An Empirical Comparison of EBLUP Estimation and Model Based Direct Estimation for Small Areas

Hukum Chandra, and Ray Chambers

1. Introduction

Unit level random effect models are often used in small area estimation, and the empirical best linear unbiased prediction (EBLUP) based approach is widely used for the estimation of small area means under such models (Rao, 2003). However, this approach does not lead to small area estimators that are weighted linear functions of the sample data from these areas. As a result, several practical advantages of using such weighted estimators are lost, with probably the most important being the relative simplicity of their mean squared error estimation. The model-based direct (MBD) approach (Chambers, 2005) overcomes some of these limitations. This approach uses sample weights derived from a population level version of the random effects model to define weighted linear small area estimators as well as a simple expression for their mean squared error. The objective of this paper is to provide some empirical results that compare the mean squared error estimates generated under the MBD approach with those generated under the EBLUP approach. The robustness of the two approaches to model misspecification is also examined.

2. Small Area Estimation Based on a Linear Mixed Model

Let \( Y_i \) be the \( N_i \times 1 \) vector of values of variable of interest in small area \( i \) and let \( X_i \) be the \( N_i \times p \) matrix of values of the associated auxiliary variables. We consider the following linear mixed model specification for the distribution of \( Y_i \) given \( X_i \):

\[
Y_i = X_i \beta + Z_i u_i + e_i .
\]

(1)

Here \( N_i \) is the number of the population units in small area \( i \), \( \beta \) is a \( p \times 1 \) vector of fixed effects, \( Z_i \) is a \( N_i \times q \) matrix of known covariates characterising differences between small areas, \( u_i \) is a random area effect associated with the \( i \)-th small area and \( e_i \) is a \( N_i \times 1 \) vector of individual level random errors. The random variables \( u_i \) and \( e_i \) are assumed to be independently distributed, with zero means and with variances \( \text{Var}(u_i) = \Sigma \) and \( \text{Var}(e_i) = \sigma^2_e I_{N_i} \) respectively. Normality of these two random variables is often also assumed. The covariance matrix of \( Y_i \) is \( \text{Var}(Y_i) = V_i = \sigma^2_e I_{N_i} + Z_i \Sigma Z_i' \), which depends on a \( k \times 1 \) vector of fixed parameters \( \theta \), usually called the variance components of the model. Finally, it is usually assumed that sampling is

\(^1\) Author #1 Division of Social Statistics, University of Southampton, Highfield, Southampton - SO17 1BJ, United Kingdom (e-mail:hchandra@soton.ac.uk); Author #2 Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail :ray@uow.edu.au)
uninformative given the values of the auxiliary variables, so the sample data also follow the population model (1).

By grouping the area-specific models (1) over the population, we are led to the population level model

\[ Y = X\beta + Zu + e \]  

where \( Y = (Y_1',\ldots,Y_m')' \), \( X = (X_1',\ldots,X_m')' \), \( Z = \text{diag}(Z_i; 1 \leq i \leq m) \), \( u = (u_1',\ldots,u_m')' \) and \( e = (e_1',\ldots,e_m')' \). The variance-covariance matrix of \( Y \) is \( V = \text{diag}(V_i; 1 \leq i \leq m) \). We assume that \( X \) has full column rank \( p \).

In practice the variance components that define \( V \) are unknown and must be estimated from the sample data. We use a “hat” to denote an estimate and put \( \hat{V} = \hat{\sigma}_e^2 I_N + Z\hat{\Sigma}Z' \). In this paper we shall only consider the restricted maximum likelihood (REML) method of variance components estimation. Further, we consider the sample and non-sample decomposition of \( Y, X, Z, e \) and \( V \) similar to Chambers (2005). We denote by \( X_s \) the \( n \times p \) matrix of sample values of the auxiliary variables, \( Z_s \) the corresponding \( n \times q \) matrix of sample components of \( Z \) and \( V_{ss} \) the \( n \times n \) covariance matrix associated with the \( n \) sample units that make up the \( n \times 1 \) sample vector \( Y_s \). We use a subscript of \( r \) to denote corresponding quantities defined by the \( N - n \) non-sample units, with \( V_{rs} \) denoting the \((N-n) \times n \) matrix defined by \( \text{Cov}(Y_r,Y_s) \). In what follows we denote 1, 1, and \( X_t \) as vectors of 1's and \( I_N, I_n \) and \( I_r \) as identity matrices of order \( N, n \) and \( N-n \) respectively. We use similar notation at the small area level by introducing an extra subscript \( i \) to denote small area. For example, we denote by \( s_i \) the set of \( n_i \) sample units in area \( i \), \( r_i \) the corresponding \( N_i - n_i \) non-sampled units in the area and put

\[ V_{iss} = \sigma_e^2 I_{n_i} + Z_{is}\Sigma Z_{is}' \] and \( V_{iros} = Z_{is}\Sigma Z_{ir} \).

Given this notation, and assuming (1) holds, the EBLUP for the \( i^{th} \) small area mean \( \dot{Y}_i \) is

\[ \hat{Y}_i^{EBLUP} = f_i \bar{Y}_{is} + (1 - f_i)[\bar{X}_i' \hat{\beta} + \bar{Z}_i \hat{\Sigma} Z_{is}' (Y_{is} - X_{is} \hat{\beta})] \]  

where \( f_i = n_i/N_i \) and \( \bar{X}_{ir} \) and \( \bar{Z}_{ir} \) are \( p \times 1 \) and \( q \times 1 \) vectors respectively of means for the \( N_i - n_i \) non-sampled units in small area \( i \). An approximately unbiased estimator of the MSE of (3) is

\[ \text{mse}(\hat{Y}_i^{EBLUP}) = (1 - f_i)^2 \left[ g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2g_{3i}(\hat{\theta}) \right] + \frac{(1 - f_i)}{N_i} \hat{\sigma}_e^2 \]  

where

\[ g_{1i}(\hat{\theta}) = \bar{Z}_i' \left( \hat{\Sigma} - \hat{\Sigma} Z_{is}' Z_{is} \hat{\Sigma} \right) \bar{Z}_{ir} \],

\[ g_{2i}(\hat{\theta}) = (\bar{X}_i' - b_i' X_{is}) \left( \sum_i X_{it}' \hat{\Sigma} Z_{is}' \right) (\bar{X}_i' - b_i' X_{is})' \],

\[ g_{3i}(\hat{\theta}) = \text{tr} \left\{ \nabla b_i' V_{ss} (\nabla b_i) \text{Var}(\hat{\theta}) \right\} \]
with \( b_i^* = \hat{Z}_i^* \hat{Z}_i \hat{Z}_i^{-1} \), \( \nabla b_i^* = \frac{\partial b_i^*}{\partial \theta} \) and where \( \text{Var}(\hat{\theta}) \) is the asymptotic covariance matrix of \( \hat{\theta} \) given by the inverse of the relevant information matrix. See Prasad and Rao (1990) and Rao (2003).

In contrast, under the population level linear mixed model (2), the sample weights that define the EBLUP for the population total of \( Y \) are

\[
\hat{w}_{EBLUP} = 1_n + \hat{H}'(X'1_N - X'_i1_n) + (I_n - \hat{H}X'_i)(\hat{V}_{ss}^{-1} + V_{sr}1_r).
\]

where \( \hat{H} = (X'_i\hat{V}_{ss}^{-1}X_s)^{-1}X'_i\hat{V}_{ss}^{-1} = (\sum_i X'_i\hat{V}_{is}^{-1}X_i)\hat{V}_{ss}^{-1} = (\sum_i X'_i\hat{V}_{is}^{-1}) \). These sample weights are calibrated on \( X \), i.e., \( X'_i\hat{w}_{EBLUP} = X'_i1_N \), and define an unbiased linear predictor of the population total of \( Y \) (Royall, 1976). They depend on the random area effects structure of the mixed model (2) via the covariance structure in the sample/population, with extension to more complex covariance structures only requiring that \( \hat{V}_{is}^{-1} \) and \( \hat{V}_{sr} \) be recomputed under these more complex models.

The MBD estimator of the \( i \)th small area mean is then defined as

\[
\hat{y}_{iMBD} = \sum_{j:s} \hat{w}_{ij} y_j / \sum_{j:s} \hat{w}_{ij},
\]

where the weights in (6) are those associated with the sample units in small area \( i \) in (5). Note that the EBLUP (3) and MBD estimator (6) are not the same at small area level. However, both methods of small area estimation lead to the same overall population estimate since

\[
\hat{N}_i \equiv \sum_{j:s} \hat{w}_{ij} y_j = \sum_{j:s} \hat{N}_i \hat{w}_{iMBD} = \sum_{j:s} \hat{N}_i \hat{w}_{iEBLUP},
\]

where \( \hat{N}_i = \sum_{j:s} \hat{w}_{ij} \) and weights are the EBLUP weights \( \hat{w}_{ij} \).

Estimation of mean squared error (MSE) of (6) follows the approach of Chambers (2005) and treats this expression as a simple weighted domain mean estimate under the usual uncorrelated model for the linear regression of \( Y \) on \( X \) at population level. This is the model that would be used to predict the population total of \( Y \), and also, typically, any domain totals associated with this variable. Under this approach the sample weights derived from (5) are treated as fixed and the prediction variance of (6) is estimated using a standard robust variance estimator. See Royall and Cumberland (1978). A “plug-in” estimate of the squared bias of (6) under this model is added to this estimated prediction variance to finally define a simple estimate of the MSE. Note that under this approach the EBLUP weights underlying (6) “borrow strength” via the assumed small area model (2), but this model is not used in inference. In contrast, standard methods for estimation of the MSE (4) of the traditional EBLUP are based on the linear mixed model (2) and can be quite complicated because of the need to account for the estimated random effects in this model (see, e.g., Prasad and Rao, 1990). The choice between these two approaches is largely philosophical and depends on how much one “believes” (2). In particular, Chambers (2005) treats (2) as a vehicle for generating estimation weights, but bases inference on the model used for population level inference, thus ensuring consistency with the way mean squared errors are estimated at population level. In particular, a robust estimator of the MBD estimator (6) is
\[ \text{mse}(\hat{Y}^{MBD}_i) = v(\hat{Y}^{MBD}_i) + \left( \text{bias}(\hat{Y}^{MBD}_i) \right)^2 \]  

(7)

where \( v(\hat{Y}^{MBD}_i) = \sum_{j=1}^{n_i} \lambda_j (y_j - x'_j \hat{\beta})^2 \) is the estimate of prediction variance of (6) under the usual population level model, with \( \lambda_j = N^{-2}_i \left( a_j^2 + (n_i - 1)^{-1} (N_i - n_i) \right) \),

\[ a_j = \left( \sum_{j=1}^{n_i} w_j \right)^{-1} \left( N_i w_j - \sum_{j=1}^{n_i} w_j \right) \]

and \( \text{bias}(\hat{Y}^{MBD}_i) = (\hat{X}^{MBD}_i - \bar{X}_i)' \hat{\beta} \) is the estimate of bias of (6) under the same model. See Royall and Cumberland (1978) and Chambers (2005). Here \( \hat{X}^{MBD}_i \) denotes the weighted average of the sample values of the auxiliary variables in area \( i \). In the next section we provide empirical results that compare (7) with MSE estimation (4) under the standard approach as set out in Prasad and Rao (1990).

3. An Empirical Study

The basis of the empirical results reported here is the same sample of 1652 Australian farms that were used for the empirical evaluation reported in Chambers (2005). In particular, we use the same target population of 81982 farms (obtained by sampling with replacement from the original sample of 1652 farms with probabilities proportional to their sample weights). The same 1000 independent stratified random samples as used in Chambers (2005) were then drawn from this (fixed) population, with total sample size in each draw equal to the original sample size (1652) and with the small areas of interest defined by the 29 Australian agricultural regions represented in this population. Sample sizes within these regions were fixed to be the same as in the original sample. Note that these varied from a low of 6 to a high of 117, allowing an evaluation of the performance of the different methods considered across a range of realistic small area sample sizes.

Our aim is to estimate average annual farm costs (A$) in these regions. These regions are grouped into three zones (Pastoral, Mixed Farming, and Coastal), with farm size (hectares) assumed known for each farm in the population. Although the linear relationship between the annual farm costs and farm size is rather weak in the original sample data, this improves when separate linear models are fitted within six post strata. These post-strata are defined by splitting each zone into small farms (farm area less than zone median) and large farms (farm area greater than or equal to zone median).

The matrix \( X \) of auxiliary variable values in (1) was defined so as to include an effect for farm size, effects for the post-strata and effects for interactions between farm size and the post strata. Two different specification for \( X \) (corresponding to whether an intercept was included or not) and two different specifications for \( Z \) (corresponding to whether a random slope on farm size was included or not) were then used to specify (1) and hence the EBLUP and MBD estimators based on this model. These four special cases of (1) are set out below.
For the farm data, models I and II are appropriate (with II fitting marginally better) while models III and IV are badly specified. We use REML estimates of random effects parameters throughout, obtained via the *lme* function in R.

Table 1 presents the average percentage relative biases, average percentage root mean squared errors (RRMSE) and average nominal 95% coverage rates (averaged over the 29 small areas), generated by the EBLUP and MBD estimators under the four different models. These results show that the relative biases of the MBD estimator are smaller than the relative biases of the EBLUP estimator for all models except IV. However, the average percent root mean square errors of the MBD estimator are marginally higher than those of the EBLUP estimator for I and II and smaller for III and IV. The coverage rates (which should nominally be around 95 percent) for the MBD estimator are relatively higher than those of the EBLUP estimator for all models. Although neither estimator dominates the other, it seems clear that the MBD approach appears to be more robust than EBLUP approach.

