# Estimation of unemployment rates in small areas of Portugal: a best linear unbiased prediction approach versus a hierarchical Bayes approach

**Luis N. Pereira[1,2] Jorge M. Mendes[3,4,5] and Pedro S. Coelho[3,4]**

[1] Escola Superior de Gestão, Hotelaria e Turismo, Universidade do Algarve, Portugal
[2] Centro de Investigação sobre o Espaço e as Organizações, Universidade do Algarve, Portugal
[3] Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Portugal
[4] Centro de Estatística e Gestão de Informação, ISEGI–UNL, Portugal
[5] Statistics Portugal, Portugal
**Email**: Lmper@ualg.pt

## Abstract

The high level of unemployment is one of the major problems in most European countries nowadays. Hence, the demand for small area labor market statistics has rapidly increased over the past few years. The Labour Force Survey (LFS) conducted by the Portuguese Statistical Office is the main source of official statistics on the labour market at the macro level (e.g. NUTS2 and national level). However, the LFS was not designed to produce reliable statistics at the micro level (e.g. NUTS3, municipalities or further disaggregate level) due to small sample sizes. Consequently, traditional design-based estimators are not appropriate. A solution to this problem is to consider model-based estimators that "borrow information" from related areas or past samples by using auxiliary information. This paper reviews, under the model-based approach, Best Linear Unbiased Predictors and an estimator based on the posterior predictive distribution of a Hierarchical Bayesian model. The goal of this paper is to analyze the possibility to produce accurate unemployment rate statistics at micro level from the Portuguese LFS using these kinds of estimators. This paper discusses the advantages of using each approach and the viability of its implementation.

**Keywords**: Best linear unbiased prediction, Hierarchical Bayes, Small area estimation, Unemployment rate

**AMS subject classifications**: 62F15, 62J05, 62P20

### 1. Introduction

The Labour Force Survey (LFS) conducted quarterly by the Portuguese Statistical Office is of capital importance nowadays and it is used for a variety of purposes. One of the most relevant purposes is the estimation of the unemployment rate, a key indicator of the whole economy, but there are also other important official statistics derived from that survey. The Portuguese Statistical Office disseminates quarterly and annually labour market statistics at the macro level (e.g. NUTS2 and national level), but there is a growing demand of such official statistics at the micro level (e.g. NUTS3, municipalities or further disaggregate level). However, the LFS was not designed to produce reliable direct estimates at micro level due to small sample sizes which lead to high levels of sampling variability. A solution to this problem is to consider model-based small area estimators that "borrow information" from related areas or past samples by using auxiliary information. The main idea of model-based estimators is to use explicit or implicit models that connect the different small areas by means of auxiliary information available from previous studies or obtained from different sources, in order to increase the effective sample size and thus precision. Depending on whether this information is available for every unit or only for the domains of interest, models will be formulated at unit or area level respectively. [5], [10] and [8] give an account of different model-based techniques to accomplish small area estimation tasks. One of the most popular

techniques is based on explicit Linear Mixed Models (LMM) that provide a link to a related small area through the use of supplementary data. The empirical best linear unbiased prediction (EBLUP) approach, using Henderson's method ([7]), is one of the most widely used technique for estimating small area parameters of interest. In this approach, the best linear unbiased predictor (BLUP) of the small area is first produced using the general theory of [7] and then the unknown variance components are estimated by a standard method (e.g. ANOVA, maximum likelihood, residual maximum likelihood, etc.). Another approach widely used to estimate small area parameters is based on a Hierarchical Bayesian (HB) framework. In the HB approach, a prior distribution on the model parameters and hyperparameters has to be specified and inferences are then based on the mean of the posterior distribution.
The aim of this paper is to analyze the possibility to produce accurate unemployment rate statistics at micro level from the Portuguese LFS using Best Linear Unbiased Predictors and predictors based on the posterior predictive distribution of a Hierarchical Bayesian model. We adopt area level models since covariables are available only at the area level.

## 2. An overview of the methodology

We now set out the small area estimation techniques used to produce the model-based unemployment rate estimates.

