FROM THE EDITOR

This "*Special Issue*" is devoted mainly to ***small area estimation (SAE)*** methods. It should be reminded that our journal has already published four issues connected with SAE: in 1994 one issue (vol. 1, Number 6, 1994) from the Warsaw International Conference held in 1992, and in 2000 two issues (Vol. 4, Number 4, 2000, and vol. 4, Number 5, June 2000) from the Riga International Conference held in 1999. An additional issue was published in 2001 (vol. 5, Number 2, June 2001).

This issue contains eleven articles devoted to SAE methods, two articles in part **Other articles** related partly to SAE, and one **Book Review.**

There are following eleven articles devoted to SAE methods:

1. ***Benchmarking Hierarchical Bayes Small Area Estimators in the Canadian Census Undercoverage Estimation*** (by Yong You, J.N.K. Rao and P. Dick, from Canada). In this paper, hierarchical Bayes (HB) unmatched area level models are considered. Posterior means and posterior variances of parameters of interest are first obtained using the Gibbs sampling method. Then the HB estimators (posterior means) are benchmarked to obtain benchmarked HB (BHB) estimators. Posterior mean squared error (PMSE) is then used as a measure of uncertainty for the BHB estimators. The PMSE can be represented as the sum of the usual posterior variance and the squared difference of HB and BHB estimators. The authors evaluate the HB and the BHB estimators in the context of 1991 Canadian census undercoverage estimation.

2. ***Some Findings of the Eurarea Project — and Their Implications for Statistical Police*** (by P. Heady and M. Ralphs from UK). The Eurarea project, funded by the EU, was intended both to research technical aspects of SAE from survey data, and to provide Eurostat and European NSIs with broad recommendations for statistical policy on SAE. This paper focuses on the implications of Eurarea's findings for statistical policy. It briefly outlines the methodology, which Eurarea used to evaluate the effectiveness of alternative approaches to SAE, and summarises some of empirical findings. In the light of these findings it discusses the choice of estimation philosophy, and the way in which alternative sample-design and record-matching policies enable or prevent the use of alternative SAE techniques.

3. ***Simultaneous Estimation under Nested Error Regression Model*** (by L. Zhang from Norway). The author of this article (Zhang, 2003) proposed a frequentist method of simultaneous small area estimation under hierarchical models. This can be useful when various ensemble characteristics of the small

area parameters are of interest in addition to area-specific prediction. In this paper the author extends the approach under the nested error regression model, which allows for use of auxiliary information at the unit level. Simulations based on monthly wage data suggest that the simultaneous estimator has much better ensemble properties than the empirical best linear unbiased predictor, without losing much of the precision of the latter in area-specific prediction.

4. ***Small Area Estimation of Disability in Australia*** (by D. Elazar from Australia). The main purpose of this paper is to discuss intended approaches for an application of existing small area methods to the topic of disability. The paper as such does not introduce any new methodological approaches to small area estimation. This paper firstly discusses the context of small area estimation in Australia. The paper then details the various small area models, which are proposed for use. Both hierarchical Bayes and frequentist methods for estimating the proposed small area models are considered.

5. ***Applying Jackknife Method of Mean Squared Prediction Error Estimation in SAIPE*** (by T. Maiti from USA). The Small Area Income and Poverty Estimation (SAIPE) project is an ongoing Census Bureau project to estimate numbers of poor school-age children by state, county and ultimately school district in the United States based upon Current Population Survey (CPS) and Internal Revenue Service (IRS) data, together with information from the latest decennial census. The current county-level methodology relies on a Fay-Herriot model fitted to log-counts (by county) of related school-age children in CPS-sampled households. The present paper discusses the measure of errors of the SAIPE estimates of county level child poverty rates.

6. ***A Generalized Class of Composite Estimators with Application to Crop Acreage Estimation for Small Domains*** (by G.C. Tikkiwal and A. Ghiya from India). This paper defines a generalized class of composite estimators, using auxiliary information, for small domains under simple random sampling and stratified random sampling schemes. The proposed class of composite estimators has desirable consistency property, and it includes a number of direct, synthetic and composite estimators. Further, this paper demonstrates the use of the estimators belonging to the generalized class for estimating crop acreage for small domains and also compares their relative performance with the corresponding direct and synthetic estimators, through a simulation study. The study suggests the use of some composite estimators at the small domains under consideration level and thus up to the level of district under certain conditions.

7. ***Ratio Estimation for Small Domain with Subsampling the Non-respondents: An Application of Rao Strategy*** (by G.A. Udofia from Nigeria). In this article, the author considers modifications of some of the procedures

for global ratio estimation in single-phase sampling with subsampling the non-respondents proposed by Rao (1986) to obtain an estimate of mean for a small domain that cuts across constituent strata of a population with unknown weights. The bias and mean-square error of each of the modified estimators are obtained for comparison. Unlike Rao (1986), the population mean of the auxiliary variable is assumed to be unknown before the start of the survey and hence double sampling is applied. Stratified simple random sampling is considered. Similar work on the ratio estimators proposed by Rao (1986) and extension to other sampling designs are the subject of an on-going research by the author.

8. ***Considerations on Optimal Sample Design for Small Area Estimation*** (by G. Dehnel, E. Gołata and T. Klimanek from Poland). The paper focuses on SAE with two-stage sampling, with special emphasis on choices that need to be made about the levels of stratification and clustering. The study tested the empirical impact of the number and size of clusters on the characteristics of direct, synthetic and composite EBLUP estimators. The optimal sample allocation for a two-stage-design, in terms of domains was found to be very close to the optimal sample allocation from the population point of view. The gains in the small-area estimation were compared with the losses in the precision of the population mean estimator.

9. ***Problems of Estimating Unemployment for Small Domains in Poland*** (by E. Golata from Poland). The paper presents results of some attempts to estimate unemployment for small domains in Poland. These are the results of the research undertaken within the Eurarea project compared with some the author's research. The properties of the estimators are discussed from the domain specific point of view.

10. ***Efficiency of Modified Synthetic Estimator for the Population Proportion: A Monte Carlo Analysis*** (by T. Jurkiewicz and K. Najman from Poland). In this paper a two-stage estimation procedure is suggested. The first stage consists of applying some distance measures to identify the degree of similarity between the sample units from the investigated domain and sample units representing other domains. In the second stage, those units, which turned out to be similar to units from the domain of interest, are used to provide sample information with specially constructed weights. Authors present results of the suggested procedure using Monte Carlo experiments based on data obtained form a continuing vocational training survey of enterprises.

11. ***Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey*** (by J. Kubacki from Poland). The author presents the application of hierarchical Bayes methods to the estimates of unemployment size for small areas applied to the Polish Labour Force Survey (PLFS). The

constructed model includes the data obtained from published results of PLFS for regions in Poland and the 2002 Population Census data. The evaluation of quality of these methods was presented in comparison to the earlier used methods (direct estimation).

The second part of this issue under the title **Other Articles** contains two articles als o partly related to small area statistics:

12. ***Optimal Stratification Using Random Search Method in Agricultural Surveys*** (by M. Kozak from Poland). The paper contains considerations on an optimal stratification given by Rivest (2002) and adapted by Lednicki and Wieczorkowski (2003) to a problem of the stratification minimizing an overall sample size subject to a fixed precision of estimation in subpopulations. Five numerical experiments were carried out to present an efficiency of the data modification and to compare the proposed algorithm with the simplex method. Data from the Agricultural Census 2002 regarding a cereals area were used.
13. ***Utilisation of Administrative Registers in the Polish Official Statistics*** (by E. Walburg and A. Prochot from Poland). The paper presents previous stages of Polish Official Statistics Information System development in the field of administrative registers, executed and planned work.

The part **Book Review** contains a short review of: J. Wywial, ***Some Contributions to Multivariate Methods in Survey Sampling (***prepared by M. Szreder from Poland). As the reviewer has stressed, Professor J. Wywial presents his own results and interpretations of essential survey sampling problems.

Jan Kordos

The Editor

STATISTICS IN TRANSITION, April 2004
Vol. 6, No.5, pp. 631—640

BENCHMARKING HIERARCHICAL BAYES SMALL AREA ESTIMATORS IN THE CANADIAN CENSUS UNDERCOVERAGE ESTIMATION

Yong You[1], J.N.K. Rao[2] and Peter Dick[3]

ABSTRACT

The Fay-Herriot (1979) area level model, based on matched sampling and linking models, and a non-linear area level model, based on unmatched sampling and linking models (You and Rao, 2002), have been used in small area estimation to obtain efficient model-based estimators of small area totals and means. It is often desirable to benchmark the model-based estimators so that they add up to reliable direct survey estimators for large areas. In this paper, hierarchical Bayes (HB) unmatched area level models are considered. Posterior means and posterior variances of parameters of interest are first obtained using the Gibbs sampling method. Then the HB estimators (posterior means) are benchmarked to obtain benchmarked HB (BHB) estimators. Posterior mean squared error (PMSE) is then used as a measure of uncertainty for the BHB estimators. The PMSE can be represented as the sum of the usual posterior variance and the squared difference of HB and BHB estimators. We evaluate the HB and the BHB estimators in the context of 1991 Canadian census undercoverage estimation. The sum of the provincial BHB census undercount estimates is equal to the direct survey estimate of the census undercount for the whole nation.

Key words: Census undercoverage, Hierarchical Bayes, Posterior mean squared error, Unmatched models.

[1] Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, yongyou@statcan.ca.

[2] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6, jrao@math.carleton.ca.

[3] Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, dickpet@statcan.ca.

1. Introduction

Sample surveys are used to provide estimates not only for the total population but also for a variety of sub-populations (domains). Direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide reliable direct estimates for specific small domains. In making estimates for small areas, it is necessary to “borrow strength” from related areas to form indirect estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and current administrative records. It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data. Small area models may be broadly classified into two types: area level and unit level models. Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002) and Rao (2003b) presented overviews and appraisals of models and methods for small area estimation; see Rao (2003a) for a comprehensive account of small area estimation,. In this paper, we focus on area level models for small area estimation. In particular, we apply the hierarchical Bayes (HB) approach to unmatched sampling and linking area level models, defined in You and Rao (2002), to obtain model-based estimators for parameters of interest in small areas.