**Table 1** Average (ARB) and median (MRB) values of relative bias (%), average (ARRMSE) and median (MRRMSE) values of relative root mean squared error (%) and average (ACR) coverage rate under model I-IV

<table>
<thead>
<tr>
<th>Model</th>
<th>Approach</th>
<th>ARB</th>
<th>MRB</th>
<th>ARRMSE</th>
<th>MRRMSE</th>
<th>ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>EBLUP</td>
<td>4.24</td>
<td>1.55</td>
<td>19.92</td>
<td>15.74</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-2.49</td>
<td>-0.82</td>
<td>20.56</td>
<td>14.45</td>
<td>0.92</td>
</tr>
<tr>
<td>II</td>
<td>EBLUP</td>
<td>2.98</td>
<td>0.61</td>
<td>19.87</td>
<td>16.40</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-2.13</td>
<td>-0.47</td>
<td>20.15</td>
<td>13.16</td>
<td>0.93</td>
</tr>
<tr>
<td>III</td>
<td>EBLUP</td>
<td>4.52</td>
<td>1.95</td>
<td>23.89</td>
<td>19.94</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-3.84</td>
<td>0.13</td>
<td>21.14</td>
<td>14.44</td>
<td>0.94</td>
</tr>
<tr>
<td>IV</td>
<td>EBLUP</td>
<td>1.17</td>
<td>-2.63</td>
<td>23.38</td>
<td>19.73</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>2.20</td>
<td>2.06</td>
<td>22.35</td>
<td>20.61</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figures 1-3 show the region-specific performances generated by the EBLUP and MBD approaches (ordered by increasing population size). Figure 1 shows the relatively better performances of both the EBLUP and MBD estimators under I and II and their relatively poor performance under IV. In particular, both estimators are badly biased under this model. Figure 2 shows that the regional estimates generated by MBD approach are comparable with those generated under the EBLUP approach in terms of efficiency, with neither approach dominating the other. Overall, with the exception of two regions (3 and 21), it seems that the MBD estimator under II performs marginally better overall.
Figure 1 Region-specific percentage relative biases for EBLUP (dashed line) and MBD (solid line) under model I-IV

Figure 2 Region-specific percentage relative RMSE for EBLUP (dashed line) and MBD (solid line) under model I-IV
In the two regions (3 and 21) where this MBD approach fails, inspection of the population and sample data indicated that this is because of a few outlying estimates. In fact, the outlying estimates for region 21 are all caused by presence of a single massive outlier (farm costs > A$30,000,000) in the original sample. This outlier was included in the simulation population (twice) and then selected (in one case, twice) in 37 of the 1000 simulation samples. If we discard the outlier driven estimates in region 21 then the MBD approach definitely seems the method of choice for regional estimation in our simulation study.

Figure 3 summarizes region-specific variation in the 2-sigma confidence interval coverage rates generated by EBLUP and MBD estimators. If we ignore the outlier driven results for region 21, the results displayed in Figure 3 show that the MBD approach gives marginally better coverage rates under I and II. A close look at these results also indicates that in the event of model misspecification (e.g. under models III and IV) the MBD coverage rate is more robust.

As pointed out above, the MBD estimators are influenced by the presence of outlier data values in some of the regions. Table 1 therefore presents the median percentage relative biases and the median percentage root mean squared errors generated by the EBLUP and MBD approaches. It is now clear that the MBD estimators based on II perform better than all the other estimators. It is also obvious that this performance does not seem to depend on the sample size in the small area.

**Figure 3** Region-specific coverage rate for EBLUP (dashed line) and MBD (solid line) under model I-IV
4. Conclusions and Future Research

Our empirical results indicate that the MBD estimator (6) performs well and represents a real alternative to the EBLUP. Furthermore, the simple MBD MSE estimator (7) appears to provide good coverage performance. However, some caution is required when interpreting these conclusions. The EBLUP predictor (3) approximates the BLUP under (1) and hence will be preferable if this model holds. Our results explore the more realistic situation where (1) is a working model, rather than the (unknown) true model underpinning the data. In this case it appears that the MBD approach is more robust and hence preferable.

It should also be pointed out that the weights underpinning the MBD estimator (6) are essentially variable specific. Furthermore, the data set used in section 3 involved skewed data as well as a potential nonlinear relationship between the survey and auxiliary variables. It is possible to adapt the MBD approach for small area estimation when variables are linear on a transformed scale. Furthermore, although the MBD weights are variable specific and therefore efficient for estimation related to the variable on which they are based, development of general-purpose (i.e. not variable specific) MBD weights has the potential to make the method even more useful. The authors are currently working on these two issues, and results obtained so far are very encouraging.

Acknowledgement

The first author gratefully acknowledges the financial support provided by a PhD scholarship from the U.K. Commonwealth Scholarship Commission.

References


An Empirical Comparison of EBLUP Estimation and Model Based Direct Estimation for Small Areas

Hukum Chandra and Ray Chambers

University of Southampton, United Kingdom and University of Wollongong, Australia
Overview

- Linear mixed models
- Weights based on linear mixed models
- The EBLUP and its MSE estimator
- The MBD estimator
- MSE estimation for the MBD estimator
- Empirical results
- Conclusions
Background

- Small area estimation (SAE) is now very common in survey sampling and several methods for the estimation have been proposed in the literature (Rao, 2003)

- Unit level random effect models are often used in SAE

- The Empirical Best Linear Unbiased Prediction (EBLUP) approach (Prasad and Rao, 1990) is the most popular model based techniques for the SAE under such models

- These approaches typically do not use the unit level survey weights in their estimators
Background

- As result, simplicity of using linearly weighted estimators are lost

- The Model-Based Direct (MBD) approach (Chambers and Chandra, 2006) overcomes these limitations

- Uses calibrated sample weights derived via population level version of linear mixed model in SAE

- MBD Estimator: Defined as linearly weighted estimator
Linear Mixed Model

- The most commonly used class of models in small area inference

\[ Y_i = X_i \beta + Z_i \mu_i + e_i; \quad i = 1, \ldots, m \]

- \( Y_i \) is the \( N_i \times 1 \) vector of values of variable of interest \( Y \) in the small area \( i \)
- \( X_i \) is a \( N_i \times p \) matrix of known auxiliary variables
- \( N_i \) is the number of the population units in the area \( i \)
- \( m \) is the number of small areas
- \( \beta \) is \( p \times 1 \) vector of fixed effects
- $Z_i$ is a $N_i \times q$ matrix of known covariates
- $u_i$ is the random area effect associated with the small area $i$
- $e_i$ is the $N_i \times 1$ vector of random errors
- $u_i$ and $e_i$ are independent random vectors, both with zero mean vectors and with $\text{Var}(u_i) = \Sigma(\theta)$, $\text{Var}(e_i) = \sigma^2 e I_{N_i}$ so that

$$\text{Var}(Y_i) = V_i = \sigma^2 e I_{N_i} + Z_i \Sigma(\theta) Z_i'$$

- Aggregating $m$-small area models, lead to population level linear mixed model (with block diagonal structure)

$$Y = X \beta + Zu + e$$

with $V = \text{Var}(Y) = \text{block diagonal}(V_i)$
Mixed Model Weighting

• Estimate variance components $\theta$ and $\sigma^2_e$ from sample data, leading to $\hat{\theta}$ and $\hat{\sigma}^2_e \Rightarrow \hat{V}_i = \hat{\sigma}^2_e I_i + Z_i \Sigma(\hat{\theta}) Z_i' \Rightarrow \hat{V} = \text{block diagonal}(\hat{V}_i)$

• Use appropriate sample/Non-sample decompositions

• Under the population level mixed model, EBLUP weights (Royall, 1976)

$$w_{EBLUP} = 1_s + H'_{EBLUP} \left( X'1_N - X's1_s \right) + \left( I_s - H'_{EBLUP} X' \right) \hat{V}^{-1}_{ss} \hat{V}_{sr} 1_r$$

$$H_{EBLUP} = \left( X'\hat{V}^{-1}_{ss} X_s \right)^{-1} X_s' \hat{V}^{-1}_{ss}$$

• $1_N$, $1_n$ and $1_r$ are vectors of 1’s of order $N$, $n$ and $(N-n)$ and $I_s$ identity matrix of order $n$
The Industry Standard

**EBLUP** of the area $i$ mean $\bar{Y}_i = N_i^{-1} \sum_{U_i} y_j$

$$\hat{Y}_i^{EBLUP} = f_i \bar{Y}_{is} + (1 - f_i) [\bar{X}'_{ir} \hat{\beta} + \bar{Z}'_{ir} \hat{\Sigma} \bar{Z}_{is} \hat{V}_{iss}^{-1} (Y_{is} - X_{is} \hat{\beta})]$$

**MSE** usually estimated via the Prasad-Rao estimator

$$mse(\hat{Y}_i^{EBLUP}) = (1 - f_i)^2 \left[ g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2 g_{3i}(\hat{\theta}) \right] + N_i^{-1} (1 - f_i) \hat{\sigma}_e^2$$

where $\hat{\theta}$ and $\hat{\sigma}_e^2$ are the estimates of the variance components and $g_{1i}$, $g_{2i}$ and $g_{3i}$ are rather complicated functions.
An Alternative

Model-Based Direct Estimator

\[ \hat{Y}_{i,MBD} = \frac{\sum_{j \in s_i} w_j y_j}{\sum_{j \in s_i} w_j} \]

- \( w_j \) are the EBLUP (mixed model) weights derived earlier

- EBLUP and MBD are not the same at small area level. However, both methods of small area estimation lead to the same overall population estimate

\[ \sum_{i=1}^{m} N_i \hat{Y}_{i,EBLUP} = \sum_{i=1}^{m} \hat{N}_i \hat{Y}_{i,MBD} = \sum_{j \in s} w_j y_j \]

where \( \hat{N}_i = \sum_{s_i} w_j \) and weights are the EBLUP weights \( w_j \)
MSE Estimation for the MBD Estimator

Adapt standard robust methods of MSE estimation for population totals and means

At population level

\[ v(\hat{Y}) = \sum_{j \in s} w_j (y_j - \hat{y}_j)^2 + \text{lower order terms} \]

• An approximately unbiased estimate of \( Var(y_j) \) is given as squared residual \((y_j - x'_j \hat{\beta})\) (Royall and Cumberland, 1978)

• **Estimate of MSE of MBD estimator** for the \( i^{th} \) small area mean (Chambers and Chandra, 2006)

\[
mse(\hat{Y}_{iMBD}) = v(\hat{Y}_{iMBD}) + \left[ \text{bias}(\hat{Y}_{iMBD}) \right]^2
\]
Estimate of Prediction variance

\[ v(\hat{Y}_{MBD}) = \sum_{j \in s_i} \lambda_j (y_j - x_j \hat{\beta})^2 \]

\[ \lambda_j = N_i^{-2} \left( a_j^2 + \frac{(N_i - n_i)}{(n_i - 1)} \right) \text{ and } a_j = \left( N_i w_j - \sum_{s_i} w_j \right) / \left( \sum_{s_i} w_j \right) \]

Estimate of bias

\[ \text{bias}(\hat{Y}_{MBD}) = (\hat{X}_{MBD} - \bar{X}_i)' \hat{\beta} \]

- Bias correction due to using the above estimator on small area level
- \( \hat{X}_{MBD} \) is the weighted average of the sample \( x_j \) in the area \( i \)
- \( \bar{X}_i \) is the population mean of \( x_j \)'s in the small area \( i \)
Empirical Evaluation

- Australian broadacre farms survey data on 1652 sample farms

- Generated a target population of 81982 farms by sampling with replacement from them with probabilities proportional to their sample weights

- 29 different Australian broadacre agricultural regions was considered as 29 small areas of interest

- Sample size within small areas varied from 6 to 117

- Regions are grouped into zones (Pastoral, Mixed Farming, Coastal)

- Six post strata: splitting each zone into small farms (farm area < than zone median) and large farms (farm area > than or equal to zone median)
Average Farm Cost vs Average Farm Area in 6 Post Strata

Variables (Y) : Annual Farm Costs (A$)
Auxiliary (X) : Farm Area
Two X Specifications

- SizeZone*Farm Area (weights constrained to reproduce population and farm area totals in each of six size by zone poststrata)
- Intercept / no intercept

Two Random Effects Specifications

- Random Intercepts specification \( (Z_i \text{ equal to a vector on one’s}) \)
- Random Slopes specification \( (Z_i \text{ equal to the design matrix for a linear regression on Farm Area}) \)
Model Specifications

Model -I (Random Intercept)

Model -II (Random slope)

Model -III (Random slope and Fixed Intercept)

Model -IV (Random slope and Zero Intercept)
Empirical Results

Average (ARB) and median (MRB) values of relative bias, average (ARRMSE) and median (MRRMSE) values of relative RMSE and average (ACR) coverage rates

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>ARB</th>
<th>MRB</th>
<th>ARRMSE</th>
<th>MRRMSE</th>
<th>ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>EBLUP</td>
<td>4.24</td>
<td>1.55</td>
<td>19.92</td>
<td>15.74</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-2.49</td>
<td>-0.82</td>
<td>20.56</td>
<td>14.45</td>
<td>0.92</td>
</tr>
<tr>
<td>II</td>
<td>EBLUP</td>
<td>2.98</td>
<td>0.61</td>
<td>19.87</td>
<td>16.40</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-2.13</td>
<td>-0.47</td>
<td>20.15</td>
<td>13.16</td>
<td>0.93</td>
</tr>
<tr>
<td>III</td>
<td>EBLUP</td>
<td>4.52</td>
<td>1.95</td>
<td>23.89</td>
<td>19.94</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>-3.84</td>
<td>0.13</td>
<td>21.14</td>
<td>14.44</td>
<td>0.94</td>
</tr>
<tr>
<td>IV</td>
<td>EBLUP</td>
<td>1.17</td>
<td>-2.63</td>
<td>23.38</td>
<td>19.73</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>MBD</td>
<td>2.20</td>
<td>2.06</td>
<td>22.35</td>
<td>20.61</td>
<td>0.97</td>
</tr>
</tbody>
</table>

- Average and median over 29 small areas
Region-specific Relative Root Mean Squared Errors

- **EBLUP**
- **MBD**

Model - I

Model - II

Model - III

Model - IV
Conclusions

- In case of model misspecification, the MBD approach appears to provide a more robust set of small area estimates.