### 2.1. The BLUP approach

Consider the following general LMM in small area estimation used by [2]:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\upsilon_i + \varepsilon_i, \tag{1}$$

where $\mathbf{X}_i$ $(n_i \times p)$ and $\mathbf{Z}_i$ $(n_i \times b_i)$ are known matrices, $\upsilon_i$ and $\varepsilon_i$ are independently distributed with $\upsilon_i \sim N(\mathbf{0}, \mathbf{G}_i)$ and $\varepsilon_i \sim N(\mathbf{0}, \mathbf{R}_i)$, $i$=1,..., $m$. We assume that $\mathbf{G}_i = \mathbf{G}_i(\psi)$ $(b_i \times b_i)$ and $\mathbf{R}_i = \mathbf{R}_i(\psi)$ $(n_i \times n_i)$ possibly depend on $\psi = (\psi_1, ..., \psi_q)'$, a $q \times 1$ vector of variance components. Arranging the data as $\mathbf{y} = col_{1\leq i\leq m}(\mathbf{y}_i)$, $\mathbf{X} = col_{1\leq i\leq m}(\mathbf{X}_i)$, $\mathbf{Z} = diag_{1\leq i\leq m}(\mathbf{Z}_i)$, $\upsilon = col_{1\leq i\leq m}(\upsilon_i)$ and $\varepsilon = col_{1\leq i\leq m}(\varepsilon_i)$, model (1) may be written as $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\upsilon + \varepsilon$, where $\upsilon$ and $\varepsilon$ are independently distributed with $\upsilon \sim N(\mathbf{0}, \mathbf{G})$ and $\varepsilon \sim N(\mathbf{0}, \mathbf{R})$, $n = \sum_{i=1}^{m} n_i$ and $b = \sum_{i=1}^{m} b_i$. The variance-covariance matrix of $\mathbf{y}$ is given by $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, where $\mathbf{G} = diag_{1\leq i\leq m}(\mathbf{G}_i)$ and $\mathbf{R} = diag_{1\leq i\leq m}(\mathbf{R}_i)$. In the context of model-based small area estimation we are interested in predicting a general mixed effect of $\theta_i = \mathbf{k}_i'\beta + \mathbf{m}_i'\upsilon$, where $\mathbf{k}_i'$ and $\mathbf{m}_i'$ are known vectors of order $p \times 1$ and $b \times 1$, respectively. Then, assuming that $\psi$ is known, we can use the Henderson's 1975 results for general LMM involving fixed and random effects ([7]) to obtain the BLUP of $\theta_i$

$$\widetilde{\theta}_i(\psi) = \mathbf{k}_i'\widetilde{\beta}(\psi) + \mathbf{m}_i'\widetilde{\upsilon}(\psi), \tag{2}$$

where $\widetilde{\beta}(\psi) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and $\widetilde{\upsilon}(\psi) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widetilde{\beta})$. However, $\psi$ is unknown and should be estimated from the data. Let $\widehat{\psi}$ be a consistent estimator of $\psi$. Then, an empirical BLUP (EBLUP) of $\theta_i$, say $\widehat{\theta}_i(\widehat{\psi}) = \mathbf{k}_i'\widehat{\beta} + \mathbf{m}_i'\widehat{\upsilon}$, is obtained from (2) with $\psi$ replaced by $\widehat{\psi}$. Model (1) covers the following area-level small area models used in this research: the Fay–Herriot model ([4]), the temporal model ([11]) and the spatial model ([9]).

### 2.2. The HB approach

Bayesian alternatives to the models referenced in previous section have been proposed (e.g. [3]; [13]). The advantage of a Bayesian approach is that it offers a coherent framework for combining different sources of information. For the direct LFS quarterly unemployment rate estimate $y_{it}$, let $\theta_{it}$ be the true unemployment rate for the $i^{th}$ small area at a particular quarter $t$. The sampling model for $y_{it}$ can be expressed as $y_{it} = \theta_{it} + \epsilon_{it}$, where $\epsilon_{it}$ is the sampling error associated with the direct estimate $y_{it}$, and is assumed to be spatially independent, but time correlated. Assuming the existence of auxiliary information on $p$ auxiliary variables at area level, the linking model for the true unemployment rate $\theta_{it}$ is written as $\log(\theta_{it}) = \mathbf{x}_{it}'\beta + \upsilon_i + u_{it}$, where $\mathbf{x}_{it}' = (x_{it1}, ..., x_{itp})$ is a vector of auxiliary variables, $\beta = (\beta_1, ..., \beta_p)$ is a vector $p$ regression coefficients (unknown fixed effects), $\upsilon_i$ is a spatial hidden process and $u_{it}$ is a