A basic area level model is the well-known Fay-Herriot (1979) model that includes a linear sampling model for direct survey estimates and a linear linking model for the parameters of interest. However, nonlinear linking models are often needed in practice to provide better model fit to the data. For example, if the parameter of interest is a probability or a rate within the range of 0 and 1, a linear linking model with normal random effects may not be appropriate. A customary linking model in this case could be a logistic regression or log-linear model. In Section 2, we consider this type of general area level models for small area estimation.

Another important problem is that the model-based estimators do not benchmark to reliable direct survey estimates for large areas. In order to protect against possible model mis-specification as well as possible overshrinkage, we benchmark the model-based HB estimates so that the benchmarked HB (BHB) estimates add up to the direct large area estimate. To measure the variability of the BHB estimators, we use the posterior mean squared error (PMSE), similar to the posterior variance associated with the HB estimators. It can be shown that the PMSE is simply equal to the sum of the posterior variance and a bias correction term, provided that the BHB estimator is a known function of the HB estimators; see Section 2.3 for details.

In Section 2, we present the unmatched sampling and linking models as well as the BHB estimators. We apply the proposed method in Section 3 to the 1991 Canadian census undercoverage estimation and obtain BHB estimates of the provincial level census undercoverage. Section 4 gives a summary and directions for further research.

2. Inference based on area level models

2.1. General area level models

Let y_i denote the direct survey estimator of the i-th small area parameter of interest θ_i. Following You and Rao (2002), we consider the following sampling model for y_i:

$$y_i = \theta_i + \varepsilon_i, \quad i = 1,...,m, \tag{1}$$

with $E(\varepsilon_i \mid \theta_i = 0)$, that is, the direct survey estimator y_i is design-unbiased for the small area parameter θ_i. The sampling variance of y_i is $V(\varepsilon_i \mid \theta_i) = \sigma_i^2$. The sampling variance, σ_i^2, is usually assumed to be known (see Section 3.3), but it may depend on the unknown parameter θ_i (You and Rao, 2002).

The unknown parameter θ_i is assumed to be related to area level auxiliary variable, x_i, through a link function $g(\cdot)$ with random area effects v_i:

$$g(\theta_i) = x_i'\beta + v_i, \quad i = 1,...,m, \tag{2}$$

where β is the vector of unknown regression parameters, and the v_i's are uncorrelated with $E(v_i) = 0$ and $V(v_i) = \sigma_v^2$, where σ_v^2 is unknown. Normality of the random effects, v_i, is also assumed.

The sampling model (1) and the linking model (2) are unmatched in the sense that they cannot be combined directly to produce a linear mixed effects model for small area estimation if the linking function $g(\cdot)$ is non-linear. If $g(\theta_i) = \theta_i$, then (1) and (2) represent the Fay-Herriot area level model.

2.2. Log-linear unmatched models

A very useful and important linking model for proportions or rates is the log-linear model, i.e.,

$$\log(\theta_i) = x_i'\beta + v_i, \quad i = 1,...,m. \tag{3}$$

The sampling model (1) and the linking model (3) can be presented in a hierarchical Bayes (HB) framework as follows:

$$y_i \mid \theta_i \sim N(\theta_i, \sigma_i^2), \quad i = 1,...,m; \tag{4}$$

and

$$\log(\theta_i) \mid \beta, \sigma_v^2 \sim N(x_i' \beta, \sigma_v^2), \quad i = 1,...,m. \tag{5}$$

The linking model (5) implies that the small area mean θ_i has a log-normal distribution with density function given by

$$f(\theta_i \mid \beta, \sigma_v^2) = \frac{1}{\sqrt{2\pi}\sigma_v \theta_i} \exp\{-\frac{1}{2\sigma_v^2}(\log\theta_i - x_i^T \beta)^2\}.$$

Prior distributions are assumed on the model parameters β and σ_v^2. By using a complete HB approach, we can obtain the posterior mean of θ_i as the HB estimator and the posterior variance of θ_i as the measure of variability of the HB estimator. Gibbs sampling method (Gelfand and Smith, 1990) with Metropolis-Hastings algorithm (Chip and Greenberg, 1995) can be used to find the posterior means and posterior variances; see You and Rao (2002) for details.

2.3. Benchmarked HB Estimators

Let $\hat{\theta}_i^{HB} = E(\theta_i \mid y)$ denote the HB estimator of θ_i and $V(\theta_i \mid y)$ the posterior variance of θ_i, where $y = (y_1,..., y_m)'$. Let $\hat{\theta}_i^{BHB}$ denote the benchmarked HB (BHB) estimator of θ_i such that $\hat{\theta}_i^{BHB}$ is a function of the HB estimators $\hat{\theta}_i^{HB}$, $i = 1,...,m$, i.e., $\hat{\theta}_i^{BHB} = h(\hat{\theta}_1^{HB},...,\hat{\theta}_m^{HB})$ for some function $h(\cdot)$, and satisfies the benchmark property:

$$\sum_{i=1}^{m} \hat{\theta}_i^{BHB} = \sum_{i=1}^{m} y_i .$$

For example, a ratio BHB (RBHB) estimator can be obtained as

$$\hat{\theta}_i^{RBHB} = \hat{\theta}_i^{HB} \frac{\sum_{k=1}^m y_k}{\sum_{k=1}^m \hat{\theta}_k^{HB}} .$$

To measure the variability associated with the BHB estimator $\hat{\theta}_i^{BHB}$, we use the posterior mean squared error (PMSE)

$$\text{PMSE}(\hat{\theta}_i^{BHB}) = E[(\hat{\theta}_i^{BHB} - \theta_i)^2 \mid y],$$

which is similar to the posterior variance associated with the HB estimator $\hat{\theta}_i^{HB}$. The PMSE of $\hat{\theta}_i^{BHB}$ is given by

$$\text{PMSE}(\hat{\theta}_i^{BHB}) = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i \mid y), \qquad (6)$$

as shown in the Appendix. Thus the PMSE of $\hat{\theta}_i^{BHB}$ is simply the sum of the posterior variance $V(\theta_i \mid y)$ and a bias correction term $(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2$. The PMSE is readily obtained from the posterior variance and the estimators $\hat{\theta}_i^{HB}$ and $\hat{\theta}_i^{BHB}$. For the ratio benchmarked estimator $\hat{\theta}_i^{RBHB}$, we have

$$\text{PMSE}(\hat{\theta}_i^{RBHB}) = [\hat{\theta}_i^{HB}(\frac{\sum_{k=1}^m y_k}{\sum_{k=1}^m \hat{\theta}_k^{HB}} - 1)]^2 + V(\theta_i \mid y).$$

3. Canadian census undercoverage estimation

3.1. Background

In Canada, a census is conducted every five years. However, the census does not enumerate all the inhabitants that should fill a census form on Census Day. In the 1991 Canadian census, it is estimated that about 3% of the population were not enumerated. Thus the census needs to be adjusted for undercoverage in order to properly represent the demographic picture of the country on Census Day. The Reverse Record Check (RRC) is used by Statistics Canada to measure the gross number of persons missed by the census. The RRC is a sample survey with a sample size of 60,000 persons, estimating the gross number of persons missed by the census. An Overcoverage Study is also conducted to measure the gross number of persons erroneously included in the census. The RRC results are combined with those of the Overcoverage Study to produce direct survey estimates of the net undercoverage for the nation and all provinces. The population estimates are based on the census counts adjusted for the estimated net undercoverage in the census.

3.2. Models and Inference

Following You and Rao (2002), we consider the following unmatched sampling and linking models to obtain HB and BHB estimates of provincial census undercoverage:

$$y_i = u_i + \varepsilon_i\,,\;\; i = 1,...,10\,,\; \varepsilon_i \sim N(0, \sigma_i^2) \qquad (7)$$

and

$$\log(u_i /(u_i + c_i)) = x_i'\beta + v_i\,,\;\; i = 1,...,10, \qquad (8)$$

where $u_i (= \theta_i)$ is the true undercoverage count for the *i*th province, y_i is the direct estimate of u_i and the sampling variances σ_i^2 are assumed to be known. The linking model (8) is a log-linear random effects model for the undercoverage rate $u_i /(u_i + c_i)$ which is a function of the undercoverage count u_i. The linking model (8) is more complex than the regular log-linear model (3). Following You and Rao (2002) and using a complete HB approach with the Gibbs sampling method, we can find the posterior means of the undercoverage counts u_i, and the associated posterior variances. Let $\hat{u}_i^{HB}$ denote the HB estimator of u_i. By using the simple ratio benchmarking approach given in subsection 2.3, we can obtain the ratio BHB estimator $\hat{u}_i^{RBHB}$.

3.3. Application and Results

We used the 1991 Canadian census undercoverage data in our analysis. We used log-transformation of the census count as the auxiliary variable, that is, $x_{1i} = \log(c_i)$, and the linking model for u_i is $\log(u_i /(u_i + c_i)) = \beta_0 + x_{1i}\beta_1 + v_i$. The sampling variances σ_i^2 were estimated through a generalized variance function of the form $V(y_i) \propto c_i^r$ and the smoothed estimates were then treated as the σ_i^2 (Dick, 1995).

To implement and monitor the convergence of the Gibbs sampler, we followed the basic approach given in Gelman and Rubin (1992). We independently simulated L=8 sequences, each of length t=2d, with d=5000. The first 5000 iterations of each sequence were deleted. To reduce the autocorrelation in the sequence, we took every 10th iteration of the remaining 5000 iterations, leading to 500 samples for each sequence.

Table 1 presents the direct, HB and RBHB undercoverage count estimates for the 10 provinces, together with the corresponding standard errors and coefficients of variations, CV = standard error/estimate. The standard error of the HB estimate is the squared root of the posterior variance, and the standard error of the RBHB estimate is the squared root of the corresponding PMSE. It follows from Table 1 that the HB and RBHB estimates have smaller CVs than the direct estimates, especially for some smaller provinces (PEI, NS). The RBHB estimates add up to the total of the direct survey estimates. The constraint RBHB estimates have slightly larger standard errors than the HB estimates due to the benchmarking property. The CVs of HB and RBHB estimates are roughly equal in this application.