- The MBD mean squared error estimator performs well and represents a real alternative to the usual EBLUP estimator.
References


Estimating real income at regional level using a CPI bias correction

Rosa Bernardini Papalia

1. Introduction

The regional Consumer Price Index (CPI) is an issue of great interest to measure the price movements for a specific local area of interest and can be used: (i) to deflate monetary measures of household living standards, and (ii) to update poverty lines, at a regional level. The issue that consumer price index bias may affect the measurement of the trend rate of poverty reduction is relevant also to countries where either the consumer price index or a variant of it, such as a price index of low-income workers, is used to update poverty lines (Deaton and Tarozzi 2000; Deaton, 2003).

In literature, much attention has been devoted to the nature and extent of different sources of bias in consumer price index compilation based on a modified Laspeyres index (Diewert, 1998; Balk 1999), and to the measurement of each particular type of bias. This type of index is known to produce a number of biases, compared to the conceptual standard of a true cost of living index (Hausman, 2003): substitution (or formula) bias; elementary index bias; outlet substitution bias; new goods and quality adjustment or linking bias. In particular, because consumers may substitute away from higher priced goods and outlets, while a Laspeyres index continues measuring the price of the higher priced items (from the original selected outlets), the consumer price index will be an upwardly biased estimate of changes in the true cost of living. In this study we introduce a method for measuring consumer price index bias at regional level that uses Engel’s law (Hamilton, 2001).

The idea is to calculate consumer price index bias by estimating the income elasticity of food using cross-sectional micro data and to employ these estimates to measure the increase in households’ real income over time by controlling for changes in relative prices and in demographic characteristics at regional level. More specifically, given that food’s budget share is inversely related to household real income, by controlling for movements in relative prices and households characteristics, it is possible to infer changes in real incomes from movements in the share of food. This approach gives reduced form estimates of the overall bias in the consumer price index, inferred from movements in food Engel curves over time.

2. Estimating framework for CPI bias measurement at regional level

The Laspeyres consumer price index which finds the cost of purchasing a fixed basket in a base period and the cost of buying the same basket in the present, is

\[ \text{Laspeyres CPI} = \frac{\text{Cost of Basket in Present}}{\text{Cost of Basket in Base Period}} \]

Department of Statistics, University of Bologna, Italy (e-mail: bernardini@stat.unibo.it).
known to produce a number of biases, such as outlet bias, quality change and new products bias.

Following the work of Hamilton (2001) we show how to use different time periods of cross-sectional micro data to identify CPI bias at regional level and to infer CPI bias from a food Engel curve.

The advantage of food as an indicator good is that its low income elasticity makes its budget share sensitive to the measurement of income, whereas goods with income elasticities close to one will have budget shares that are unchanged through time even if income growth is mismeasured. Food is also a non-durable, implying that expenditures in one period cannot provide a flow of consumption in another, and is likely to be separable from other goods in consumers’ utility functions.

We start by introducing the Leser-Working form of the Engel model:

\[ w_{i,j,t} = \phi + \gamma (\ln P_{F,j,t} - \ln P_{NF,j,t}) + \beta (\ln Y_{i,j,t} - \ln P_{j,t}) + X^\prime \theta + u_{i,j,t} \]  \hspace{1cm} (1)

where \( w_{i,j,t} \) is the budget share of food for household \( i \) in region \( j \) and time period \( t \), \( P_{F,j,t} \), \( P_{NF,j,t} \), and \( P_{j,t} \) represent the true but unobserved prices of food, non-food, and all goods, \( Y \) is the household’s total income (which is measured by total expenditure), \( X \) is a vector of individual household characteristics and \( u \) is the residual.

The true cost of living is treated as a geometric weighted average of the prices of food and non-food:

\[ \ln P_{j,t} = \alpha \ln P_{F,j,t} + (1 - \alpha) \ln P_{NF,j,t} \]  \hspace{1cm} (2)

and it is also assumed that all prices of a good \( G \) (either food, non-food, or all goods) are measured with error:

\[ \ln P_{G,j,t} = \ln P_{G,j,0} + \ln(1 + \Pi_{G,j,t}) + \ln(1 + E_{G,t}) \]  \hspace{1cm} (3)

where \( \Pi_{G,j,t} \) represents the cumulative percentage increase in the CPI-measured price of good \( G \) from period 0 to period \( t \) and \( E_{G,t} \) is the period-t percent cumulative measurement error in the cost-of-living index since the base period.

By substituting equation (3) into (2), we obtain:

\[ \ln(1 + E_t) = \alpha \ln(1 + E_{F,t}) + (1 - \alpha) \ln(1 + E_{NF,t}) \]  \hspace{1cm} (4)

Assuming that CPI bias does not vary geographically, substituting equations (2), (3) and (4) into equation (1) gives:

\[ w_{i,j,t} = \phi + \gamma \left[ \ln(1 + \Pi_{F,j,t}) - \ln(1 + \Pi_{NF,j,t}) \right] + \beta \left[ \ln Y_{i,j,t} - \ln(1 + \Pi_{j,t}) \right] + X^\prime \theta + \gamma \left[ \ln(1 + E_{F,t}) - \ln(1 + E_{NF,t}) \right] - \beta \ln(1 + E_t) + \gamma \left[ \ln P_{F,j,t} - \ln P_{NF,j,t} \right] - \beta \ln P_{j,t} + u_{i,j,t} \]  \hspace{1cm} (5)
By using a cross-section/time series database with micro data on income and food expenditure as well as other variables such as family composition which would influence food expenditure, as well as cross-section CPI for all consumption, for food and non-food, over the entire data period for a single region, the following empirical version of equation (5) can be estimated:

$$w_{i,j,t} = \hat{\phi} + \gamma [\ln(1 + \Pi_{F,j,t}) - \ln(1 + \Pi_{NF,j,t})] + \beta [\ln Y_{i,j,t} - \ln(1 + \Pi_{j,t})] + X'\theta + \sum_{i=1}^{r} \delta_i D_i + u_{i,j,t}. \quad (6)$$

where $D_i$ is a dummy variable equal to 1 in period $t$, $\delta_i$ is its coefficient, and $\hat{\phi}$ is the intercept from equation (5), plus the coefficients of the omitted time dummies.

The key role of the time dummy variables to the measurement of CPI bias emerges because:

$$\delta_t = \gamma [\ln(1 + E_{F,t}) - \ln(1 + E_{NF,t})] - \beta \ln(1 + E_t). \quad (7)$$

Under the assumption that the relative bias between food and non-food is constant across time periods and writing the previous expression in terms of the cumulative bias in the CPI for all goods, then we have:

$$\ln(1 + E_t) = \frac{\delta_t}{-\beta - \frac{\gamma(1-r)}{1-\alpha(1-r)}} . \quad (8)$$

It follows that the bias can be identified up to an unknown parameter, $r$, which is the ratio of the CPI bias in food and non-food, and also depends on $\alpha$, which is foods’ share in the cost-of-living index. The equation (8) can be reduced to:

$$\ln(1 + E_t) \approx -\frac{\delta_t}{\beta} . \quad (9)$$

if either $\gamma$ or $(1-r)$ is close to zero. If $r<1$ as seems plausible — food is less badly biased than nonfood - then (9) understates the bias, increasingly as $r$ falls increasingly below unity. Thus, a lower bound for cumulative percentage CPI bias at period $t$ is given by a simple ratio of estimated coefficients from equation (6):

$$1 - \exp\left\{-\frac{\delta_t}{\beta} \right\} . \quad (10)$$

If the assumptions underlying equation (9) are satisfied, the cumulative CPI bias is found by dividing the coefficient on the dummy variable for the round by the income coefficient; the average monthly bias can be found by dividing the difference between cumulative bias estimates by the number of months separating them.

To identify the parameter on food prices, $\gamma$, it is possible to use cross-sectional variations in inflation rates, rather than price levels. Nevertheless, this period by
period variation in an aggregate price index for food relative non-food is perfectly correlated to the time dummy variables, \(D_t\), and an appropriate estimator has to be used. To this end, a Generalized Maximum Entropy (GME) estimation approach (Bernardini Papalia 2004; Golan et al. 1996) is here suggested. The maximum entropy-based estimators are most efficient relative to traditional estimators in particular when data constraints for each observation are included in the maximum entropy-based problem formulations. Second, they are able to produce estimates in models where the number of parameters exceeds the number of data points and in models characterized by a non-scalar identity covariance matrix. Third, prior information can be introduced by adding suitable constraints in the formulation without imposing strong distributional assumptions.

Without geographic variation in the price of food, equation (6) becomes:

\[
\frac{\ln(\hat{t}_{jt}T_{jt})}{\beta} + D_t + u_{i,j,t}. \tag{11}
\]

This model specification can be used when cross-sectional variation in food prices is unavailable. Here, the dummy variables measure not just the CPI bias of equation (7) but also the effect on budget shares of intertemporal variation in the measured inflation rate for food relative to non-food. Hence, the cumulative percentage CPI bias at time \(t\) is calculated from:

\[
1 - \exp\left\{-\frac{\delta_t - \bar{\pi}_t[\ln(1 + \pi_{F,t}) - \ln(1 + \pi_{NF,t})]}{\beta}\right\} \tag{12}
\]

where \(\bar{\pi}\) has to be computed from outside of the estimated parameters for equation (11).

Using regionally disaggregated data for the food and nonfood inflation rates, equations (6) and (9) provide the basic framework, following the approach of Hamilton (2001) of employing food and nonfood inflation rates rather than price levels to identify \(\bar{\pi}\).

3. An empirical application

3.1 Estimation of regional price indexes

The Laspeyres CPI is defined as a fixed-quantity price index that measures the price change in a fixed market basket of consumption goods and services that are purchased by the reference population (Turvey, 2002).

Let \(\hat{e}_g^n\) be an estimator of the expenditure on commodity group \(g\) (\(\hat{e}_g^n = \sum_g e_g^0\)); the Laspeyres CPI at time \(t\) with base year 0 is given by:
\[ I_t^g = \sum_g I_t^g w_0^g = \sum_g I_t^g e_0^g / \bar{e}^0 \]

\[ \sum_g w_0^g = 1, \]

\[ w_0^g = \frac{\sum_{i \in H^0} e_i^0 w_{ig}^0}{\sum_{i \in H^0} e_i^0} \]

\[ e_i^0 = \sum_g e_{ig}^0 \]

where: \( I_t^g \) is the price index of commodity group (or expenditure item) \( g \), \( (g=1,...,G) \), \( e_0^i \) denotes household \( i \) total expenditures \( (i=1,...,H) \), \( w_0^g \) is the budget shares for expenditure category \( g \) in the CPI, and \( w_{ig}^0 \) denotes household \( i \) budget share for expenditure category \( g \). The budget shares for expenditure category \( g \) in the CPI, \( w_0^g \), equals the share of the expenditure on \( g \) in the total base year expenditure on all commodities for a specific group of households.

Expenditure shares for each good are treated as if they were those of an aggregate “super-household” representative of the specific population group. More specifically, the CPIs use weights which reflect the composition of the estimate aggregate values of the reference population. Each household contributes to these weights by an amount proportional to its expenditure. Such weighting is named “plutocratic” in contrast with the “democratic” type of weighting which gives equal importance to all households by averaging consumption value proportions over the whole reference population. The aggregate CPI is computed with the weights which reflect the expenditure of an average household. The resulting Laspeyres price index, known as the plutocratic Laspeyres price index (Prais, 1959), measures the price change of the total base period consumption and it may be interpreted as the price change of the “representative” household’s base period consumption relative to the reference population group.

The regional index is computed by calculating: (i) city indices of product; (ii) regional indices of product, as weighted average of the elementary city indices of product with weights equal to the population in the cities; (iii) regional index, as weighted average of the regional indices of product with weights equal to the expenditures in the region. Weights are obtained from a Family Budget Survey and from other sources in combination with National Accounts data on regional household consumption within the Italian economic territory.

In this view, the regional consumer price index is calculated for the Umbria region by using monthly data over the period January to June 2001 with regard to the 2000 basket and prices, and by assuming the month of December 2000 as the time base. The price indices for each product in four different cities (Perugia, Terri, Città di Castello, Orvieto) are combined in the regional price index for product using weights which represent the relative importance of the expenditure in each city and since expenditure data are not available at the city level, population weights are used (Bernardini Papalia 2004). Then, the price indices for all products are combined to obtain the regional index using weights which represent the relative importance of the expenditure for each product in the region. The data base used to estimate the expenditure shares for subgroups of the reference population is derived from the Italian Expenditures Survey and contains detailed information about the expenditure, together with a great number of household characteristics of a sample of households in 2001.
The results of the computation of the Consumer Price Index for the Umbria region are presented in Table 1. From January to June 2001 the CPI for the Umbria region rose by 1.9%; it rose at a slower pace than the Italian CPI and it was above the Italian CPI for the entire period; the Italian CPI seems to have escalated at a faster rate than the RCPI for all the main consumption groups. Over the entire period, the greatest price increases at the main component level were in Food and Beverages, Alcohol and Tobacco, Transportation, Restaurant and Hotels, while the greatest declines were for Housing and Communication.