space-time hidden process whose role is to borrow strength over spatially adjacent small areas and consecutive time periods (taking advantage of the rotation pattern of the LFS sample which overlap 5/6 of the sample between consecutive quarters in the same area). The hidden process $v$ which introduces spatially structured random effects follows an Intrinsically conditional, nearest neighbour Autoregressive (IAR) model ([1]) defined over the small areas. The hidden process $u$ which introduces space-time random effects follows a random walk process over time, that is $u_{it} = u_{it-1} + \varepsilon_{it}$, where $\varepsilon_{it}$ is white noise. The regression coefficients $\beta$ are assumed to follow a non-informative Gaussian prior distribution and variance of the hyperparameters are assumed to follow non-informative Gamma prior distributions. We use the special case of Markov Chain Monte Carlo (MCMC) methods known as Gibbs sampling ([6]) to evaluate the joint posterior distribution numerically.

## 3. Application

The target population of the LFS includes all persons living in private households on the Portuguese economic territory. The survey follows a stratified two-stage cluster design. The primary sampling units (PSU) are census areas (*areas da amostra mãe*) which are grouped into seven strata (NUTSII). For each PSU selected, a sample of private households is drawn and all of their residents aged 15 or more are inquired. The target parameter of interest was defined as the unemployment rate, for each small area at each time period. In all models we have considered the unemployment rate accounted by the Institute for Employment and Vocational Training (IEFP) as the auxiliary variable. The data were available on a quarter basis from 18 time points ($t$=1, ..., 18). In this research 24 small areas were used ($i$=1, ..., 24). These small areas of interest, designated as Labour Market Areas (LMA), are defined by aggregations of municipalities.

In this research we have used the following EBLUP estimators: the Fay–Herriot estimator, the spatial estimator and the temporal estimator. Further, we also have used an estimator based on the mean of the posterior distribution, as described in section 2.2 (the model runs a chain of 100,000 iterations and the first 10,000 were deleted as "burn-in" period; we sampled every 100 simulated and use a sample of 900 observations of the posterior distribution to compute the posterior mean). The precision of these model-based estimators was compared with the precision of the direct design-based estimator. In order to summarize results, table 1 presents the average coefficient of variation (CV) of the unemployment estimates for all small areas and time periods. For the HB modelling approach the CV was computed as the ratio of standard deviation and mean of the posterior distribution of $\theta$. The estimated CV shows that model-based estimates have a higher degree of precision when compared to the direct estimates. This is true for both BLUP and HB approaches, however it is not clear which approach is better. In order to make an in-depth analysis of the relative performance of the model-based estimators, we are conducting a design-based simulation study in which 1,000 samples were drawn from a pseudo-population based on a real population and a realistic sampling design.

Table 1: Global average CV of the unemployment estimates

| **Estimator** | Direct | Fay–Herriot EBLUP | Spatial EBLUP | Temporal EBLUP | HB estimator |
|---|---|---|---|---|---|
| **CV** | 20.6% | 15.2% | 16.8% | 11.0% | 15.7% |

**Acknowledgements**: This paper was partially financed by *Fundação para a Ciência e a Tecnologia.*

## References

1.Besag, J., Green, P., Higdon, D. and Mengeresen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* 10, 3–66.
2.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, 613–627.
3.Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association* 94, 1074–1082.
4.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.
5.Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science* 9, 55–93.
6.Gilks, W.R., Richardson, S. and Spiegelhalter, D. (1996). *SMarkov Chain Monte Carlo in practice. Interdisciplinary Statistics*, Chapman & Hall, Boca Raton.
7.Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447.
8.Jiang, J. and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test* 15, 1–96.
9.Petrucci, A., Pratesi, M. and Salvati, N. (2005). Geographic information in small area estimation: small area models and spatially correlated random area effects. *Statistics in Transition* 7, 609–623.
10.Rao, J.N.K. (2003). *Small area estimation*, John Wiley & Sons, New Jersey.
11.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, 511–528.
12.You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology* 34, 19–27.
13.You, Y. and Rao, J.N.K. (2002). Hierarchical Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics* 30, 3–15.
14.You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-Based unemployment Rate Estimation for the Canadian Labour Force Survey: A Hierarchical Bayes Approach. *Survey Methodology* 29, 25–32.