Table 1. 1991 Canadian census undercoverage estimation

	Estimate			Standard Error			CV		
Province	Direct	HB	RBHB	Direct	HB	RBHB	Direct	HB	RBHB
NFLD	11566	10782	10925	1846	1471	1478	0.16	0.14	0.14
PEI	1220	1486	1506	366	289	290	0.30	0.19	0.19
NS	17329	17412	17643	3475	2474	2485	0.20	0.14	0.14
NB	24280	18948	19200	3333	3294	3304	0.14	0.17	0.17
QUE	184473	189599	192119	15400	15105	15314	0.08	0.08	0.08
ONT	381104	368424	373321	32260	31316	31697	0.08	0.08	0.08
MAN	20691	21504	21790	4310	3077	3090	0.21	0.14	0.14
SASK	18106	18822	19072	3416	2550	2562	0.19	0.14	0.14
ALTA	51825	55392	56128	7553	6591	6632	0.15	0.12	0.12
BC	92236	89929	91124	9096	8109	8197	0.10	0.09	0.09

NFLD: Newfoundland, PEI: Prince Edward Island, NS: Nova Scotia, NB: New Brunswick, QUE: Quebec, ONT: Ontario, MAN: Manitoba, SASK: Saskatchewan, ALTA: Alberta, BC: British Columbia.

3.4. Test of Model Fit

Following Datta et al. (1999) and You and Rao (2002), we used the posterior predictive *p* value to test the adequacy of model fit. The posterior predictive *p* value is defined as $p = \Pr(T(y^*,\theta) > T(y_{obs},\theta) \mid y_{obs})$, where y^* is a sample from the posterior predictive distribution $f(y \mid y_{obs})$, $T(y,\theta)$ is a discrepancy measure depending on the data *y* and on parameters θ and y_{obs} is the observed *y*. Let θ^* represent a draw from the posterior distribution of θ and let y^* represent a draw from $f(y \mid \theta^*)$, then marginally $y^* \sim f(y \mid y_{obs})$. Note that the probability is with respect to the posterior distribution given the observed data. If a model fits the observed data, then the two values of the discrepancy measure should be similar. In other words, if the given model adequately fits the observed data, then $T(y_{obs},\theta)$ should be near the central part of the histogram of the $T(y^*,\theta)$ values if y^* is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive *p* value is expected to be near 0.5 if the model adequately fits the data. Extreme *p* values (near 0 or 1) suggest poor fit. Computing the posterior predictive *p* value is relatively easy using the posterior simulations from the Gibbs sampler. For each simulated value θ^*, we draw y^* from $f(y \mid \theta^*)$ and then compute $T(y^*,\theta^*)$ and $T(y_{obs},\theta^*)$. The posterior predictive *p* value is estimated by the proportion of times $T(y^*,\theta^*)$

exceeds $T(y_{obs},\theta^{*})$. In the present application, the discrepancy measure that we used for overall fit is $T(y,\theta)=\sum_i (y_i-\theta_i)^2/\sigma_i^2$, where $\theta_i=u_i$. A similar discrepancy measure was used in Datta, et al. (1999). For the 1991 census undercoverage data, the estimated posterior predictive *p* value is 0.383, which suggests the adequacy of the model.

4. Summary

In this paper, we studied benchmarked HB (BHB) estimators for small area estimation based on unmatched sampling and linking models proposed by You and Rao (2002). The BHB estimates add up to the direct survey estimates for large areas to protect against possible model mis-specification and possible overshrinkage of the direct survey estimates. The benchmarking property is very appealing to survey practitioners. We used posterior MSE (PMSE) as a measure of variability of the BHB estimators. The PMSE is very easy to compute using the HB and BHB estimates and the posterior variance. We applied the BHB approach to the 1991 Canadian census undercoverage estimation and obtained the BHB undercoverage estimates for the 10 provinces across Canada.

For a future study, we propose to apply the BHB estimation approach to produce small domain estimates of missed persons in the 2001 census. Dick (2001) used an empirical Bayes (EB) approach with the estimates calibrated to direct estimates for large areas. However, the MSE approximation used in Dick (2001) did not account for this calibration. The proposed study should provide useful comparisons of the EB and BHB approaches. We also plan to study the problem of estimated sampling variances, as discussed in Dick and You (2003), for the BHB approach, using the HB models studied by You and Chapman (2003) to the census undercoverage problems.

APPENDIX

Proof of $\text{PMSE}(\hat{\theta}_i^{BHB}) = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i \mid y)$:

We have

$$\begin{aligned}
&\text{PMSE}(\hat{\theta}_i^{BHB}) = E[(\hat{\theta}_i^{BHB} - \theta_i)^2 \mid y] \\
&= E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB} + \hat{\theta}_i^{HB} - \theta_i)^2 \mid y] \\
&= E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 \mid y] + 2E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})(\hat{\theta}_i^{HB} - \theta_i) \mid y] + E[(\theta_i - \hat{\theta}_i^{HB})^2 \mid y] \\
&= (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i \mid y),
\end{aligned}$$

by noting that the cross-product term is equal to 0:

$$E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})(\hat{\theta}_i^{HB} - \theta_i) \mid y] = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})E[(\hat{\theta}_i^{HB} - \theta_i) \mid y] = 0.$$

REFERENCES

CHIP, S. and GREENBERG, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327—335.

DATTA, G.S., LAHIRI, P., MAITI, T. and LU, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074—1082.

Dick, J.P. (2001). Small domain estimation of missed persons in the 2001 census. In *Proceedings of the Survey Method Section, Statistical Society of Canada*, pp 37-46.

DICK, J.P. and YOU, Y. (2003). Methods used for small domain estimation of census net undercoverage in the 2001 Canadian census. Presented at the 2003 Federal Committee on Statistical Methodology Research Conference, Arlington, Virginia.

FAY, R.E. and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269—277.

GELFAND, A.E. and SMITH, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398—409.

GELMAN, A. and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457—472.

GHOSH, M. and RAO, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 55—93.

PFEFFERMANN, D. (2002). Small area estimation — new developments and directions. *International Statistical Review*, 70, 125—143.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175—186.

RAO, J.N.K. (2003a). *Small Area Estimation*. New York: Wiley.

RAO, J.N.K. (2003b). Some new developments in small area estimation. *Journal of Iranian Statistical Society*, 2, 145—169.

YOU, Y. and RAO, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3—15.

YOU, Y. and CHAPMAN, B. (2003). Small area estimation using area level models and estimated sampling variances. Methodology Branch Working paper, HSMD-2003-005E, Statistics Canada, Ottawa.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 641—653

SOME FINDINGS OF THE EURAREA PROJECT — AND THEIR IMPLICATIONS FOR STATISTICAL POLICY

Patrick Heady, Martin Ralphs[1]

ABSTRACT

The EURAREA project, funded by the EU, was intended both to research technical aspects of small area estimation (SAE) from survey data, and to provide Eurostat and European NSIs with broad recommendations for statistical policy on SAE. This paper focuses on the implications of Eurarea's findings for statistical policy. It briefly outlines the methodology which Eurarea used to evaluate the effectiveness of alternative approaches to SAE and summarises some of empirical findings. In the light of these findings it discusses the choice of estimation philosophy, and the way in which alternative sample-design and record-matching policies enable or prevent the use of alternative SAE techniques.

The purpose of the Eurarea project (outlined by Heady and Hennell (2001) in an earlier volume of this journal) was to investigate the performance of standard and innovative methods for small area estimation (henceforth SAE) in the European context, and to provide advice to Eurostat, and to European NSIs, on the appropriate use of SAE methods in the context of official statistics. The full range of results, including methodological findings, specimen programs and recommendations regarding statistical policy will be published in the Eurarea project reference volume, and made available on the project website[2]. The purpose of the present paper is to outline the main implications for statistical policy. While it will indicate the basis for these recommendations, readers are referred to the project reference volume for details of the methodological findings[3].

[1] The authors work at the Office for National Statistics, London.

[2] http://www.statistics.gov.uk/eurarea, which will be available for public access in summer 2004.

[3] The authors gratefully acknowledge the work of all partners in the Eurarea consortium, on whose joint work the findings of this paper are based. However, the opinions expressed are the authors' own, and this paper should not be taken as an official statement of the views either of the ONS or of the Eurarea consortium. A fuller review of the project's implications, expressing the views of the consortium as a whole, will be published in the forthcoming Eurarea reference volume.

We consider two main aspects of statistical policy:

- The choice of estimation approach;
- The data systems required to support small area estimation.

To provide the context for these general themes we first provide a brief review of Eurarea's assessment procedures, and what they have shown us about the performance of different estimators in two contexts:

- where the emphasis is on the (average) quality of the estimates for each individual area;
- where the objective is to describe the overall distribution of area values.

The basis of Eurarea's assessment procedures

The fundamental idea behind Eurarea was to assess the performance of different SAE methods by drawing simulated samples from real population data, as given in population registers or censuses, in six European countries[1]. The population data was available both in aggregated form for all the local areas in the country concerned (or part of it) and also in disaggregated form for every individual and address in those parts of the country included in the study. The data set included three target variables (unemployment, household income, and household type) and a range of covariates that could be used to predict these target variables. Target variables were given for both individual and area level[2]. The same applied to covariates so far as practicable, though some covariates were available at area level only. The areas whose characteristics (means and proportions) we were trying to estimate were NUTS3 areas, and also smaller areas corresponding to NUTS4 or 5 areas.

The simulation process consisted of drawing samples, using approximations to the sampling designs that would be used in practice, applying the various estimation procedures, and comparing their estimates to the true values for all areas in the study data-set. The process was repeated a large number of times (typically 500) and the distance between the estimates and true values was summarised using a number of criteria — most prominently *average empirical mean squared error*, which we define thus:

$$\text{AEMSE}\left(\hat{\bar{Y}}_d\right)=\frac{1}{DK}\sum_{d=1}^{D}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_d^{(k)}-Y_d\right)^2$$

where D is the number of areas in the country, and K is the total number of replicates.

[1] Spain, Italy, Poland, Finland, Sweden, Britain.

[2] In most instances the target variables were included in the original population data-base. However, some of the target variables had to be imputed using models derived from survey data.