Tab. 1 - Consumer Price Index of the Umbria region (December 2000=100, January - June 2001)

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>Jan-01</th>
<th>Feb-01</th>
<th>Mar-01</th>
<th>Apr-01</th>
<th>May-01</th>
<th>Jun-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food and Beverages</td>
<td>101.1</td>
<td>101.6</td>
<td>102.1</td>
<td>102.6</td>
<td>103.3</td>
<td>103.7</td>
</tr>
<tr>
<td>Alcohol and Tobacco</td>
<td>100.1</td>
<td>100.1</td>
<td>100.3</td>
<td>103.2</td>
<td>103.2</td>
<td>103.3</td>
</tr>
<tr>
<td>Clothing and Footwear</td>
<td>100.0</td>
<td>100.0</td>
<td>100.3</td>
<td>100.7</td>
<td>100.9</td>
<td>100.9</td>
</tr>
<tr>
<td>Housing</td>
<td>100.1</td>
<td>99.7</td>
<td>99.9</td>
<td>100.1</td>
<td>98.8</td>
<td>98.7</td>
</tr>
<tr>
<td>Household furnishings</td>
<td>100.1</td>
<td>100.6</td>
<td>100.7</td>
<td>100.8</td>
<td>101.8</td>
<td>101.8</td>
</tr>
<tr>
<td>Medical care</td>
<td>100.6</td>
<td>100.6</td>
<td>100.6</td>
<td>100.7</td>
<td>100.7</td>
<td>100.7</td>
</tr>
<tr>
<td>Transportation</td>
<td>99.2</td>
<td>99.4</td>
<td>99.5</td>
<td>99.9</td>
<td>101.0</td>
<td>101.2</td>
</tr>
<tr>
<td>Communication</td>
<td>99.5</td>
<td>99.2</td>
<td>99.1</td>
<td>99.9</td>
<td>98.7</td>
<td>98.7</td>
</tr>
<tr>
<td>Recreation</td>
<td>102.0</td>
<td>102.0</td>
<td>102.0</td>
<td>102.1</td>
<td>102.3</td>
<td>102.2</td>
</tr>
<tr>
<td>Education</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Restaurants and Hotels</td>
<td>102.7</td>
<td>103.2</td>
<td>103.5</td>
<td>105.1</td>
<td>105.2</td>
<td>105.4</td>
</tr>
<tr>
<td>Other goods and services</td>
<td>101.1</td>
<td>101.2</td>
<td>101.4</td>
<td>102.0</td>
<td>102.1</td>
<td>102.1</td>
</tr>
<tr>
<td>All items</td>
<td>100.6</td>
<td>100.8</td>
<td>101.0</td>
<td>101.5</td>
<td>101.8</td>
<td>101.9</td>
</tr>
</tbody>
</table>

The consumption expenditure of some special groups of the reference population: (i) low-income household (FamL), (ii) high-income household (FamH), (iii) one-person type family (Single), and (iv) family without children (No-kids), is quite different from the average expenditure of the general population (see Tab. 2).

Tab. 2 - Distribution of total average expenditures for the major components of the Consumer Price Index (Umbria region. Relative share in percentage. Household Survey 2000)

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>W Umbria</th>
<th>W H-F</th>
<th>W L-F</th>
<th>W No-kids</th>
<th>W Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food and Beverages</td>
<td>18.0</td>
<td>16.1</td>
<td>23.1</td>
<td>18.4</td>
<td>17.4</td>
</tr>
<tr>
<td>Alcohol and Tobacco</td>
<td>1.6</td>
<td>1.5</td>
<td>1.9</td>
<td>1.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Clothing and Footwear</td>
<td>7.6</td>
<td>8.3</td>
<td>5.6</td>
<td>7.5</td>
<td>6.7</td>
</tr>
<tr>
<td>Housing</td>
<td>27.8</td>
<td>24.2</td>
<td>38.2</td>
<td>29.5</td>
<td>34.9</td>
</tr>
<tr>
<td>Household furnishings</td>
<td>7.8</td>
<td>9.5</td>
<td>3.0</td>
<td>7.0</td>
<td>4.5</td>
</tr>
<tr>
<td>Medical care</td>
<td>3.5</td>
<td>3.4</td>
<td>3.7</td>
<td>3.9</td>
<td>4.8</td>
</tr>
<tr>
<td>Transportation</td>
<td>14.6</td>
<td>15.8</td>
<td>11.1</td>
<td>13.8</td>
<td>11.2</td>
</tr>
<tr>
<td>Communication</td>
<td>2.6</td>
<td>2.4</td>
<td>3.1</td>
<td>2.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Recreation</td>
<td>5.3</td>
<td>6.0</td>
<td>3.4</td>
<td>5.2</td>
<td>5.5</td>
</tr>
<tr>
<td>Education</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Restaurants and Hotels</td>
<td>2.4</td>
<td>2.7</td>
<td>1.7</td>
<td>2.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Other goods and services</td>
<td>8.6</td>
<td>9.7</td>
<td>5.2</td>
<td>8.0</td>
<td>6.8</td>
</tr>
<tr>
<td>All items</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

More specifically, an examination of Table 2 shows that, at the level of Housing, Communication, and Medical care, the total average expenditure relative to some sub-groups of the reference population is higher than that of the reference population.
for low-income households, no-kids and single type families. In the case of high-income households, the relative share of total expenditures for the Clothing, Furnishings and Transportation categories are 8.3, 9.5 and 15.8 percent as compared to 7.6, 7.8 and 14.6 percent for the reference population and were 9.3%, 22% and 8.2% higher. Low-income households’ average expenditures on Food, Housing, and Medical care categories were 23.1, 38.2 and 3.7 percent as compared to 18, 27.8 and 3.49 percent for the reference population and were 28.5%, 37.2% and 6.0% higher.

3.2. Estimation of CPI bias at regional level

The CPI bias is estimated from the food regression using a sample of household of the Umbria region according to the equation (6). A Generalized maximum entropy estimator is used to deal with the collinearity problems (Bernardini Papalia, 2004).

The dependent variable used is the share of expenditures devoted to all food; control variables include the real total expenditures, relative price changes, demographic characteristics, time dummies. Using a more limited set of demographic controls does not affect the bias estimates, so we include only the household size neglecting control variables relative to the work status of the head household, number of children, etc. Augmenting the model with a quadratic expenditure term does not change the dummy variable coefficients showing the downward drift in the food Engel curve, and in fact the quadratic terms are not statistically significant in the estimation results.

The regression results yield reasonable estimates of expenditures and price elasticities; control variables have the expected sign. Food and non-food inflation rates at regional level are computed according to the methodology we presented in section 3.1. Table 3 summarizes cumulative bias estimates and presents estimates corrected for relative price changes.

Tab. 3 – Estimates of cumulative CPI bias in Umbria region, (January - June 2001)

<table>
<thead>
<tr>
<th></th>
<th>Bias – estimates adjusted for relative price changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-01</td>
<td>0.809</td>
</tr>
<tr>
<td>Feb-01</td>
<td>0.923</td>
</tr>
<tr>
<td>Mar-01</td>
<td>0.000</td>
</tr>
<tr>
<td>Apr-01</td>
<td>0.891</td>
</tr>
<tr>
<td>May-01</td>
<td>0.838</td>
</tr>
<tr>
<td>June-01</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Some differences emerged when the mean family income is deflated using the arithmetic mean over six months of the regional CPI index with and without the bias adjustment. For example, deflating income with the regional CPI yields a real income of 44381. When the bias-adjusted regional CPI index is used to deflate mean family expenditure yields a real income of 44636 euros. The result is a -254 difference in purchasing power, or approximately a -0.6 percent difference. This evidence emerged in analyzing all specific sub-groups of the reference population.
4. Conclusions

In this paper, following the work of Hamilton (2001) the idea is to compute a consumer price index bias by estimating the income elasticity of food using cross-sectional micro data at regional level. This approach gives reduced form estimates of the overall bias in the regional consumer price index, inferred from movements in food Engel curves over time.

We have estimated Engel functions for the food budget share of a sample of households living in a specific region of Italy, based on data from January to June 2001 from the Italian Family Budget Survey.

We find an average CPI bias relative to the Umbria region of about two percentage points per month during the period we analyze. The cumulative effect of this bias causes an understatement of the growth performance of this region. We find that, the level of real per capita GDP in 2001 may be undersated by up to 0.6% compared with using a bias-corrected deflator.

References


The Use of Small Area Methodology in Micro-enterprise Revenue Estimation

Tomasz Piasecki

1

1. Introduction

The paper describes results of a research carried out at the Centre for Statistical Surveys Organization in Statistical Office in Lodz. The presented analysis and calculations concern an application of selected small area methods in micro-enterprise revenue estimation in two-dimensional structure by territorial units (voivodships) and kinds of activity.

Calculations were made on the basis of data from the Polish small enterprise survey in Poland (called SP-3 survey). The SP-3 is a sample survey of small enterprises conducted systematically once a year by Polish Official Statistics. Its results are disseminated in regular publications. The paper presents attempts to obtain estimates in more detailed desegregation than they are currently published. In this level of desegregation samples in domains are too small to use a direct estimator – its precision is not sufficient. To improve the estimators’ precision small area methods and data from tax system as auxiliary information are used.

Two methods of estimation are considered: an EBLUP model estimator and a synthetic biproportional estimator that takes into account a two-dimensional structure of population division by domains.

Calculations presented here are not included in the official publications. They should be treated only as an illustration of methods under consideration. These analyses are supposed to contribute to the recognition of possibilities of their implementation into the survey’s practice.

The calculations were made with the use of statistical software R.

2. Problem description

In Poland, the enterprise survey is divided into two parts. Big enterprises (over 9 employed persons) are subject to a full-scale survey (called SP survey). Small enterprises (up to 9 employed persons) are surveyed by means of sampling methods (it is SP-3 survey).

The SP-3 is a sample survey with a stratified sampling. The frame consists of about 2.8 million units. A sample size is about 3% of population. The percentage differs in individual strata.

Results are published in one-dimensional structure by:

- territorial units (16 voivodships),
- kinds of activity (14 NACE sections).

1Statistical Office in Lodz, 93-176 Lodz, ul. Suwalska 29, Poland (t.piasecki@stat.gov.pl)
Data in two-dimensional structure by territorial units and kinds of activity cannot be published. The sample size in most cells of the division defined in such a way is too small to obtain estimates of a suitable precision. The application of methods that can improve the precision is the subject to analysis.

**Enterprise net revenue recorded in SP-3 Survey** is a variable of interest. Variable mean in cells of two-dimensional classification is estimated by using small area methods. As an auxiliary variable **enterprise net revenue from tax return** is used. **Small domains** are defined as cells of a two-dimensional classification by:
- territorial location (16 voivodships) and
- kind of activity in two variants:
  - 14 NACE sections (Level 1),
  - 77 detailed units called NZ (Level 2).

There are two variants of analysis, with different levels of details as regards the classification by kinds of activity. The more detailed level uses a peculiar classification unit called NZ. There are 77 NZ-s in the whole classification of activities.

In the first variant (less detailed level of aggregation – level 1) there are 223 domains, in the second (level 2) – 1230.

Information from two data sets is used in the analysis. The first data set is a source of information on the variable of interest, the second one – on the auxiliary variable. The set of data collected in the SP-3 survey includes about 45465 records, reflecting the examined enterprises. The taxation dataset contains data from the taxation system and covers information from tax returns.

An application of the following estimators is compared:
- Design based direct estimator ($\hat{y}_{i}^{direct}$)
- Synthetic regression estimator ($\hat{y}_{i}^{reg}$)
- Model based EBLUP estimator ($\hat{y}_{i}^{EBLUP}$)
- Synthetic biproportional estimator ($\hat{y}_{i}^{biprop}$)

At the more detailed level of aggregation (level 2) biproportional estimator was not used. Its application makes sense when the precision of one-dimensional marginal distributions is satisfactory. The one-dimensional distribution by NZ does not meet this condition.

### 3. Methods description

#### 3.1. Model based estimation

In case of a model based estimation, the Fay-Herriot model on area level was applied. It assumes for each domain:

$$\hat{y}_{i}^{direct} = \alpha + \beta \bar{x}_{i} + u_{i} + \varepsilon_{i}, \quad (2.1)$$

where:
- $i$ – area (domain) identifier,
- $\bar{x}_{i}$ – true auxiliary variable mean,
- $u_{i}$ – area specific effect, $D^{2}(u_{i}) = s^{2}$,
\[ \varepsilon_i \] — sampling error of a direct estimator, \( D^2(\varepsilon_i) = \sigma^2_i \).

The first variance component \( \sigma^2_i \), associated with a sampling error, is a variance of direct area mean estimator. In this case, a design based estimation for each area was used.

The second component \( s^2 \), associated with an area specific effect, was estimated by using a simple moment estimator (Rao, 2004):

\[
\hat{s}^2 = \frac{1}{m-2} \left[ \sum_{i=1}^{k} (\tilde{y}_i - \bar{y}_i^{\text{direct}})^2 - \sum_{i=1}^{k} \hat{\sigma}^2_i \left( 1 - [1 \; x_i] (X^T X)^{-1} [1 \; x_i]^T \right) \right]
\]  
(2.2)

where:

- \( \tilde{y}_i \) — fitted value from model (2.1) estimated by OLS (without reflecting of error term heteroscedasticity)
- \( X \) — model matrix with explanatory variables values ([1 \; x_i] are its rows)
- \( m \) — number of domains

A regression estimator is obtained with formula:

\[
\bar{y}_i^{\text{reg}} = \hat{\alpha} + \hat{\beta} x_i
\]  
(2.3)

where \( \hat{\alpha} \) and \( \hat{\beta} \) are estimated by GLS (with reflecting of estimated error term variance \( S^2 + \sigma^2_i \)).

The regression estimator allows for the diversity of differences between total variances of error terms in particular domains, but it doesn’t take the structure of variance components into consideration.

The EBLUP estimator can be treated as a weighted mean of the direct and regression estimator. It is given by a formula:

\[
\bar{y}_i^{\text{EBLUP}} = \gamma \bar{y}_i^{\text{direct}} + (1 - \gamma) \bar{y}_i^{\text{reg}}
\]  
(2.4)

where \( \gamma = \frac{\hat{s}^2}{\hat{s}^2 + \hat{\sigma}^2_i} \).

The \( \gamma \) parameter shows the contribution of area effect component to the total variance as well as the share of the direct estimator in the EBLUP estimator.

### 3.2. Biproportional estimation

Synthetic biproportional estimation makes it possible to preserve the marginal distributions of totals identical as in the case of direct estimation.

Biproportional estimators of domain totals are obtained in an iterative process:

\[
y_{ij}^{\text{biprop}(0,2)} = x_{ij}
\]  
(2.5)

\[
y_{ij}^{\text{biprop}(n,1)} = y_{ij}^{\text{biprop}(n-1,2)} \frac{y_{i\cdot}}{\sum_k y_{ik}^{\text{biprop}(n-1,2)}} \quad \text{for each } i, j
\]  
(2.6)

\[
y_{ij}^{\text{biprop}(n,2)} = y_{ij}^{\text{biprop}(n,1)} \frac{y_{\cdot j}}{\sum_k y_{kj}^{\text{biprop}(n,1)}} \quad \text{for each } i, j
\]  
(2.7)
where:
- $i, j$ – row (voivodship) and column (kind of activity) identifiers,
- $x_{ij}$ – domain total of an auxiliary variable,
- $y_{i*}, y_{*j}$ – direct row and column totals,
- $n$ – iteration number.

