat{\boldsymbol{\psi}})\right] = g_{1it}(\hat{\boldsymbol{\psi}}) + g_{2it}(\hat{\boldsymbol{\psi}})$.

On the basis of the methodological aspects, this work intends to compare the performances of standard and proposed methods to estimate the mean price of the habitation transaction at NUTS III level. A Monte Carlo simulation study was carried out to perform empirical analysis. The simulation study has been based on 1000 longitudinal independent samples drawn from a pseudo population of 4655 Portuguese companies of real state mediation, using a stratified-cluster sample design (NUST III-companies of real state mediation). 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) were used. By means of a set of precision and bias measures, this simulation was used to analyse the relative merits of the estimators proposed in comparison to various other direct and indirect estimators that have been used in small area estimation. The statistical properties of the estimators were evaluated under a design-based approach. The results show that the proposed estimator performs better than other direct and indirect estimators usually used in small area estimation with respect to the bias and precision measures. Further, the heterogeneous covariance structure and the chronological autocorrelation of random effects may be used for improving the direct survey estimates of the mean price of the habitation transaction at NUTS III level. The results of the empirical study also show that the naïve estimator of the model-based MSE of the proposed presents significant bias in a great part of the domains.

Key words: Small area estimation; Chronological autocorrelation; Combined estimator; Model-assisted estimation; Empirical best linear unbiased predictor; Estimation of variance components.

References:

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,* 269-277.

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.

Pantula, S.G. and Pollock, K.H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics, 41,* 909-920.

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), 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), 511-528.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance components*. New York: John Wiley & Sons.

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