In drawing the simulated samples we used approximations to the sample designs used by real surveys in the country concerned. The covariates we used, which were broadly comparable from one country to another, were also ones that would be available to statisticians in the country concerned. However, our treatment of the countries' data-systems departed from realism in one respect. Given our detailed data-sets, all data-matches required by the estimation process — whether of sampled units to areas, or of sampled units to unit-level covariate values — could be made without difficulty. In practical situations the availability of such matched data depends on the organisation of the country's statistical system, and on whether the statistician making the estimates is located within or outside the NSI. By making it possible to connect data in ways that might not be feasible in practice, the Eurarea simulations allow us to estimate the cost, in terms of lost precision, of the real-life restrictions on data-matching.

Some exemplary estimators

By and large, the estimators investigated by EURAREA were fairly simple. For instance the models we used had random intercepts but not random slopes, and we concentrated on frequentist approaches. This is not because we thought that the methods we restricted ourselves to were necessarily the best from among the whole family of model-based methods (see Pfeffermann 2002 and Rao 2003 for reviews). But we did think that the difference between these other kinds of model-based approach and those that we used would be less important than the differences between design-based and model-based approaches (Särndal 1984), and between methods within these two broad categories that drew on different ranges of data. Since the issues facing European ISIs at the moment are whether to use model-based approaches at all, and how to make the best use of the data provided by the different national statistical systems, it was appropriate to concentrate on relatively straightforward methods.

We considered the performance of four basic estimator types, which we define below[1]:

1. Direct Estimator: $\hat{\bar{Y}}_d^{\text{DIRECT}} = \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}y_{id}$, where $\hat{N}_d = \sum_{i\in s_d} w_{id}$

2. Generalised Regression Estimator (GREG):

$$\hat{\bar{Y}}_d^{\text{GREG}} = \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}y_{id} + \left(\bar{\mathbf{X}}_d - \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}\mathbf{x}_{id}\right)^T \hat{\boldsymbol{\beta}} \quad \text{where}$$

$\bar{\mathbf{X}}_d = (\bar{X}_{d,1},...,\bar{X}_{d,p})^T$ is a vector of p population mean covariates.

[1] Note that capitals denote population means or totals, while lower case letters refer to sampled quantities and to random effects.

3. Area-level Synthetic Estimator

A linear model with area-level covariates is fitted to the sample area means of the target variable.

The model is $\bar{y}_{.d} = \bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} + \xi_d$, and the estimator is $\hat{\bar{Y}}_d^{SYNTH} = \bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}$,

where the ξ_d are independent variables with mean *0* and variance

$$\left(\sigma_u^2 + \sigma_e^2 \big/ n_d \right),$$

and n_d is the sample size of area *d*. Note that $\hat{\sigma}_e^2$ is calculated using a pooled estimate of within-area variance.

4. Composite (EBLUP) Estimator

This is a weighted combination of estimators 1 and 3 above:

$$\hat{\bar{Y}}_d^{\text{EBLUP}} = \gamma_d \hat{\bar{Y}}_d^{\text{DIRECT}} + (1-\gamma_d)\bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}$$

where $\gamma_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$, and the parameter estimates are the same as for the area-level synthetic estimator.

Figure 1 shows how these different estimators perform[1] in practice, based on results of simulations in six European countries and considering two different geographical levels — NUTS 3 regions and smaller NUTS 4 and 5 areas. The underlying factors that explain their performance are, of course, the different sample sizes and the explanatory power of the covariates used in the models. The figure shows that even with the sample sizes that are typical for NUTS3 areas, synthetic and composite estimators have some tendency to produce estimates with lower MSEs than design-based estimators. The composite estimators, which in this case are optimally weighted combinations of direct and synthetic estimators, perform best in all cases. This finding is a good deal more pronounced at NUTS 4 and 5, where small sample sizes mean that design-based estimators tend to be subject to high levels of error.

[1] In the column graphs we present in this paper, estimators appear in performance order with the best at the top and the worst at the base of the column. The proportion of the column taken up by each estimator indicates its level of success based on performance rankings — the best estimators (with the lowest empirical MSE scores) are usually first in the rankings and as a result the sum of ranks across multiple simulations is low. They therefore occupy less volume than poorer performers.

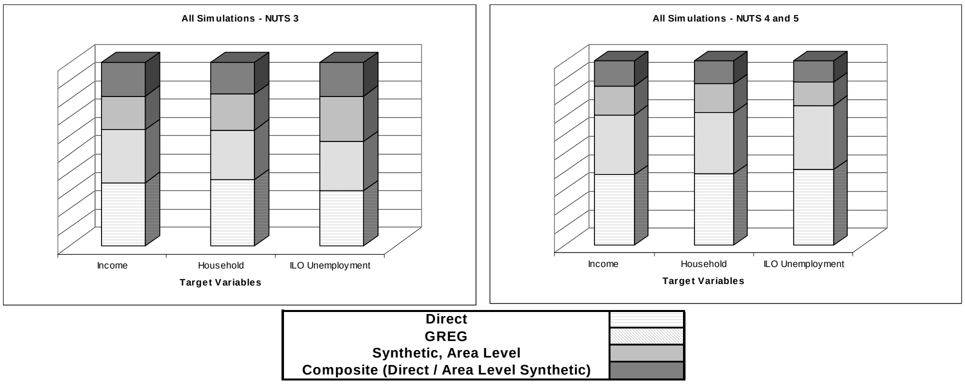


Figure 1. Performance of design and model-based estimators at NUTS 3 and NUTS4/5. Estimators appear in performance order, with the best at the top and the worst at the base of the column.

Variations on the model-based approaches

There are numerous ways of specifying the fixed effect part of model based estimators. A key distinction that we have investigated in Eurarea is the difference between models fitted at area level and those fitted at individual or unit level. In Figure 2 we compare the results achieved by these approaches and see that, in the majority of cases, area-level synthetic estimators tend to produce better results than their unit-level counterparts.

This is probably caused by the well-known “ecological effect”: the fact that regression and correlation coefficients calculated using data aggregated to area level, differ from those that would be calculated using the same data at unit level. In calculating the value of a synthetic estimator, the coefficients of the regression model are applied to area-level values of the covariates. If the coefficients have also been estimated using area-level data, as in the case of an area-level synthetic estimator, no problem arises. However, if the regression coefficients have been calculated at unit level, it is likely that they do not correctly estimate the relationship between the area-level averages on which synthetic estimation depends. This effect is discussed at greater length in Heady et al. (2000). It should be noted that if individual level models are expanded to include the corresponding area-level covariates they are no longer subject to this ecological bias. The models can, of course, only be expanded in this way if the modeller knows the identity of the area from which the individual sample units were drawn.

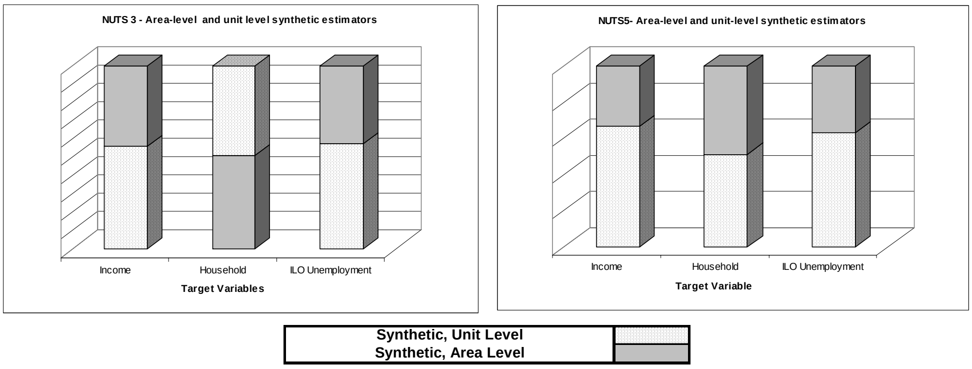


Figure 2. Ranked performance of unit and area-level synthetic estimators at NUTS3 and NUTS4/5. Estimators appear in performance order, with the best at the top and the worst at the base of the column.

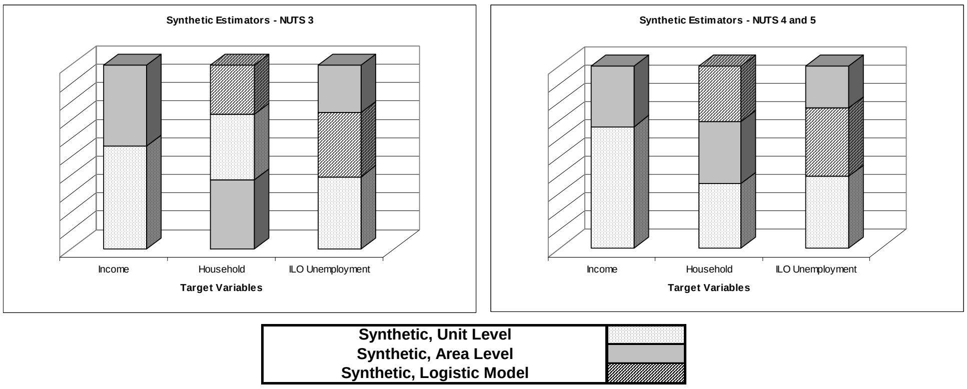


Figure 3. Ranked performance of unit level, area level and logistic synthetic estimators at NUTS3 and NUTS4/5.

In Figure 3, we introduce a third synthetic estimator based on an area-level logistic regression model. We thought that this might provide better estimates than the standard area-level synthetic estimator when the target variable is a proportion. In practice, neither estimator emerged as the clear winner in terms of minimising root MSE. The reasons for this are analysed in the forthcoming Eurarea reference volume.

Borrowing strength over space and time

The performance of the EBLUP can be enhanced considerably by including data from past years for the same areas — either by modelling the temporal autocorrelation of random effects, or by means of fixed effects models which include indicator terms for each time period. This simpler fixed effects approach seems to be equally effective in some practical situations.

Building in spatial autocorrelation has brought less benefit, either because we needed more sophisticated distance metrics, or because the spatial autocorrelation was already largely accounted for by the covariates. (Again, details of these findings will be given in the Eurarea reference volume, and on the project web site.)