The biproportional mean estimator is obtained as $y_{ij}^{biprop} = y_{ij}^{biprop} / N_{ij}$, where $N_{ij}$ – domain population size.

The biproportional estimator can be treated as generalization of a ratio estimator.

The ratio estimator applied in rows can be expressed as a product of a total value in the domain and a certain constant:

$$y_{ij}^{ratio} = r_i x_{ij}$$ (2.8)

where $r_i$ is constant for row and $\sum_j r_j x_{ij} = y_{i*}$. 

The biproportional estimator can be seen as a product of a total value for the domain and certain coefficients, constant for each row and column:

$$y_{ij}^{biprop} = r_i x_{ij} s_j$$ (2.9)

where $r_i$ are constant for rows, $s_j$ are constant for columns and for each column and each row:

$$\sum_j r_j x_{ij} s_j = y_{i*} \quad \text{for each } i$$ (2.10)

$$\sum_i r_i x_{ij} s_j = y_{*j} \quad \text{for each } j$$ (2.11)

The estimator was designed by using a biproportional correction method, applied among others in econometrics while correcting forecasted matrices of Input-Output coefficients and called the RAS method. The method used is also related to the SPREE method (structure preserving estimation).

### 3.3. MSE assessing

The mean squared error (MSE) of considered estimators is assessed by a simulation with use of the bootstrap method at an individual level.

Bootstrap samples are generated at individual level with reflection of a sample structure (stratification). Observations are generated from distribution $F_i$, where $i$ is the strata identifier. For $F_i$ an empirical distribution in the strata is assumed.

For each bootstrap sample all estimator values are calculated. Assessments of a root mean squared error (RMSE) for particular estimators are obtained by formula:

$$RMSE(\bar{y}_i^{est}) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\bar{y}_{ib}^{est} - \bar{y}_i)}$$ (2.12)

where:
- $i$ – domain identifier,
- $b$ – bootstrap sample identifier,
There were 1000 replications in this simulation (B = 1000).

4. Results

Here are presented results of the calculations. For a model based estimation there is some information about model fitting and variance decomposition. For all estimators obtained estimates and MSE assessments are compared. Table 4.1 shows results of model fitting by GLS method at both levels of aggregation. The relationship between variables is significant. Model fitting is better in case of more detailed classification.

Table 4.1. Results of GLS model fitting.

<table>
<thead>
<tr>
<th>Level 1 (NACE section)</th>
<th>Coefficients</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.067e+05</td>
<td>7.717e+03</td>
<td>13.826</td>
<td>&lt;2e-16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.435e-01</td>
<td>5.119e-02</td>
<td>6.709</td>
<td>1.62e-10</td>
<td></td>
</tr>
</tbody>
</table>

Residual standard error: 0.9368 on 221 degrees of freedom

Multiple R-Squared: 0.1692
Adjusted R-squared: 0.1655

F-statistic: 45.02 on 1 and 221 DF, p-value: 1.619e-10

<table>
<thead>
<tr>
<th>Level 2 (NZ)</th>
<th>Coefficients</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.370e+05</td>
<td>7.424e+03</td>
<td>18.45</td>
<td>&lt;2e-16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.452e-02</td>
<td>2.546e-02</td>
<td>25.34</td>
<td>&lt;2e-16</td>
<td></td>
</tr>
</tbody>
</table>

Residual standard error: 0.7637 on 1228 degrees of freedom

Multiple R-Squared: 0.3434
Adjusted R-squared: 0.3429

F-statistic: 642.3 on 1 and 1228 DF, p-value:<2.2e-16

The results of variance decomposition are presented by values of γ parameter, which represents relation between two variance components. Figure 4.1 shows histograms of parameter value among domains at both levels. Figure 4.2 presents the two-dimensional distribution of γ parameter at more detailed level of aggregation (level 2). In rows there are kinds of activity, in columns – voivodships. Dark fields represent values close to zero (synthetic estimator), the bright ones – values close to one (direct estimator). The bright fields prevail. However, the horizontal, darker strips can be observed. They exemplify kinds of activity with lower share of the direct estimator in the EBLUP estimator. For these kinds of activity the precision of the direct estimator is distinctly worse.
Fig. 4.1. Histogram of $\gamma$ coefficient values

- **Level 1**
- **Level 2**

<table>
<thead>
<tr>
<th>Number of Domains</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Fig. 4.2. Values of $\gamma$ parameter by voivodship and kind of activity
Table 4.2 contains an aggregated comparison of individual estimator values. It presents median of \( \frac{\hat{y}_i^A - \hat{y}_i^B}{\hat{y}_i^A} \), where \( A \) denotes an estimator placed in a row and \( B \) estimator placed in a column.

Tab. 2. Comparison of mean estimates – median of absolute percentage difference between estimates.

<table>
<thead>
<tr>
<th></th>
<th>Level 1 (NACE section)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct estimator</td>
<td>Regression estimator</td>
<td>EBLUP estimator</td>
<td>Biproportional estimator</td>
</tr>
<tr>
<td>Direct estimator</td>
<td>X</td>
<td>43.23</td>
<td>2.96</td>
<td>21.69</td>
</tr>
<tr>
<td>Regression estimator</td>
<td>47.49</td>
<td>X</td>
<td>32.97</td>
<td>43.32</td>
</tr>
<tr>
<td>EBLUP estimator</td>
<td>2.95</td>
<td>30.32</td>
<td>X</td>
<td>19.27</td>
</tr>
<tr>
<td>Biproportional estimator</td>
<td>20.95</td>
<td>41.60</td>
<td>18.29</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Level 2 (NZ)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct estimator</td>
<td>Regression estimator</td>
<td>EBLUP estimator</td>
<td></td>
</tr>
<tr>
<td>Direct estimator</td>
<td>X</td>
<td>51.71</td>
<td>1.39</td>
<td></td>
</tr>
<tr>
<td>Regression estimator</td>
<td>50.00</td>
<td>X</td>
<td>39.69</td>
<td></td>
</tr>
<tr>
<td>EBLUP estimator</td>
<td>1.37</td>
<td>37.60</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 presents summary comparison of RMSE assessments for individual estimators. A relative RMSE (coefficient of variation) is defined as \( \nu(est)_i = \frac{RMSE(\bar{y}^est)_i}{\bar{y}} \), with notations as in the formula (2.12). For all estimators the characteristics of relative RMSE distribution among domains and percentage of domains in which RMSE of particular estimator is smaller then RMSE of the direct estimator are shown.

Tab. 3. Relative RMSE – summary of values for domains

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimator</td>
<td>Direct estimator</td>
<td>Regression estimator</td>
<td>EBLUP estimator</td>
</tr>
<tr>
<td>min</td>
<td>0.055</td>
<td>0.064</td>
<td>0.072</td>
<td>0.051</td>
</tr>
<tr>
<td>max</td>
<td>1.665</td>
<td>7.807</td>
<td>1.291</td>
<td>13.677</td>
</tr>
<tr>
<td>mean</td>
<td>0.267</td>
<td>0.799</td>
<td>0.259</td>
<td>0.564</td>
</tr>
<tr>
<td>median</td>
<td>0.218</td>
<td>0.477</td>
<td>0.215</td>
<td>0.258</td>
</tr>
<tr>
<td>( \text{RMSE &lt; RMSE of direct estimator} ) (percentage of domains)</td>
<td>25.6</td>
<td>43.9</td>
<td>41.7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Level 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimator</td>
<td>Direct estimator</td>
<td>Regression estimator</td>
<td>EBLUP estimator</td>
</tr>
<tr>
<td>min</td>
<td>0.079</td>
<td>0.036</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>5.115</td>
<td>212.980</td>
<td>2.832</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.327</td>
<td>1.686</td>
<td>0.309</td>
<td></td>
</tr>
<tr>
<td>median</td>
<td>0.285</td>
<td>0.603</td>
<td>0.277</td>
<td></td>
</tr>
<tr>
<td>( \text{RMSE &lt; RMSE of direct estimator} ) (percentage of domains)</td>
<td>25.6</td>
<td>60.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Conclusions

- Mean estimates for domains obtained with the use of the synthetic estimators differ considerably from the direct estimates. The results show that synthetic estimators seem to be biased.
- The EBLUP estimator gives the best results in terms of the MSE. It improves the results obtained by direct estimation, despite the weak relation between the variable of interest and the auxiliary one. However, the improvement is not considerable.
- Improvement resulting from the EBLUP estimation concerns mainly these domains for which the quality of the direct estimation is the worst.
- For most domains γ parameter approaches 1. The EBLUP estimator is close to the direct one what reflects a bias of the synthetic estimator.
- Regularities in distribution of γ parameter by domains concern rather kinds of activity than voivodships and are they are better noticeable at the lower level of aggregation.
- The lower level of aggregation allows better modeling of relation between the variable of interest and the auxiliary one.
- Comparing the synthetic estimators, the biproportional estimator seems to be better than the regression one, in respect of the MSE. Taking into consideration the two-dimensional structure results in considerable improvement of the precision.
- The applied method of the MSE assessing is biased to favour of the direct estimator by assuming direct estimates as true values. The MSE of the indirect estimators may be slightly overestimated.

References

Rao J. N. K. (2003), Small Area Estimation, Wiley
R Development Core Team (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
The use of small area methodology in micro-enterprise revenue estimation

Tomasz Piasecki
Statistical Office in Lodz
Contents

- SP-3 Survey
- Datasets used
- EBLUP estimator
- Biproportional estimator
- Presentation of results
- Conclusions
SP-3 Survey

- Survey of small enterprises
  (up to 9 employed persons)

- Sample survey
  - Stratified sampling
  - Population size: 2,8 M
  - Sample size: about 3% of population
    (different percentage in particular strata)
  - Successfully examined: about 50% of sample
SP-3 Survey

Publication of results:
- by territorial units (voivodships)
- by kinds of activity (NACE section)

Problem:
To estimate revenue in two-dimensional space:

territorial unit x kind of activity
Problem

- **Variable of interest**: 
  *enterprise net revenue recorded in SP-3 Survey*

- **Auxiliary variable**: 
  *enterprise net revenue from tax return*

- **Small area definition**: 
  Levels of two-dimensional classification by:
  - voivodship
  - kind of activity in two variants:
    - 14 units (*NACE section*)
    - 77 units (called *nz*)
Datasets

- Two datasets:
  - Data collected in SP-3 Survey
    - 45,465 records
    - contains variable of interest and information about grouping/stratification

- Taxation dataset
  Data from taxation system, that covers information from tax returns
  - contains information about auxiliary variable
Estimators

- Design based direct estimator (\( \overline{y}_{i}^{\text{direct}} \))
- Synthetic regression estimator (\( \overline{y}_{i}^{\text{reg}} \))
- Model based EBLUP estimator (\( \overline{y}_{i}^{\text{EBLUP}} \))
- Synthetic biproportional estimator (\( \overline{y}_{i}^{\text{biprop}} \))

Two levels of aggregation:

- **Level 1**: voivodship x section (223 domains)
  Estimators used: \( \overline{y}_{i}^{\text{direct}}, \overline{y}_{i}^{\text{reg}}, \overline{y}_{i}^{\text{EBLUP}}, \overline{y}_{i}^{\text{biprop}} \),

- **Level 2**: voivodship x nz (1230 domains)
  Estimators used: \( \overline{y}_{i}^{\text{direct}}, \overline{y}_{i}^{\text{reg}}, \overline{y}_{i}^{\text{EBLUP}} \)
Model based estimation

- Fay-Herriot model on area level

\[
\bar{y}_i^{\text{direct}} = \alpha + \beta \bar{x}_i + u_i + \epsilon_i
\]

\[
D^2(u_i) = s^2 \quad D^2(\epsilon_i) = \sigma_i^2
\]

- \(i\) – area (domain) identifier
- \(\bar{x}_i\) – area mean of auxiliary variable
  (true value in population)
Model based estimation

- $\sigma_i^2$ estimation - design based for each area (variance of direct area mean estimator)
- $s^2$ estimation - simple moment estimator
- Regression estimator
  \[ \bar{y}_i^{\text{reg}} = \hat{\alpha} + \hat{\beta} x_i \]
  $\hat{\alpha}, \hat{\beta}$ estimated by GLS with reflection of $(u_i + \varepsilon_i)$ variance
- EBLUP estimator
  \[ \bar{y}_i^{\text{EBLUP}} = \gamma \bar{y}_i^{\text{direct}} + (1 - \gamma) \bar{y}_i^{\text{reg}} \]
  \[ \gamma = \frac{s^2}{\hat{s}^2 + \hat{\delta}_i^2} \]
Biproportional estimator

- $y_{ij}^{biprop}$ is the result of iterative process:
  (iterative proportional correction of domain totals in rows and columns)

\[
\begin{align*}
  y_{ij}^{biprop(0,2)} &= x_{ij} \\
  y_{ij}^{biprop(n,1)} &= y_{ij}^{biprop(n-1,2)} \frac{y_{i\bullet}}{\sum_k y_{ik}^{biprop(n-1,2)}} \quad \text{for each } i, j \\
  y_{ij}^{biprop(n,2)} &= y_{ij}^{biprop(n-1,1)} \frac{y_{j\bullet}}{\sum_k y_{kj}^{biprop(n-1,1)}} \quad \text{for each } i, j
\end{align*}
\]

- $i,j$ - row (vivodship) and column (kind of activity) identifiers
- $x_{ij}$ - domain total of auxiliary variable
- - direct row and column totals
- $n$ - number of iteration

- Biproportional mean estimator is obtained as:

\[
\bar{y}_{ij}^{biprop} = y_{ij}^{biprop} / N_{ij} , \quad \text{where } N_{ij} - \text{domain population size}
\]
**Biproportional estimator**  
as generalization of synthetic ratio estimator