Estimating the shape of the distribution of area values

An important consideration for policy makers is how effectively different SAE methods capture the variability between areas. This is particularly relevant for those resource allocation problems in which one or other tail of the area distribution is of interest. An example of such a problem is the allocation of European Union funding for regional and sub-regional support. Such funds are typically distributed to member states on the basis of the number of its (sub)-regions which score highly in terms of some indicator of need. If the (sub)regional scores are estimated it is important that the estimates preserve the shape of the underlying distribution, so that the correct proportion of areas fall within the upper tail.

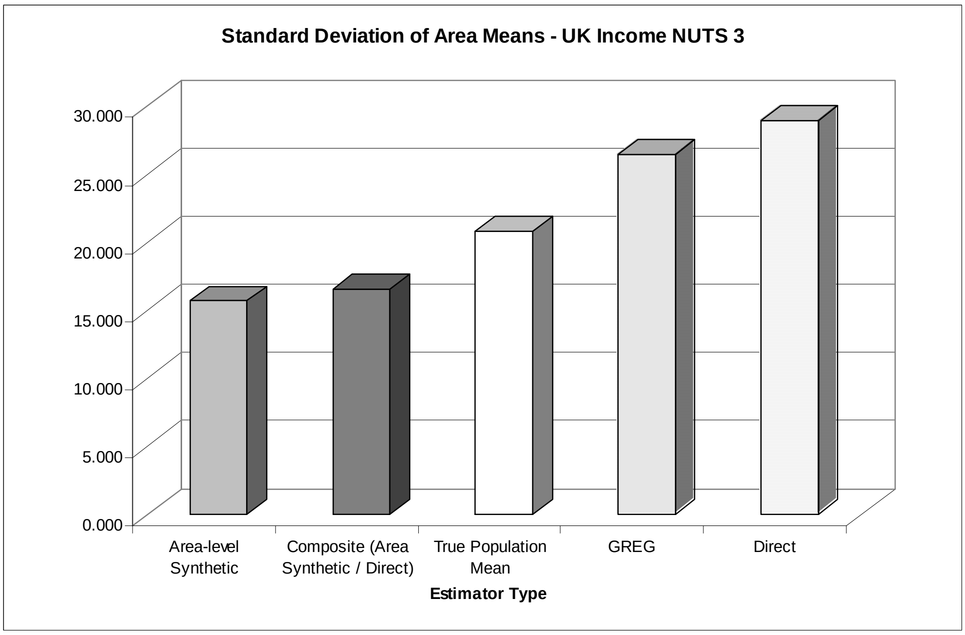


Figure 4. Comparing the standard deviations of the area means produced by different estimation methods

The tendency of small area estimates to distort the underlying pattern of between-area variation was discussed by Spjøtvoll and Thomsen (1987), who noted that the direct estimates for a set of areas were more scattered than the true underlying values, while model-based Empirical Bayes estimates were "over-shrunk" towards the predicted values and therefore understated the number of areas with exceptionally high or low values on the target variable. Similar calculations on our data confirm this contrast between the performance of design-based and model-based estimators, and also allow us to assess the relative degree of over-dispersion and over-shrinkage of the figures produced by different estimators within these two families.

Figure 4 explores how each estimation method approximates the true variability between areas by comparing the standard deviation of the estimated means with the standard deviation of the set of true population means. As any one sample may be untypical, we compare the true standard deviation with the average of the standard deviations calculated from the individual sample replicates.

We see that design-based estimators tend to overestimate between-area variability, while the synthetic estimates tend to underestimate it. The composite estimator provides the closest approximation to the truth, a finding which was repeated in all of our experiments, but even in this case the true variation is usually over or understated to some degree.

The reasons for this pattern are linked fundamentally to the way in which the estimators operate. Design-based estimators are highly vulnerable to the size and representativeness of samples. Variability in sample quality and, in particular, small sample sizes, mean that they are likely to lose much of the underlying pattern of true values, especially in small NUTS 4 and 5 areas. For the same reasons, the distribution they produce is also more extreme — with more areas with very low and very high proportions than is the case in reality. The distribution of synthetic estimates is usually closer to reality, but it too loses some definition. In this case, though, the estimator tend to smooth out real differences because of bias (very low values are overestimated, while very high values are underestimated), resulting in overshrinkage.

With model-based estimates, including synthetic, we can correct for overshrinkage by adding back in an appropriate amount of between-area estimation. But to do this, one has to have constructed the model and estimated its parameters. (See Shen and Louis 1998, and Rao 2003: 211—214.) This kind of adjustment is not possible with design-based estimators.

Design-based and model-based approaches

Alongside the literature on the technical properties of alternative small area estimators, there exists a certain amount of literature on the issues that arise when applying these estimators in practice. Schaible (1996) and Gambino and Dick

(1999) give accounts of the application of these techniques in the USA and Canada respectively. In this literature disquiet is often expressed (for instance by Gambino and Dick) about the use of model-based estimates, design-based methods being considered preferable. Marker (2001) specifically reviews, and recommends, methods that avoid the need to use model-based methods. A feeling that bias is worse than other kinds of error seems to be somewhere near the root of this desire to minimise reliance on model-based methods.

We share the concern with the way estimators work in practice, but not (as our use of the AEMSE criterion shows) the feeling that bias is somehow worse than other kinds of error. Starting from this desire to minimise error from all sources, what conclusions can we draw from Eurarea's findings about the use of different kinds of estimators at NUTS3 level and below. In particular, are design-unbiased methods or model-based methods to be preferred?

The findings reviewed above are fairly unambiguous. As we have just seen, model-based methods are preferable because:

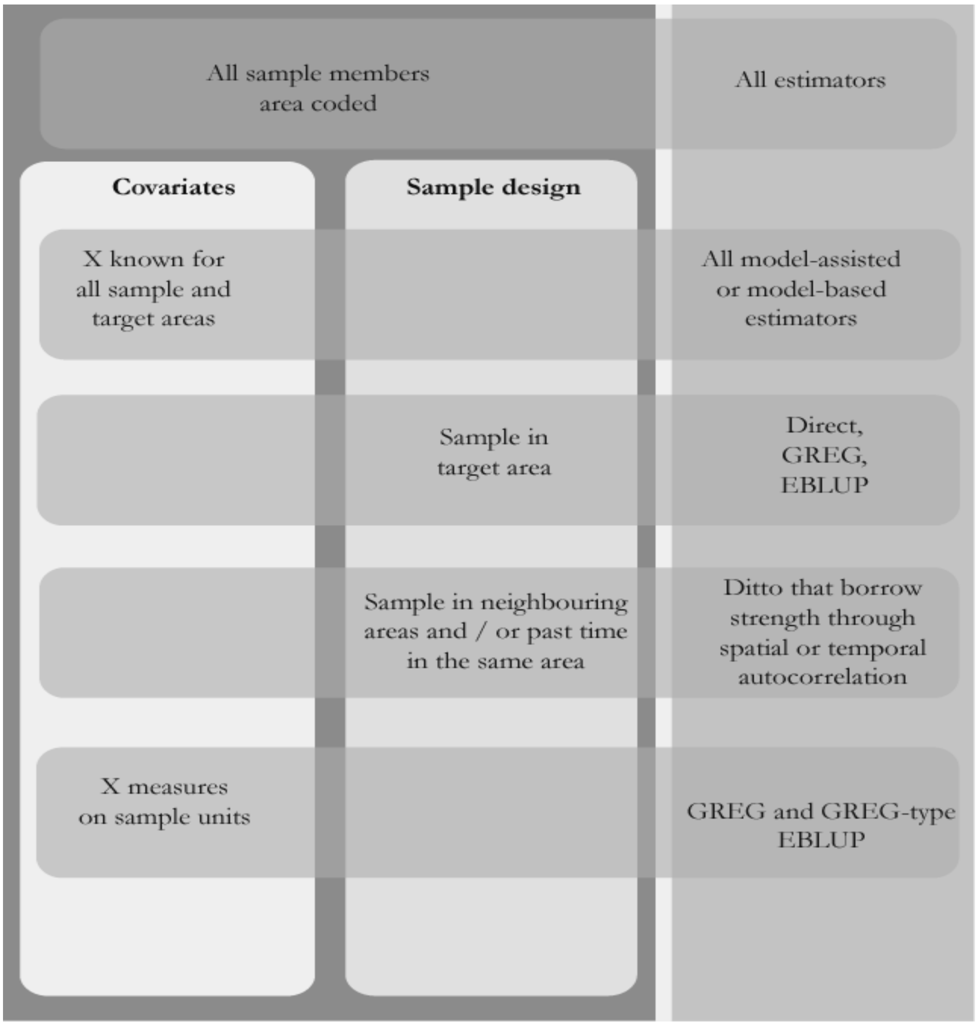
- they provide better (lower MSE) estimates, particularly at NUTS4/5 level, and
- they also make it possible to address the task of describing the overall population distribution of area means.

On the other hand, synthetic estimates which are based purely on the fixed effects part of the model are not optimal. Composite EBLUPs, which are in a sense a compromise between synthetic estimates and design-based estimators, are better. Composite estimators that make use of data from previous time points are best of all.

Data-requirements and statistical systems

The choice of appropriate estimators is only one aspect of the statistical policy required to produce reliable figures at regional and sub-regional level. The collection and management of appropriate data is at least as important. Indeed this is an aspect of active concern to European NSIs as shown by the efforts devoted to reviewing and enhancing the supply of sub-national data. (For current developments in Britain see Allsopp 2003.) In the final part of this article we consider how the choice of appropriate local estimators interacts with the structure of the data-system constructed by the NSI and available to analysts. In particular we consider the impact of alternative sample designs, and of the ways in which it is possible to match sample and covariate data.

Figures 5 and 6 set out the data requirements and analysis options for the families of small area estimation methods that we have considered in Eurarea. In Figure 5, we explore the dependencies of specific families of estimator in terms of covariates and sample designs. The diagram is divided into two sections. On the left, we consider data issues. On the right, we consider estimators. The horizontal bands indicate sets of linked requirements in terms of covariates, sample designs and estimator families, so for example any model-assisted or model-based estimator requires that area-level covariate data be available for all sample and target areas.



All sample members area coded		All estimators
Covariates	**Sample design**	
X known for all sample and target areas		All model-assisted or model-based estimators
	Sample in target area	Direct, GREG, EBLUP
	Sample in neighbouring areas and / or past time in the same area	Ditto that borrow strength through spatial or temporal autocorrelation
X measures on sample units		GREG and GREG-type EBLUP

Figure 5. Data requirements by estimator type

The first row of the table reminds us that a basic requirement of all the methods considered here is that the data for each sample member should be area-coded. While this is implicit in the formulae for all the estimators given above, it is worth noting that the beta parameters of a unit-level synthetic estimator can be estimated without reference to the location of the sample units. The reason why we have nevertheless indicated that area-coding is a requirement of all reliable estimation methods is that, as we have seen above, unit-level models (without additional area-level covariates) are subject to strong ecological biases — and therefore should not be used.

If one wishes to use a composite estimator, such as an EBLUP, it is necessary to have a sample in the area concerned which means that, if composite estimators are to be the standard approach, it is necessary to have samples in all areas at that level of geographic concentration. This effectively implies a geographically unclustered sample if the areas concerned are at NUTS4 or 5 level. If the EBLUP is enhanced by using data from previous time points, it is necessary that these too should be available for all or most target areas, which further strengthens the attraction of unclustered samples.

Data-matching at individual level is less important than either access to area-level covariates or unclustered samples. The reason for this is that GREGs, the main kind of estimator for which unit-level covariates are required, are less effective in terms of AEMSE than either synthetic estimators or EBLUPs. However, access to unit-level covariate data does have some value, primarily because it permits the use of composite estimators in which the direct component is replaced by a more accurate GREG.

In Figure 6, we revisit this information in a whole-system context and consider how sampling and covariate availability will determine which estimation methods are possible. Adjacent hexagons in the diagram represent specific analytical paths, all of which begin with the central requirement that we have access to area-coded survey data and end with possibilities for small area estimation given covariate and sampling configurations. For example, if we have access to area coded survey data, area-level covariates and clustered samples, we are restricted to the use of synthetic estimates for small area estimation.

This diagram provides a quick reference for deciding which small area estimation methods are available given the data configuration in the country concerned. Thus the situation in the UK mostly corresponds to the upper left-hand arm of the diagram, which means that (except for estimates derived from the unclustered Labour Force Survey) the only available approach is synthetic estimation (Heady, Clarke and others 2003). In contrast the data systems in Sweden and Finland — with their unclustered samples which NSI statisticians can link to individual-level covariate data in their population-registers - correspond to the bottom right hand arm of the diagram and permit all of the estimation strategies outlined above.

The diagram can also be taken as a representation of the alternative data strategies that any NSI might chose to adopt. As the results above have shown, countries situated on the upper left-hand arm could considerably improve their local estimates by de-clustering their samples and so moving to the lower left-hand arm. The further move to the Scandinavian position would only bring a fairly limited further improvement to the quality of survey-based local estimates.

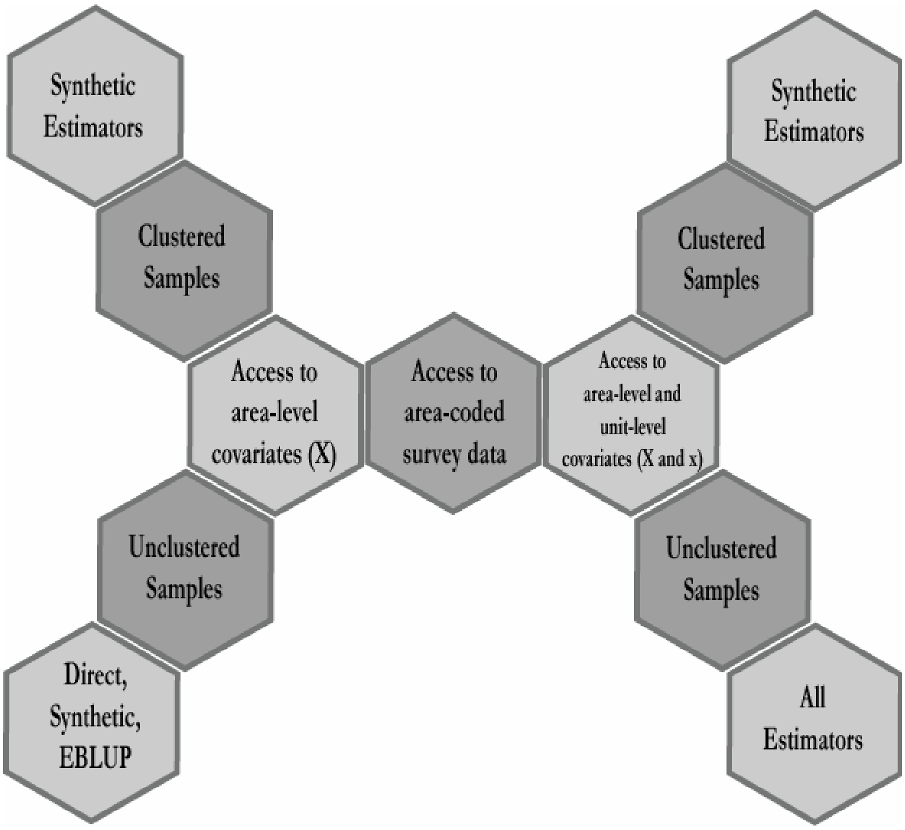


Figure 6. Data and sampling strategy requirements for SAE methods

Small area estimation and statistical policy

Until recently, the mixture of user requirements, cost restrictions and secondary dependencies arising from survey data collection processes have meant that optimisation of data collection policy for small area estimation has not been a primary concern for policy makers. However, the recent focus on sub-regional statistics in many EU member states and the requirement for increasingly detailed local information to inform regional policy indicate that this position is changing — and both data and estimation strategies are currently under review. In this paper, we hope to contribute to this process by providing some clarification of how the interdependencies between SAE and data collection policies will impact on local information agendas.

REFERENCES

ALLSOPP C. (2003) Review of statistics for economic policy making. Norwich: HMSO.

GAMBINO J., DICK P., (1999) Small area estimation practice at Statistics Canada. Statistics in Transition 4: 597—610.

HEADY P., CLARKE P., BROWN D., D'AMORE A., MITCHELL B. (2000) Small area estimates derived from surveys: ONS central research and development programme. Statistics in Transition 4: 635—648.

HEADY P., HENNELL S. (2001) Enhancing small area estimation techniques to meet European needs. Statistics in Transition 5: 195—203.

HEADY P., CLARKE P. & OTHERS (2003) Small area estimation project report. Model-based small area estimation series No.2. London: Office for National Statistics.

MARKER D.A. (2001) Producing small area statistics from national surveys: methods for minimising use of indirect estimators. Survey Methodology 27: 183—188.

PFEFFERMANN D. (2002) Small area estimation — new developments and directions. International Statistical Review 70: 125—143.

RAO J.N.K. (2003) Small Area Estimation. Hoboken NJ: Wiley.

SÄRNDAL C.E. (1984) Design-consistent versus model-dependent estimation for small domains. JASA (79) 642—631.

SCHAIBLE W.L. (Ed) (1996) Indirect estimators in U.S. Federal Programmes. Lecture Notes in Statistics. Springer-Verlag: New York.

SHEN W., LOUIS T.A. (1998) Triple-goal estimates in two-stage hierarchical models. JRSS(B) 60: 455—471.

SPJØTVOLL E., THOMSEN I. (1987) Application of some empirical Bayes methods to small area statistics. Bulletin of the International Statistical Institute 2: 435—449.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 655—665

SIMULTANEOUS ESTIMATION UNDER NESTED ERROR REGRESSION MODEL

Li-Chun Zhang[1]

ABSTRACT

Zhang (2003) proposed a frequentist method of simultaneous small area estimation under hierarchical models. This can be useful when various ensemble characteristics of the small area parameters are of interest in addition to area-specific prediction. In this paper we extend the approach under the nested error regression model (Battese, Harter and Fuller, 1988), which allows for use of auxiliary information at the unit level. Simulations based on monthly wage data suggest that the simultaneous estimator has much better ensemble properties than the empirical best linear unbiased predictor, without losing much of the precision of the latter in area-specific prediction.

***Key words**:* area-specific prediction; ensemble statistics; bootstrap.

1. Introduction

In small area estimation problems, the ensemble characteristics of the small area estimators (Judkins and Liu, 2000), i.e. when these are viewed as a collection of statistics, is often of as much interest as each area-specific estimator. Such ensemble characteristics include the variance, the rank ordering, the mini- and maximum, the range, the percentiles, *etc.* of the small area parameters. Estimators that are optimal for prediction of each specific area may have unsatisfactory ensemble properties. For instance, the between-area variation of the estimates can be much smaller than the true variation in the population, which is known as over-shrinkage. Various constrained Bayes approaches have been developed (Louis, 1984; Spjøtvoll and Thomsen, 1987; Lahiri, 1990; Ghosh, 1992). For a two-stage hierarchical model without auxiliary covariates, Shen and Louis (1998) proposed "triple-goal" estimators that produce good ranks, a good distribution and good area-specific estimators. The authors also noted that the approach can be generalized under models with a regression intercept and slope.

[1] Statistics Norway, Kongensgt. 6, P.B. 8131 Dep., N-0033 Oslo, Norway. E-mail: lcz@ssb.no

Zhang (2003) proposed a frequentist method of simultaneous estimation under basically the same two-stage hierarchical model. Suppose we are interested in, say, the mean of a variable from a large number of small areas, denoted by θ_i, for $i = 1,...,m.$ At the lower level of the model, we assume a parametric distribution of θ_i; at the upper level, we assume a conditional distribution of the data given θ_i. The simultaneous estimates are derived in two steps. Firstly, since the number of small areas is finite, one of them must have the smallest value among all the θ_i's, another must have the second smallest value, and so on. Let $\theta_{(i)}$ be the ith order statistic of $\{\theta_i\}$, i.e. the set of all the θ_i's, where $\theta_{(1)} \le \theta_{(2)} \le \Lambda \le \theta_{(m)}$. Denote the expectation of $\theta_{(i)}$ by

$$\eta_i = E[\theta_{(i)}; \xi]$$

where ξ is the parameter of the distribution of θ_i. It follows that η_1 is the best predictor for $\theta_{(1)}$, and η_2 is the best predictor for $\theta_{(2)}$, and so on, and $\{\eta_i\}$ is the best ensemble predictor of $\{\theta_i\}$. Let $\hat{\xi}$ be an estimator of ξ and $\hat{\eta} = E[\theta_{(i)}; \hat{\xi}]$, then $\{\hat{\eta}_i\}$ is the estimated best ensemble predictor.