- **For ratio estimator applied in rows:**
  \[ y_{ij}^{ratio} = r_i x_{ij} \]
  where \( r_i \) is constant for row and
  \[ \sum_j r_i x_{ij} = y_i. \]

- **For biproportional estimator**
  \[ y_{ij}^{biprop} = r_i x_{ij} s_j \]
  where \( r_i \) are constant for rows, \( s_j \) are constant for columns and for each column and each row:
  \[ \sum_j r_i x_{ij} s_j = y_i \text{ for each } i \]
  \[ \sum_i r_i x_{ij} s_j = y_j \text{ for each } j. \]
Results

- Values of estimates
- Structure of variance components for model based estimation ($\gamma$ coefficient values)
- Mean squared error obtained with bootstrap method
  - RMSE (root mean squared error of estimator)
  - Coefficient of variance based on RMSE (relative RMSE)
    \[
    v(\bullet)_i = \frac{RMSE(\bar{y}_i^*)}{\bar{y}_i}
    \]
    where $\bar{y}_i$ - true mean of domain $i$ in bootstrap model
  - RMSE of particular estimator in relation to the direct one
    \[
    \frac{v(\bullet)}{v(direct)}_i = \frac{RMSE(\bar{y}_i^*)}{RMSE(\bar{y}_i^{direct})}
    \]
The results of GLS model fitting

**Level 1** *(section)*

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.067e+05</td>
<td>7.717e+03</td>
<td>13.826</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>x</td>
<td>3.435e-01</td>
<td>5.119e-02</td>
<td>6.709</td>
<td>1.62e-10</td>
</tr>
</tbody>
</table>

*Residual standard error: 0.9368 on 221 degrees of freedom*

*Multiple R-Squared: 0.1692  Adjusted R-squared: 0.1655*

*F-statistic: 45.02 on 1 and 221 DF, p-value: 1.619e-10*
EBLUP estimation

Histogram of $\gamma$ coefficient values

Level 1 (section)
EBLUP estimation
Means of $\gamma$ values
Level 1 (section)

Grouped by voivodship

Grouped by section

---

Means of $\gamma$ values for different sections and voivodships.
The results of GLS model fitting

**Level 2** ($n_z$)

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p-value(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>$1.370e+05$</td>
<td>$7.424e+03$</td>
<td>18.45</td>
<td>$&lt;2e-16$</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>$6.452e-02$</td>
<td>$2.546e-02$</td>
<td>25.34</td>
<td>$&lt;2e-16$</td>
</tr>
</tbody>
</table>

**Residual standard error**: 0.7637 on 1228 degrees of freedom

**Multiple R-Squared**: 0.3434  
**Adjusted R-squared**: 0.3429

**F-statistic**: 642.3 on 1 and 1228 DF, p-value: $<2.2e-16$
EBLUP estimation

Histogram of $\gamma$ coefficient values

Level 2 ($nz$)
EBLUP estimation
Means of $\gamma$ values
Level 2 ($nz$)

Grouped by voivodship

Grouped by $nz$
EBLUP estimation
\( \gamma \) values by voivodship and kind of activity \((nz)\)
Level 2 \((nz)\)
Comparison of mean estimates

**Level 1 (sections)**

Median of \( \frac{\bar{y}_i^A - \bar{y}_i^B}{\bar{y}_i^A} \) for domains (in %)

<table>
<thead>
<tr>
<th>( \frac{\bar{y}_i^A}{\bar{y}_i^B} )</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
<th>Bipropotional estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>X</td>
<td>43.23</td>
<td>2.96</td>
<td>21.69</td>
</tr>
<tr>
<td>Regression</td>
<td>47.49</td>
<td>X</td>
<td>32.97</td>
<td>43.32</td>
</tr>
<tr>
<td>EBLUP</td>
<td>2.95</td>
<td>30.32</td>
<td>X</td>
<td>19.27</td>
</tr>
<tr>
<td>Biproportional</td>
<td>20.95</td>
<td>41.60</td>
<td>18.29</td>
<td>X</td>
</tr>
</tbody>
</table>
### Mean estimates for 10 random selected domains

**Level 1** (*section*)

<table>
<thead>
<tr>
<th>voivodship</th>
<th>section</th>
<th>Population size</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
<th>Biprop. estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>I</td>
<td>13440</td>
<td>141795</td>
<td>133041</td>
<td>140420</td>
<td>129681</td>
</tr>
<tr>
<td>6</td>
<td>I</td>
<td>10929</td>
<td>117739</td>
<td>134874</td>
<td>118304</td>
<td>157186</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>5875</td>
<td>64395</td>
<td>123745</td>
<td>65033</td>
<td>67060</td>
</tr>
<tr>
<td>10</td>
<td>K</td>
<td>22307</td>
<td>107437</td>
<td>132523</td>
<td>108549</td>
<td>93956</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>797</td>
<td>100462</td>
<td>120023</td>
<td>104115</td>
<td>94316</td>
</tr>
<tr>
<td>18</td>
<td>E</td>
<td>64</td>
<td>191965</td>
<td>165299</td>
<td>181580</td>
<td>257689</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>1494</td>
<td>127090</td>
<td>122920</td>
<td>125239</td>
<td>120897</td>
</tr>
<tr>
<td>20</td>
<td>H</td>
<td>1998</td>
<td>96761</td>
<td>126316</td>
<td>98264</td>
<td>116450</td>
</tr>
<tr>
<td>22</td>
<td>A</td>
<td>1704</td>
<td>72486</td>
<td>123021</td>
<td>74174</td>
<td>132899</td>
</tr>
<tr>
<td>22</td>
<td>I</td>
<td>15600</td>
<td>173416</td>
<td>144536</td>
<td>170807</td>
<td>225678</td>
</tr>
</tbody>
</table>
## Comparison of mean estimates

**Level 2** ($n_Z$)

Median of

$$\left| \frac{\bar{y}_i^A - \bar{y}_i^B}{\bar{y}_i^A} \right|$$

for domains (in %)

<table>
<thead>
<tr>
<th></th>
<th>$\bar{y}_i^A$</th>
<th>$\bar{y}_i^B$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct estimator</td>
<td>Regression estimator</td>
</tr>
<tr>
<td>Direct</td>
<td>X</td>
<td>51.71</td>
</tr>
<tr>
<td>Regression</td>
<td>50.00</td>
<td>X</td>
</tr>
<tr>
<td>EBLUP</td>
<td>1.37</td>
<td>37.60</td>
</tr>
</tbody>
</table>
Mean estimates for 10 random selected domains

**Level 2** \((nz)\)

<table>
<thead>
<tr>
<th>voivodship</th>
<th>nz</th>
<th>Population size</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>29</td>
<td>4754</td>
<td>262074</td>
<td>207140</td>
<td>260362</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>624</td>
<td>245047</td>
<td>182656</td>
<td>233092</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>230</td>
<td>48585</td>
<td>172147</td>
<td>48903</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>6813</td>
<td>609635</td>
<td>320551</td>
<td>565087</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>1577</td>
<td>52411</td>
<td>171404</td>
<td>52510</td>
</tr>
<tr>
<td>18</td>
<td>44</td>
<td>1421</td>
<td>332394</td>
<td>235786</td>
<td>316475</td>
</tr>
<tr>
<td>20</td>
<td>70</td>
<td>741</td>
<td>39097</td>
<td>152516</td>
<td>39146</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>110</td>
<td>642762</td>
<td>272655</td>
<td>440887</td>
</tr>
<tr>
<td>28</td>
<td>45</td>
<td>525</td>
<td>1178753</td>
<td>464551</td>
<td>810326</td>
</tr>
<tr>
<td>32</td>
<td>39</td>
<td>4366</td>
<td>580053</td>
<td>353925</td>
<td>519978</td>
</tr>
</tbody>
</table>
## Relative RMSE for domains

### Level 1 (section)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
<th>Biproportional estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>min</strong></td>
<td>0.055</td>
<td>0.064</td>
<td>0.072</td>
<td>0.051</td>
</tr>
<tr>
<td><strong>max</strong></td>
<td>1.665</td>
<td>7.807</td>
<td>1.291</td>
<td>13.677</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>0.267</td>
<td>0.799</td>
<td>0.259</td>
<td>0.564</td>
</tr>
<tr>
<td><strong>median</strong></td>
<td>0.218</td>
<td>0.477</td>
<td>0.215</td>
<td>0.258</td>
</tr>
</tbody>
</table>

| RMSE < RMSE of direct estimator (percentage of domains) | 25.6 | 43.9 | 41.7 |
Relative RMSE for 10 random selected domains

**Level 1 (section)**

<table>
<thead>
<tr>
<th>voivodship</th>
<th>section</th>
<th>Population size</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
<th>Biprop. estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>I</td>
<td>13440</td>
<td>0.2636</td>
<td>0.0679</td>
<td>0.2443</td>
<td>0.1183</td>
</tr>
<tr>
<td>6</td>
<td>I</td>
<td>10929</td>
<td>0.1228</td>
<td>0.2358</td>
<td>0.1258</td>
<td>0.3726</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>5875</td>
<td>0.1321</td>
<td>1.1257</td>
<td>0.1341</td>
<td>0.1524</td>
</tr>
<tr>
<td>10</td>
<td>K</td>
<td>22307</td>
<td>0.2019</td>
<td>0.6275</td>
<td>0.2008</td>
<td>0.1103</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>797</td>
<td>0.3893</td>
<td>0.3720</td>
<td>0.3720</td>
<td>0.1706</td>
</tr>
<tr>
<td>18</td>
<td>E</td>
<td>64</td>
<td>0.3066</td>
<td>0.1710</td>
<td>0.2719</td>
<td>0.6731</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>1494</td>
<td>0.5626</td>
<td>0.1118</td>
<td>0.4641</td>
<td>0.1894</td>
</tr>
<tr>
<td>20</td>
<td>H</td>
<td>1998</td>
<td>0.1838</td>
<td>0.4369</td>
<td>0.1848</td>
<td>0.2476</td>
</tr>
<tr>
<td>22</td>
<td>A</td>
<td>1704</td>
<td>0.2086</td>
<td>0.8958</td>
<td>0.2106</td>
<td>0.7260</td>
</tr>
<tr>
<td>22</td>
<td>I</td>
<td>15600</td>
<td>0.1591</td>
<td>0.1490</td>
<td>0.1548</td>
<td>0.3343</td>
</tr>
</tbody>
</table>
Relative RMSE of direct estimator

Level 1 (section)
Relative RMSE of regression estimator

Level 1 (section)
Relative RMSE of EBLUP estimator

Level 1 (section)
Relative RMSE of biproportional estimator

Level 1 (*section*)
RMSE of regression estimator in relation to the direct one

Level 1 (section)
RMSE of EBLUP estimator in relation to the direct one

Level 1 (section)
RMSE of biproportional estimator in relation to the direct one

Level 1 (*section*)

![Histogram of RMSE values](image)

- X-axis: log scale
- Y-axis: Frequency
- Values range from 0.25 to 84
## Relative RMSE for domains

**Level 2** \((nz)\)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{min})</td>
<td>0.079</td>
<td>0.036</td>
<td>0.078</td>
</tr>
<tr>
<td>(\text{max})</td>
<td>5.115</td>
<td>212.980</td>
<td>2.832</td>
</tr>
<tr>
<td>(\text{mean})</td>
<td>0.327</td>
<td>1.686</td>
<td>0.309</td>
</tr>
<tr>
<td>(\text{median})</td>
<td>0.285</td>
<td>0.603</td>
<td>0.277</td>
</tr>
</tbody>
</table>

### RMSE < RMSE of direct estimator (percent of domains)

|                | 25.6 | 60.2 |

The table above compares the performance of different estimators in terms of RMSE for domains at Level 2. It shows the minimum, maximum, mean, and median values for each estimator, indicating that the EBLUP estimator generally performs better than the direct estimator, with 25.6% of domains having a lower RMSE compared to the direct estimator and 60.2% showing a higher RMSE.
### Relative RMSE for 10 random selected domains

**Level 2** ($nz$)

<table>
<thead>
<tr>
<th>voivod ship</th>
<th>nz</th>
<th>Population size</th>
<th>Direct estimator</th>
<th>Regression estimator</th>
<th>EBLUP estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>29</td>
<td>4754</td>
<td>0.1659</td>
<td>0.1059</td>
<td>0.1635</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>624</td>
<td>0.5209</td>
<td>0.0992</td>
<td>0.4739</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>230</td>
<td>0.3143</td>
<td>4.2687</td>
<td>0.3150</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>6813</td>
<td>0.1699</td>
<td>0.4673</td>
<td>0.1563</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>1577</td>
<td>0.1383</td>
<td>3.0235</td>
<td>0.1384</td>
</tr>
<tr>
<td>18</td>
<td>44</td>
<td>1421</td>
<td>0.3171</td>
<td>0.2209</td>
<td>0.2951</td>
</tr>
<tr>
<td>20</td>
<td>70</td>
<td>741</td>
<td>0.1817</td>
<td>5.8043</td>
<td>0.1819</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>110</td>
<td>0.3804</td>
<td>0.5519</td>
<td>0.3181</td>
</tr>
<tr>
<td>28</td>
<td>45</td>
<td>525</td>
<td>0.2029</td>
<td>0.6306</td>
<td>0.2211</td>
</tr>
<tr>
<td>32</td>
<td>39</td>
<td>4366</td>
<td>0.2506</td>
<td>0.3966</td>
<td>0.2281</td>
</tr>
</tbody>
</table>
Relative RMSE of direct estimator
Level 2 ($n_z$)
Relative RMSE of regression estimator

Level 2 ($nz$)
Relative RMSE of EBLUP estimator
Level 2 (nz)
RMSE of regression estimator in relation to the direct one

Level 2 (section)
RMSE of EBLUP estimator in relation to the direct one
Level 2 (section)
Conclusions

- Mean estimates for domains obtained with the use of synthetic estimators differ considerably from direct estimates. The results show that synthetic estimators seem to be significantly biased.

- EBLUP estimator gives the best results, as regards the MSE. It improves the results obtained by direct estimation, despite the weak relation between variable of interest and the auxiliary one. However, the improvement is not significant.