Secondly, we match $\{\hat{\eta}_i\}$ with the small areas. Let $\hat{\theta}_i$ be the estimated best area-specific predictor, for $i = 1,...,m.$ Instead of using $\hat{\theta}_i$ directly, we obtain the rank of $\hat{\theta}_i$ among all the $\hat{\theta}_i$'s, denoted by $r_i = rank(\hat{\theta}_i)$. These are now used to match the estimated best ensemble predictors, and the simultaneous estimator of area i is given by

$$\overset{\&}{\theta}_i = \hat{\eta}_{r_i}.$$

In this way, the simultaneous estimates have the same rank ordering as the area-specific estimates. In the special case of ties among the $\hat{\theta}_i$'s, we assign the corresponding $\hat{\eta}_i$'s randomly.

The simultaneous estimator is not optimal for area-specific prediction. However, they have better ensemble properties due to the use of the best ensemble estimators $\hat{\eta}_i$'s. Empirical results (Zhang, 2003), validated by the true population values, suggest that the simultaneous estimator performs similarly as the Bayesian alternatives, i.e. producing estimates with good ensemble as well as area-specific properties. Exactly how big is the trade-off between the gain in ensemble statistics and the loss in area-specific precision, however, depends on the particular situation and must be evaluated on a case-to-case basis.

In this paper, we extend the approach of Zhang (2003) to regression models. In our derivation we concentrate on the one-fold nested error regression model

(Battese, Harter and Fuller, 1988), which allows us to incorporate auxiliary information at the individual level. We also allow for nonparametric specification of the distribution of the random errors, which is another difference from the case of two-stage hierarchical model above. In Section 2, we show that the best linear unbiased predictor (BLUP) entails loss of between-area variation under the nested error regression model. In Section 3 we present the simultaneous estimator. Section 4 contains a simulation study based on the monthly wage data, where the simultaneous estimator is compared to the empirical best linear unbiased predictor (EBLUP) and the direct estimator, both in situations with and without auxiliary covariates. A short summary is given in Section 5.

2. Between-area variation under nested error regression model

The nested error regression model to be considered is given as

$$y_{ij} = x_{ij}^T \beta + u_{ij} \qquad \text{and} \qquad u_{ij} = v_i + e_{ij} \tag{2.1}$$

where *j* is the subscript for unit *j* within area i, y_{ij} is the variable of interest and x_{ij} is the vector of unit-level covariates, and β contains the regression coefficients. The random error u_{ij} is the sum of an area-level effect v_i and a unit-level effect e_{ij}. The random errors v_i's and e_{ij}'s are assumed to be independent, with zero mean and variance σ_v^2 and σ_e^2, respectively. Apart form the first two moments, we do not require full specification of the distribution of the random errors in general.

Suppose we are to estimate the small area means of y_{ij}, denoted by $\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, where N_i is the size of area i. Let $\theta_i = \overline{X}_i^T \beta + v_i$, where $\overline{X}_i$ is the mean of x_{ij} within area i, which is the expected area mean conditional on v_i. The difference between θ_i and $\overline{Y}_i$ is the within-area population average of the unit-level random effects. Given the covariates of the population, we have

$$E[\frac{1}{m-1} \sum_{i=1}^{m} (\theta_i - \overline{\theta})^2 \mid \overline{X}_1, \ldots, \overline{X}_m] = \Delta + \sigma_v^2 \tag{2.2}$$

where $\overline{\theta}$ is the average of all the θ_i's, and $\overline{X}$ is the average of all the $\overline{X}_i$'s, and

$$\Delta = \frac{1}{m-1} \beta^T \sum_i (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})^T \beta .$$

From (2.2), it is seen that the variation in θ_i decomposes into two parts, where the first part is what can be accounted for by the covariates through the regression model, and the second part is what needs to be attributed to random effects. By comparing Δ with σ_v^2, we may get an idea of how good the covariates of the model are in a particular situation.

Let $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ be the sample mean of y_{ij} within area i, where n_i is the within-area sample size, which is an unbiased direct estimator of θ_i conditional on v_i. The BLUP of θ_i is given by

$$\hat{\theta}_i = \bar{X}_i^T \beta + \gamma_i(\bar{y}_i - \bar{x}_i^T \beta) = \bar{X}_i^T \beta + \gamma_i(v_i + \bar{e}_i)$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 / n_i)$ when β, σ_v^2 and σ_e^2 are all known. In the special case of $n_1 = \Lambda = n_m$, let $\varphi = \sigma_e^2 / n_i$ be a constant for all the areas. We have

$$E[\frac{1}{m-1}\sum_i (\hat{\theta}_i - \bar{\hat{\theta}})^2 \mid \bar{X}_1, \mathrm{K}, \bar{X}_m] = \Delta + \gamma^2(\sigma_v^2 + \varphi) = \Delta + \gamma\sigma_v^2 < \Delta + \sigma_v^2 .$$

In other words, the BLUPs are under-dispersed compared to the θ_i's. Notice that, in the absence of auxiliary variables, i.e. $\Delta = 0$, the result reduces to that in Zhang (2003).

3. Simultaneous estimator

The empirical best linear unbiased predictor (EBLUP) $\hat{\theta}_i$ can be written as

$$\hat{\theta}_i = \bar{X}_i^T \hat{\beta} + \hat{v}_i \qquad \text{where} \qquad \hat{v}_i = \hat{\gamma}_i(\bar{y}_i - \bar{x}_i^T \hat{\beta}) .$$

Loss of between-area variation is essentially due to over-shrinkage in estimation of v_i. This amounts to using too small shrinkage-factor $\hat{\gamma}_i$. We now consider two adjustments, depending on whether the distribution of v_i is fully specified or not.

In the first place, we might assume a fully parametric distribution of v_i, denoted by $G(v;\xi)$ with parameters ξ. Let $\{v_{(i)}\}$ be the order statistics of $\{v_i\}$, for $i = 1,...,m.$ Let

$$\eta_i = E[v_{(i)};\xi]$$

be the expectation of $v_{(i)}$. It follows that $\{\eta_i\}$ is the best ensemble predictor of $\{v_i\}$. Let $\hat{\xi}$ be an estimator of ξ and $\hat{\eta} = E[v_{(i)};\hat{\xi}]$, then $\{\hat{\eta}_i\}$ is the estimated best ensemble predictor. Instead of using $\hat{v}_i$ directly, we obtain the rank $\hat{v}_i$ of among all the $\hat{v}_i$'s, denoted by $r_i = rank(\hat{v}_i)$. The simultaneous estimator of the random effect of area i is then given by

$$\tilde{v}_i = \hat{\eta}_{r_i}$$

and the simultaneous estimator of θ_i is

$$\tilde{\theta}_i = \overline{X}_i^T \beta + \tilde{v}_i.$$

In the case when all the parameters are known, we have, as $m \to \infty$,

$$E[\frac{1}{m-1}\sum_i (\tilde{\theta}_i - \bar{\tilde{\theta}})^2 \mid \overline{X}_1, \ldots, \overline{X}_m] \to \Delta + \sigma_v^2 = E[\frac{1}{m-1}\sum_i (\theta_i - \bar{\theta})^2 \mid \overline{X}_1, \ldots, \overline{X}_m]$$

provided $\frac{1}{m-1}\sum_i (\overline{X}_i - \overline{X})^T \beta \eta_{r_i} = 0$, because $\frac{1}{m-1}\sum_i (\eta_i - \overline{\eta})^2 \to \sigma_v^2$ as $m \to \infty$.

In many situations, however, it may be too difficult or restrictive to fully specify the distribution of v_i. Let $\tau_{\hat{v}}^2 = (m-1)^{-1}\sum_i (\hat{v}_i - \hat{\bar{v}})^2$ be the empirical variance of the EBLUP $\hat{v}_i$'s. We observe over-shrinkage of the EBLUPs if

$$\tau_{\hat{v}}^2 < \hat{\sigma}_v^2 .$$

A simple nonparametric simultaneous estimator can be given as

$$\tilde{v}_i = \hat{\bar{v}} + (\hat{v}_i - \hat{\bar{v}})\hat{\sigma}_v / \tau_{\hat{v}} \qquad \text{and} \qquad \tilde{\theta}_i = \overline{X}_i^T \beta + \tilde{v}_i. \tag{3.1}$$

Notice that, in this way, the empirical variance of the $\tilde{\theta}_i$'s always equals to $\hat{\sigma}_v^2$.

To evaluate the MSE (or variance) of $\tilde{\theta}_i$, we use a bootstrap procedure. Firstly, we fix the parameters of the model at the estimated values. Secondly, we generate a bootstrap sample in area i:

1) let $\theta_i^* = \overline{X}_i^T \hat{\beta} + v_i^*$, where v_i^* is drawn randomly and with replacement from $\{\tilde{v}_i\}$;
2) let $y_{ij}^* = x_{ij}^T \hat{\beta} + v_i^* + e_{ij}^*$, where e_{ij}^* is drawn randomly and with replacement from $\{\tilde{e}_{ij}\}$, and $\tilde{e}_{ij} = \hat{e}_{ij}\hat{\sigma}_e / \tau_{\hat{e}}$, and $\hat{e}_{ij} = y_{ij} - x_{ij}^T \hat{\beta} - \hat{v}_i$, and $\tau_{\hat{e}}^2$ is the empirical variance of the $\hat{e}_{ij}$'s.

Finally, based on a bootstrap sample $\{y_{ij}^*\}$, we re-estimate the model (2.1) and derive the simultaneous estimates in the same way as based on the original sample, denoted by $\tilde{\theta}_i^*$. A bootstrap replicate of the error in the original simultaneous estimator is given by $\tilde{\theta}_i^* - \theta_i^*$. Independent bootstrap replicates can then be used to produce Monte Carlo approximation to the bootstrap MSE (or variance).