- Improvement, resulting from EBLUP estimation, concerns first of all these domains, for which the quality of direct estimation is the worst.

- For most domains $\gamma$ parameter approaches one. EBLUP estimator is close to the direct one, what reflects considerable bias of synthetic estimator.
Conclusions

- Regularities in distribution of $\gamma$ parameter by domains concern rather kinds of activity than voivodships and are better noticeable at the lower level of aggregation.

- The lower level of aggregation allows better modeling of relation between the variable of interest and the auxiliary one.

- Comparing the synthetic estimators, biproportional estimator seems to be better than the regression one, in respect of the MSE. Taking into consideration the two-dimensional structure results in considerable improvement of precision.
References


1. Introduction

1.1. Scope

This paper looks at why smoothing methods are needed for area statistics and what benefits that smoothing can provide. Some simple smoothing solutions are explored in the context of smoothing mortality statistics at ward level. The investigation has two main objectives. The first one is to establish whether spatial data smoothing methods appear to provide more stable estimates of underlying patterns of mortality than the approach undertaken presently by Health Statistics. The second is to identify which, if any, of the approaches is a suitable candidate for application and could be taken forward.

1.2. Why use smoothed data?

The smoothing solution brings benefits that go beyond enhancing the clarity of the information contained in a map of SMRs: firstly, smoothing helps stabilise rates based on small numbers at the desired level of spatial (dis)aggregation; secondly, smoothing reduces noise caused by different population sizes used in the calculation of rates. The ultimate gain is the increased ability of the user to discern systematic patterns in the spatial variation of underlying risk (Waller and Gotway, 2004; Kafadar, 1996). While the estimate of risk in any single area is optimal when location is not seen as relevant and independence across space is assumed, it is possible to derive improved estimates of the relative risk by building estimators that borrow strength in space. This means a change of approach from regarding each area separately to taking into account the ensemble of areas under study. This refinement is important when using SMRs as measures of the underlying area-specific risks, since they do not allow for spatial correlation ‘which may be induced through dependence on shared but unmeasured risk factors which themselves vary systematically in space’ (Best et al., 1999: 132). Spatially smoothed estimates are, therefore, the most appropriate ones for the judgement of geographic variation (Veugelers and Hornibrook, 2002).

It is important to recognise, however, that there are also disadvantages associated with smoothing. Firstly, smoothing introduces autocorrelation (i.e. correlation among...
neighbouring values) in the data, and may be seen as replacing unstable estimates with correlated ones. Secondly, the differences between the smooth and the raw values may be a source of concern for users. As Waller and Gotway (2004) warn, ‘[M]any people are leery of statistically adjusted numbers, particularly if money or power is to be allocated based on them’ (p.97). On balance, given how misleading inferences based on maps from raw rates can be, smoothing is regarded as preferable (ibid.).

1.3. Project objectives

In this project, we focus on the smoothing of SMRs, across time, across space and across both time and space. Our objectives are to:

- Define a selection of simple time and space smoothing techniques and apply them to ward-level SMRs;
- Evaluate the performance of the simple smoothing solutions defined;
- Identify the most suitable simple smoothing solution for the SMR data under analysis;
- Indicate the direction for future work in relation to smoothing SMRs.

We chose to concentrate on simple smoothing techniques for two reasons. On the one hand, as this is a first project devoted to the exploration of smoothing solutions in the Office, we aimed to start by clarifying the basic ideas of smoothing and on conveying the essence of the technique, best reflected in the most simple of the smoothers. On the other hand, we expected the users to use the smooth values in an exploratory way, for which purpose ‘a smoother that can be calculated quickly without complex estimation rules, difficult or tedious programming, and specification of extra parameters is preferred’ (Waller and Gotway, 2004: 98). More complicated solutions, closely associated with inferential statistics (such as Bayesian ones) are beyond the scope of the current project.

1.4. Introduction to SMR data

The framework of this project is determined by the nature of the data analysed (SMRs) and of the techniques applied to them (smoothing). We discuss each in turn briefly in this introduction.

By definition, the Standardised Mortality Ratio (SMR) is the ratio of the observed to the expected number of deaths in an area, multiplied by 100:

\[
SMR = \frac{\text{observed deaths}}{\text{expected deaths}} \times 100
\]  

(1)

**Observed** deaths are those which actually occurred in the local population in the year in question. **Expected** deaths are those which would be expected given the age structure and size of the local population, if it had the same age-specific death rates as the reference population.
In this project, the local population is that of each ward, while the reference population is that of England and Wales (male and female, below the age of 85). Observed deaths are calculated for each ward as the total deaths (i.e. the sum of deaths, across age groups up to 85, and gender).

The smoothing techniques applied here are linear in nature and, in essence, involve simple and weighted averaging over predefined neighbourhoods of values. These neighbourhoods include past values in time and sets of contiguous neighbours in space.

1.5. Smoothing methodology

The essence of smoothing is that, when data are noisy, an underlying component – free of error – can be made more visible or approximated via smoothing.

While this project fits in the broad view of smoothing, the details of the theoretical setup that underpins our work are defined by the specific aim of this analysis, namely that of finding the best estimator of underlying mortality in any particular area. In particular, the approach that we have taken to smoothing in this project rests on the view of the nature of the SMR data under analysis that we set out below.

Each observation \( y_{it} \) at every area or ward \((i)\) at every moment in time \((t)\) has, by definition, two components: mortality proneness or underlying mortality \( (u_{it}) \) and random mortality \( (e_{it}) \). Formally:

\[
\begin{align*}
  y_{it} &\equiv u_{it} + e_{it} \\
  e_{it} &\sim (0, \sigma_{e(i)}) \\
  E(e_{it}u_{it}) &= 0 \\
  E(e_{it}e_{ir}) &= 0
\end{align*}
\]

1.6. Performance criteria

In pragmatic terms, it is of interest to assess how much smoothing each technique imposes on the data. When measuring smoothness, we concentrate on the stability of
the smoothed data. This can be assessed by comparing, in pairs, outcomes from the application of each technique to data at different points in time. A stable smoother is one for which the difference in the outcomes from one point in time to another is small.

The comparison between any two smoothers is based on the mean squared differences between smooth values in repeated applications of the same technique at successive moments in time, t and (t+1):

\[ M\hat{SD} = \frac{1}{N} \sum \left( \hat{u}_{t+1,k} - \hat{u}_{t,k} \right)^2 \]  

where \( N \) is the total number of wards. We can compare any smoothing technique (k) to a chosen benchmark technique (b), by computing the following ratio \( M\hat{SDR} \), where the numerator is the \( M\hat{SD} \) for approach k and the denominator is the \( M\hat{SD} \) for the benchmark. A value of \( M\hat{SDR} \) below 1 for the ratio indicates that option k is more stable than the benchmark one.

A smooth series with high **predictive power** is one that correlates relatively highly with the un-smoothed data at the next point in time (one step ahead). Evaluation entails the comparison of the predictive power of more than one predictor on a given outcome. A relatively small difference between actual and predicted values is taken to mean good predictive power; the smaller the difference, the better the smoothing technique. We evaluate the correlation between the smooth at one moment in time (t) and the rough data one step ahead (t+1). Specifically, between any two smoothing solutions k and m (of which the latter could be a benchmark approach), the one with the lowest mean squared estimated error is to be regarded as the better one. The comparison involves:

\[ M\hat{SE}_k = \frac{1}{N} \sum (y_{t+1} - \hat{u}_{t,k})^2 \] vs. \[ M\hat{SE}_m = \frac{1}{N} \sum (y_{t+1} - \hat{u}_{t,m})^2 \]

The pair of estimated mean squared errors can be subject to the ratio comparison described above.

It is clear that a smoother estimate of underlying mortality does not necessarily mean a better estimate. Indeed, it is possible to over-smooth, by either miss-specifying the structural part or by losing some of the information contain in the random terms. Indeed, in extremis, the national mean is the smoothest estimate for each ward, which does not imply that it also is the best estimate as far as any individual area is concerned. It is important to be able to tell when an over-smooth has been produced.

**1.7. Some basic smoothing estimators**

The choice of methods to be implemented was dictated by the nature of the data, the software packages available and the time scale of this project. The smoothing solutions used in this project belong to the broad family of linear smoothers. The choice is justified on grounds of relative ease of implementation and of applicability to the type of

\footnote{We report the average \( \log(M\hat{SD}) \), in order to make proportional changes visible and facilitate comparisons.}
data under analysis. In most general terms of this approach, the smooth value of a target variable \( y_i \) is defined as:

\[
\hat{u}_i = \frac{\sum w_{ij} y_j}{\sum w_{ij}}
\]  

(5)

where \( w_{ij} \) are the weights that link any two observations \( i \) and \( j \), while \( \hat{u}_i \) is the smooth value for observation \( i \).

The specific approach we take is that of building time and space moving averages over three year time intervals and geographical neighbourhoods defined as the target area and the set of its contiguous neighbours, where the area is the ward.

The smoothing solutions detailed below are specific cases of (5), where values are given to the weights \( w_{ij} \) and the identity of the elements \( j \) (past points or neighbouring values) is specified.

1.7.1. **Smoothing over time**

In the data set we are using, observations are available only for five separate years. A time series of five annual observations is very short by the standards of time series analysis and therefore is not suitable for anything other than simple smoothing techniques.

The simplest scenario is one where all the values included in the calculation of the smooth have the same weight. Over time, equal weights do not distinguish between the present and the past in terms of the contribution to the smooth estimate, the generic form is specified as:

\[
\hat{u}_t = \sum_{k=0}^{T-1} \alpha_k y_{t-k}
\]

(6)

where the subscript \( t \) refers to the year to which the smoothing is applied and \( T \) denotes the length of the time-horizon for smoothing and \( \alpha_k \) represent weights.

Declining weights assign more importance to recent data points and less to points further in the past. We specify two sets of weights, corresponding to high persistence or memory (relatively high weights for the past) and low persistence or memory (relatively low weights for the past).

1.7.2. **Smoothing in space**

In spatial analysis terminology, ward data belongs to the category of lattice data, i.e. observations from a random process over a countable collection of contiguous spatial regions of varying size (the wards), together with a neighbourhood structure. In the neighbourhood structure used here, the neighbours are defined as wards that border
on each other. This structure means that, for each individual ward, the areas that matter in terms of spatial interaction are the immediate neighbours. By analogy with smoothing over time, the simplest way of smoothing in space is by replacing each individual value (for the target area) by a weighted average of itself and its neighbours. In our study we consider neighbours defined by first order (immediate) and second order (neighbours of first order neighbours) contiguity and with two types of ad hoc weights: equal and tapering over space, to mirror the structure of the time smoothing solutions described above. In general, neighbourhoods based on contiguity are specified through a matrix of spatial interactions $W$ with neighbours identified through their sharing a common border:

$$w_{ij} = \begin{cases} 1 & i \text{ and } j \text{ share common border} \\ 0 & \text{otherwise} \end{cases}$$

(7)

This approach is illustrated in Figure 1, which contains the map of a set of areas and the corresponding spatial interactions matrix using first-order neighbours. In this matrix, a value of 1 indicates that two areas are adjacent to one another, while a 0 indicates that the areas are not adjacent. The first line of the matrix contains information about the identity of the neighbours for area 1 (who shares a border with areas 2 and 3, the only non-zero values in the row). The second line contains the relevant information for area 2, and so on.

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure 1 Simple contiguity setup

2. ResultD

2.1. Summary of performance

The best space-time smooth identified has three main strengths: it recognises the importance of accounting for spatial and temporal structure in estimating mortality ratios; it has conceptual simplicity, being easy to explain and present and benefits from computational simplicity, as it is easy and quick to implement for large numbers of areas and time points.

Graphically, the overall information is summarised in Figure 2, which depicts the relative performance of the best smoothing solutions identified against that of the raw data, which is set as the base (equal to 1). The gain from smoothing in terms of both stability and predictive power is clearly visible for pure time, pure space and time-space

---

4 Spatial interaction among neighbours can also be expressed as a function of distance between the centroids of the areas involved, as a function of shared border or as a function of some flow variable (e.g. migration). Also, the weights can be specified ad hoc or defined through the use of a kernel. See section 5.3 of Carrington et al (2005) for further details.
smoothing. The largest increase in stability comes from the application of space-time smoothing. This chart is to be interpreted keeping in mind that, as far as predictive power is concerned, what matters is the relative positions, not the specific magnitudes.

As far as the status quo is concerned, it has the best performance amongst the time smoothing solutions evaluated. Further benefits in terms of stability and predictive power can be obtained through the spatial smoothing of the time-smoothed values.

**Figure 2: Benefits from smoothing (overall)**

An important problem with mapping SMRs by administrative areas is that the rates are estimated with very different precisions, depending on the population in each area. The methods explored here - and among which the best performing one has been identified - provide estimates that have less variance than the SMRs, but with differences that can be quite large between rural and urban areas. These differences undermine the interpretation of patterns across space (in a map), because estimates of different precision are being compared. Both the mean and the variance of the raw and smooth values are systematically lower in the case of rural areas.

The results for all cases of interest – raw values and best performing method in each category – indicate that systematic differences in performance are present. As a starting point, the raw data for urban areas is much more stable and has almost twice the predictive power than that for rural areas.

Details of the performance on urban and rural subsets appear in Figure 3. It is important to note that the relative positions in terms of performance for the methods compared are the same for the two subsets of data. The stability (smoothness) of the urban and rural values increases to comparable levels. The increase in predictive power due to smoothing is also present in both cases. In the case of urban values, the gain in predictive power from time smoothing is marginally higher than that from space-time smoothing (0.6990 vs. 0.6925). This is not the case overall or for the rural areas. This may be regarded as an indication that a spatial smoothing solution that differentiates
between urban and rural areas in terms spatial structure and interactions would be more suitable.

Figure 3: Benefits from smoothing (urban vs. rural)

Figure 4 illustrates the visual impact of smoothing in a pair of maps: one of raw values (left) and one of smooth values (right) produced with the best space-time approach identified, for the year 2001. It is clear from a visual comparison of the two maps that, on a global scale, smoothing makes the geographic pattern of variation more visible and that the contrast between extreme values is more easily apparent.
Figure 4: Raw data 2001 SMRs (left) and best space-time smooth 2001 SMRs (right).
2.2. Concluding comments

In general, we conclude that it is beneficial to apply some form of smoothing to annual SMR series if the objective is to extract the underlying variability in the data and to remove local noise. While most gain can be obtained from smoothing data over time and space, we also note that, compared to the raw data, well chosen spatial smoothing brings a fair increase in stability and some increase in predictive power. This relatively lower level of gain is, however, augmented by the fact that pure spatial smoothing on annual data eliminates the need for revisions from one year to the next, which can be the case for some time smoothing solutions. Further exploratory work in the use of more sophisticated techniques and their implications, as well as consideration of alternative evaluation techniques is desirable.

References


Smoothing Data for Small Areas
An evaluation of some simple techniques

Anca Carrington and Martin Ralphs
Office for National Statistics, UK
Overview of Presentation

- Why do we need smoothing methods for area statistics?
- What are the benefits of smoothing?
- Smoothing solutions explored
- Further work
What is smoothed data useful for?

- Extraction of broad patterns and general trends from noisy data
- Clearer visualisation of variation across space and time
- Maximising access to data that would otherwise be suppressed (disclosure control)
Why smooth small area statistics?

- Data become more sparse for:
  - smaller and smaller area geographies
  - sub-categories (e.g. mortality ratios by cause; property prices by dwelling type)
- Figures in areas with low incidence are unstable and potentially misleading – apparent differences may be wholly due to noise
Data used for illustration

- Standardised Mortality Ratios (SMRs) in England and Wales (at ward level, 8800 areas)

- HM Land Registry house price information in England and Wales (at Middle Layer Super Output Area level, 7193 areas)
Benefits of smoothing

- Support users in interpreting unstable data
- Reveal underlying trends in noisy data – useful for policy makers
- Increase data availability – by providing smoothed values in areas / categories where data was previously suppressed
Limitations

- Smoothing transforms the original data
- Smooth data values are correlated by construction
- On balance: *the visual inferences obtained from maps of raw rates can often be so misleading that smoothing is a better choice* (Waller & Gotway, 2004)
Project objectives

- Develop and evaluate methods suitable for smoothing area statistics – from simple to complex
- Produce programs to implement procedures
- Produce a guide to the procedures
Definition of Smoothing

- Observed event data like house sales, deaths, births, crimes etc. ($y$) at a moment in time ($t$) and in a given location ($i$) are the combination of two components:

$$y_{it} = u_{it} + e_{it}$$

- the measure of the underlying phenomenon of interest ($u$)
- error or noise ($e$)
Smoothing methods

- Methods which aim to retrieve the underlying component $u$ are generically known as SMOOTHING
- The same data can be smoothed in more than one way
- There is no ready-made smoothing tool available to do this
Smoothing mechanism

- In most general terms, the smooth value of a target variable $y_i$ is defined as a weighted average:

$$\hat{u}_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}}$$

$w_{ij}$ are the weights that link target observation $i$ and any other observation $j$.
Specification

- We evaluated a range of specifications across SPACE, across TIME, and across BOTH

- Differentiated by:
  - The identity of the areas included in the smooth
  - The weights allocated to the areas included
SPACE smoothing

- Neighbourhood structure: simple contiguity

```
W =
```

```
[1 2 3 4 5 6]
0 1 1 0 0 0
1 0 1 1 1 0
1 1 0 0 1 1
0 1 0 0 1 0
0 1 1 1 0 1
0 0 1 0 1 0
```
What makes a good smooth?

- We compared the performance of the smoothing methods against results achieved by the raw, unsmoothed data.

- We concentrate on two global measures of performance:
  - Predictive power: smooth(t) vs. raw(t+1)
  - Smoothness (stability): smooth(t) vs. smooth(t+1)
## Comparative Performance

<table>
<thead>
<tr>
<th>STABILITY</th>
<th>Correlation Coef.</th>
<th>Gain</th>
<th>PERFORMANCE</th>
<th>Correlation Coef.</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>0.5757</td>
<td>raw</td>
<td>raw</td>
<td>0.5757</td>
<td></td>
</tr>
<tr>
<td>space-smoothed</td>
<td>0.7359</td>
<td>27.8%</td>
<td>space-smoothed</td>
<td>0.6157</td>
<td>6.9%</td>
</tr>
<tr>
<td>time-smoothed</td>
<td>0.9280</td>
<td>61.2%</td>
<td>time-smoothed</td>
<td>0.6609</td>
<td>14.8%</td>
</tr>
<tr>
<td>time-space-smoothed</td>
<td>0.9610</td>
<td>66.9%</td>
<td>time-space-smoothed</td>
<td>0.6614</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

**Overall performance**

![Overall performance chart](chart.png)

- **raw data**
- **spatial smoothing, two lags, high spatial link**
- **time smoothing, 3 years with equal weights (status quo)**
- **space-time smoothing**

The chart illustrates the comparison of predictive power and smoothness across different smoothing techniques.
Summary of findings

- Most gain can be obtained from smoothing data over time and space.
- Well chosen spatial smoothing brings a fair increase in stability and some increase in predictive power.
- Findings are robust to alternative specifications of the boundary systems used, as well as to small changes in measurement error in the data.
Some Final Points

- An investigation of local properties shows that there are gains in both rural and urban areas, but that rural areas have higher instability in the raw data:

![Subset performance graph](image)
Quality implications

- Smooth data are more suitable for mapping measures at small area level
- Comparability of smooth values across areas is influenced by the smoothing solution employed
- The quality of geographical data is essential to the correct quantification of spatial relationships
Further work

- Refinement of methods (accounting for population variability; Bayesian approaches)
- Exploration of methods for individual rather than area data
- Alternative evaluation approaches (optimal smoothing parameters and cross-validation)
Evaluation Criteria

The STABILITY of a smoother is based on the mean squared differences between smooth values in repeated applications of the same technique at successive moments in time, $t$ and $(t+1)$:

$$MSD = \frac{1}{N} \sum \left( \hat{u}_{t+1,k} - \hat{u}_{t,k} \right)^2$$

To evaluate PREDICTIVE POWER, we measure the mean squared difference between the smooth at one moment in time $(t)$ and the rough data one step ahead $(t+1)$:

$$MSD_k = \frac{1}{N} \sum \left( y_{t+1} - \hat{u}_{t,k} \right)^2$$
Theoretical basis

\[ y_{it} \equiv u_{it} + e_{it} = f(u_{i^*t^*}) + \delta_{it} + e_{it} \]

- \( y_{it} \) = observation at area \( i \) in year \( t \)
- \( u_{it} \) = underlying component of \( y \)
- \( e_{it} \) = random part of \( y \) [noise]
- \( \delta_{it} \) = (random) part of \( y \) that remains unexplained by the best prediction based on the structural part
- \( S \) = a set of areas \( (i^*) \) and points in time \( (t^*) \) to which the current value in area \( i \) is related
TIME smoothing (ii)

- Moving average over three years, with equal weights
  \[ \hat{u}_t = \frac{1}{T} \sum_{k=0}^{T-1} y_{t-k} \]

- Moving average over three years, with declining weights
  \[ \hat{u}_t = \sum_{k=0}^{T-1} \alpha_k y_{t-k} \quad \sum_{k=0}^{T-1} \alpha_k = 1 \]

  (i) \( \alpha_{0h} = 0.5; \alpha_{1h} = 0.3; \alpha_{2h} = 0.2 \) (high memory)
  (ii) \( \alpha_{0l} = 0.7; \alpha_{1l} = 0.2; \alpha_{2l} = 0.1 \) (low memory)
SPACE smoothing (i)

- Spatially weighted average over contiguous areas, with equal weights

\[ \hat{u}_i = \frac{1}{n_i + 1} \left( y_i + \sum_{j \neq i}^{n_i} y_j \right) \]

- Spatially weighted average over contiguous areas, one spatial lag, unequal weights

\[ \hat{u}_i = \alpha y_i + \left(1 - \alpha\right) \frac{\sum_{j \neq i}^{n_i} y_j}{n_i}, \quad n_i \neq 0 \]

\[ \alpha = 0.9; \quad \alpha = 0.7; \quad \alpha = 0.5; \]
SPACE smoothing (ii)

- Spatially weighted average over contiguous areas, one spatial lag, unequal weights

\[ \hat{u}_i = \sum_{k=0}^{2} \alpha_k L_k(y_i) \quad \text{and} \quad \sum_{k=0}^{2} \alpha_k = 1 \]

\[ L_k(y_i) = W_k y_i \]

\[ W_0 = I \]
General Restriction Estimator in Small Area Estimation

Kaja Sõstra
Statistics Estonia
European Conference on Quality in Survey Statistics
24-26 April 2006
Small area estimation problem

• National sample surveys are usually designed for estimating at national level and in large groups.

• Sample size in small area (geographical or subgroup) is often too small for reliable direct estimates.

• Several methods have been designed for “borrowing strength” from neighbouring regions.
General regression estimator (GREG)

The generalised regression estimator (GREG) for an area is obtained by adjusting the direct estimator using the standard linear model.

\[
\hat{Y}_{d}^{\text{GREG}} = \frac{\hat{N}_d}{\hat{N}_d} \sum_{i \in u_d} w_i y_i + \left( x_d - \frac{\hat{N}_d}{\hat{N}_d} \sum_{i \in u_d} w_i x_i \right)^T \hat{\beta}
\]

\[
\hat{N}_d = \sum_{i \in u_d} w_i
\]
General regression estimator (GREG)

\( X_d = (X_{d1}, \ldots, X_{dp})^T \) is the vector of true totals of \( p \) covariates in the area \( d \)

\( \hat{\beta} \) is the least squares regression estimate assuming a standard linear model

\[ y_i = x_i \beta + \varepsilon_i \]

with independent errors \( \varepsilon_i \sim N(0, \sigma^2) \)
SAE and population total estimator

SAE perform differently depending on the size of area

For better results it is appropriate to use different estimates for smaller areas and large sub-populations

Obtained estimates do not satisfy the criteria that the sum of estimated small area totals is equal to estimated population totals:

\[ \hat{Y} \equiv \sum_{d=1}^{D} \hat{Y}_d \]

One solution of the problem is general restriction estimator
General restriction (GR) estimator

Consider a $k$-vector of unbiased estimators

$$\hat{\theta}_s = \left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right)'$$

The nonsingular covariance matrix is $V_\theta$

The parameters have to obey the set of $m$ linear restrictions

$$R \theta = c$$

where $R$ is $m \times k$ matrix of rank $m$
GR estimator (2)

The general restriction estimator that satisfies restrictions above is (Knottnerus, 2003)

\[ \hat{\theta}_s^R = \hat{\theta}_s + K\left(c - R\hat{\theta}_s\right) \]

\[ K = V_{\theta}R'(RV_{\theta}R')^{-1} \]

\[ V_{\theta|R} = Cov\left(\theta - \hat{\theta}_{s|R}\right) = (I - KR)V_{\theta} \]
GR estimator (example)

Consider two different samples from the same population.
Estimates for population totals $Y$ and $Z$ are obtained from sample 1 and for population totals $U$ and $Z$ from sample 2.

Parameters: $\theta_1, ..., \theta_4$

Restriction: $R = (0, 1, 0, -1)$
GR estimator for SAE (1)

Vector of estimated small area totals and vector $R$

$$\theta_s = (Y_1, \ldots, Y_D)'$$

$$R = (1, \ldots, 1)$$

Restriction: small area totals have to sum up to population total $Y$

$$R \theta = Y$$

$$\sum_{d=1}^{D} Y_d = Y$$
GR estimator for SAE (2)

For calculating restriction estimators the covariance matrix of the parameter vector is needed. Assuming the independence of regions the covariance matrix takes a form

\[ V_{\theta} = \text{diag}(\hat{V}_{Y_1}, \ldots, \hat{V}_{Y_D}) \]
GR estimator of small area total

Assuming that population total $Y$ is known from external sources or estimated from larger survey with high precision the GR estimator for small area $d$ takes a form

$$
\hat{Y}_d^{GR} = \hat{Y}_d + \frac{Y - \sum_i \hat{Y}_i}{\sum_i \hat{V}_{Y_i}} \cdot \hat{V}_{Y_d}
$$
Variance of GR estimator

Variance estimator of GR estimator of small area $d$

$$\hat{V}_{Y_d}^{GR} = \frac{\hat{V}_{Y_1} + \ldots + \hat{V}_{Y_{d-1}} + \hat{V}_{Y_{d+1}} + \ldots + \hat{V}_{Y_D}}{\sum_i \hat{V}_{Y_i}} \cdot \hat{V}_{Y_d}$$

Variance of GR estimator is decreased by

$$\hat{V}_{Y_d}^{GR} - \hat{V}_{Y_d} = \frac{\hat{V}_{Y_d}^2}{\sum_i \hat{V}_{Y_i}}$$
Simulation study

Population was generated for simulation study with target variable $Y$ and auxiliary variable $X$

Population size $N = 10,000$

Number of small areas $D = 3$

$Y$ and $X$ were normally distributed with correlation 0.85

1000 samples were drawn from population with sample size $n = 100$

Small area totals were estimated using GREG estimator.

GR estimators were calculated using known population total (GR1) and population total GREG estimator (GR2)
Performance of estimators

Measures for comparing the performance of estimators:
1) Absolute relative bias (ARB)
2) Relative root mean square error (RRMSE)

\[
ARB_d = \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\hat{Y}_d^{(m)} - Y_d}{Y_d} \right|
\]

\[
RRMSE_d = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left( \frac{\hat{Y}_d^{(m)} - Y_d}{Y_d} \right)^2}
\]
Performance criteria by small area

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARB</td>
<td></td>
<td></td>
<td></td>
<td>RRMSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GREG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GR1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GR2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

• GR estimator is a possibility for calibration the small area estimators to meet certain conditions.

• Simulation study showed that GR estimator performed slightly better than GREG estimator.