4. Simulations

4.1. Data

The Norwegian Wage Survey (NWS) is based on a yearly sample of clusters of wage earners. The clusters correspond to establishments enlisted in the Establishment Register, stratified according to the size of the establishment. The NWS includes all the employees from each selected establishment. The primary variable of interest is the monthly wage, classified by sex, age, education, type of position, and so on. For our simulations, we use the sample from industry group 52 (retailing) and occupation group 5 (sales, service) in 2000, 2001 and 2002. We use the municipalities as small areas, and estimate the average monthly wage in each municipality.

4.2. Case without auxiliary information

In this case model (2.1) contains only an intercept, denoted by μ. We fit the model separately for men and women in all the 3 years. Estimates of the model parameters are given in Table 1. As expected, the estimated overall average monthly wage, i.e. $\hat{\mu}$, increases from 2000 to 2002, and is higher for men than for women. The estimated variance components vary from one year to another, with $\hat{\sigma}_v^2$ having the largest variation. The estimated residuals are far from normal. Student-t distribution, on the other hand, appears to fit the estimated residuals quite well, albeit after deletion of a few largest and/or smallest values. In the simulations below, we shall consider only the nonparametric version of the simultaneous estimates.

We now set up the model parameters for simulation based on the sample in 2001. We use the mean and variance of the observed area sample means as the true μ and σ_v^2. Whereas we use the variance of the observed within-area deviations, i.e. $e_{ij} = y_{ij} - \bar{y}_i$, calculated across all the areas as the true σ_e^2. We draw a simulated sample in two steps:

a) draw θ_i^* randomly and with replacement from all the observed area sample means;

b) draw e_{ij}^* randomly and with replacement from $\{e_{ij}\}$ and set $y_{ij}^* = \theta_i^* + e_{ij}^*$, for all (i, j).

Table 1. Estimated model parameters in 2000, 2001 and 2002. (Data: NWS)

Men					
Year	Sample size	m	$\hat{\mu}$	$\hat{\sigma}_v$	$\hat{\sigma}_e$
2002	6062	316	20954	573.4	4601.5
2001	6353	305	19768	780.1	4589.3
2000	5444	303	19318	623.9	3916.6

Women					
Year	Sample size	m	$\hat{\mu}$	$\hat{\sigma}_v$	$\hat{\sigma}_e$
2002	10424	365	19318	359.6	3494.1
2001	10659	269	18475	1025.0	4556.9
2000	9025	366	17552	503.9	2949.2

The within-area sample sizes are the same as the observed ones in 2001. Given each simulated sample, we estimate the model parameters, and derive the direct estimate, the EBLUP, and the simultaneous estimate of all the θ_i^* 's. The results for the parameter estimators based on 1000 simulated samples are given in Table 2.1. Both the estimator for μ and σ_e^2 seem to be unbiased. Whereas the estimator of σ_v^2 ("fitting-of-constants" method, Rao, 2003) appears to be slightly downward biased. In addition, $\hat{\sigma}_v$ has large (just below 30%) relative standard error. The estimation of the between-area variation is more demanding than the estimation of the within-area variation.

Table 2.1. Simulation results for parameter estimators, 1000 simulations.

Parameter	Men			Women		
	$\hat{\mu}$	$\hat{\sigma}_e$	$\hat{\sigma}_v$	$\hat{\mu}$	$\hat{\sigma}_e$	$\hat{\sigma}_v$
True value	19852	4500	2679	18287	4493	1515
Expectation	19856	4475	2592	18285	4474	1425
Relative standard error (%)	1.0	2.6	27.1	0.7	8.2	28.4

The three small area estimators are compared to each other with respect to (i) the average, and maximum, *absolute relative error (ARE)* given, respectively, by

$$m^{-1}\sum\nolimits_{i=1}^{m} |\,\hat{\theta}_i^* / \theta_i^* - 1\,| \qquad \text{and} \qquad \max_{i=1,\mathrm{K}\,,m} |\,\hat{\theta}_i^* / \theta_i^* - 1\,|;$$

(ii) the average *absolute relative distributional error (ARDE)* given by

$$m^{-1}\sum\nolimits_{i=1}^{m} |\,\hat{\theta}_{(i)}^* / \theta_{(i)}^* - 1\,|$$

where $\theta^*_{(i)}$ is the *i*th order statistic of $\{\theta^*_i\}$, and $\hat{\theta}^*_{(i)}$ is the *i*th order statistic of $\{\hat{\theta}^*_{(i)}\}$; and (iii) the *relative error (RE)* of the range estimator given by

$$(\max_i \hat{\theta}^*_i - \min_i \hat{\theta}^*_i)/(\max_i \theta^*_i - \min_i \theta^*_i) - 1.$$

Table 2.2. Simulation results for small area estimators, 1000 simulations

	Men			Women		
Estimator	Direct	EBLUP	Simultaneous	Direct	EBLUP	Simultaneous
Average ARE	8.6	5.7	6.4	6.8	3.9	4.4
Maximum ARE	89.7	42.0	46.5	116.4	39.2	45.8
Average ARDE	4.5	2.1	2.0	4.1	2.0	1.2
RE in range	47.9	-17.1	6.3	121.7	-27.6	6.2

Based on the results given in Table 2.2, we observe that, (I) on average, the model-based estimators improve the area-specific estimation compared to the direct estimator. They are also more robust since the maximum AREs are much smaller than that of the direct estimator. The simultaneous estimator is slightly worse than the EBLUP, without losing the essential gains of the modeling approach. (II) The model-based estimators are also much better than the direct estimator for estimation of the distribution of the small area means, both with respect to the average ARDE and the range. (III) Not unexpectedly, the simultaneous estimators have better ensemble properties than the EBLUPs. In the present simulation, the gains are substantial with respect to the range (and variance) of the small area means.

4.3. Case with auxiliary information

To create the case with auxiliary information, we take the joint sample of 2001 and 2002, which contains 8459 persons of both sex. We now treat the monthly wage in NWS 2001 as the known covariates, denoted by x_{ij}, and the monthly wage in NWS 2002 as the variable of interest. The parameters of the model (2.1) are fixed as follows. Firstly, we obtain the sample means $\bar{x}_i$ and $\bar{y}_i$. The ordinary least square fit of regressing $\bar{y}_i$ on $\bar{x}_i$ (including an intercept term) yields the regression coefficients for the simulations below, denoted by β. Whereas the residuals will be used as the area-level random effects, denoted by $v_i = \bar{y}_i - \bar{x}_i^T \beta$. Finally, we obtain the within-area deviations as $\varepsilon_{ij} = x_{ij} - \bar{x}_i$ and $e_{ij} = y_{ij} - \bar{y}_i$. For the particular data we used, we find

$$\Delta/(\Delta + \sigma_v^2) = 0.430,$$

such that the covariate accounts for about half of the variation in the variable of interest.

For each simulation, we generate the population and the sample as follows:

1) draw $\overline{X}_i^*$ randomly and with replacement from $\overline{x}_i$, and draw v_i^* randomly and with replacement from $\{v_i\}$, and set $\theta_i^* = \overline{X}_i^{*T} \beta + v_i^*$;
2) draw ε_{ij}^* randomly and with replacement from $\{\varepsilon_{ij}\}$, and set $x_{ij}^* = \overline{X}_i^* + \varepsilon_{ij}^*$;
3) draw e_{ij}^* randomly and with replacement from $\{e_{ij}\}$, and set $y_{ij}^* = x_{ij}^{*T} \beta + v_i^* + e_{ij}^*$.

The within-area sample sizes are the same as in the observed panel. Based on each simulated sample, we derive the parameter estimates, the direct estimates, the EBLUP and the simultaneous estimates of all the θ_i^*'s. The results based on 1000 simulations are given in Table 3.1 and 3.2.

The parameter estimators perform similarly as in the case without auxiliary information. The estimators of β and $\hat{\sigma}_e$ are apparently unbiased. Whereas $\hat{\sigma}_v$ appears to be slightly downward biased, and is associated with the largest uncertainty.

Table 3.1. Simulation results for parameter estimators with auxiliary information, 1000 simulations.

	$\hat{\beta}$	$\hat{\sigma}_e$	$\hat{\sigma}_v$
True value	(8613.9, 0.603)	3713.9	1195.2
Expectation	(8623.4, 0.602)	3705.0	1157.6
Relative standard error (%)	(2.6, 1.8)	2.1	29.5

Table 3.2. Simulation results for small area estimators with auxiliary information, 1000 simulations.

Estimator	Direct	EBLUP	Simultaneous
Average ARE	7.2	3.1	3.6
Maximum ARE	73.7	62.8	55.8
Average ARDE	4.3	1.1	0.9
RE in range	53.0	-23.6	-8.8

Next, we compare the three estimators with respect to the average and maximum ARE, the average ARDE and the RE in range (Table 3.2). The conclusions are similar to those in the case without auxiliary information: (i) the model-based estimators improve the direct estimator both in terms of area-specific and ensemble properties, and (ii) the simultaneous estimator improves the ensemble properties of the EBLUP, without losing much of the precision of the EBLUP in area-specific prediction. Notice that the nonparametric simultaneous estimator can be expected to give good estimation of the variance of the small area means, since it is based on an adjustment of the empirical variance of the EBLUPs. The simulation results above suggest that this typically also leads to better estimation of the other ensemble statistics such as the range.

5. Summary

We considered a nested error regression model, which is a very basic model for small area estimation. We showed, in theory as well as by empirical example, that the (empirical) best linear unbiased prediction entails loss of the between-area variation of the small area means. In general, estimators that are optimal for area-specific prediction may have unsatisfactory ensemble properties. We extend the simultaneous estimation approach of Zhang (2003) for the nested error regression model. This allows us to make use of auxiliary information at the unit-level when it is available. Simulations suggest that the simultaneous estimator may substantially improve the estimation of the ensemble characteristics of the small area parameters, without losing much of the precision in area-specific prediction. Our approach provides a frequentist alternative to the existing Bayesian methods.

REFERENCES

BATTESE, G.E., HARTER, R.M. and FULLER, W.A. (1988). An error component model for prediction of County Crop Areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28—36.

JUDKINS, D.R. and LIU, J. (2000). Correcting the bias in the range of a statistic across small areas. *Journal of Official Statistics*, 16, 1—13.

GHOSH, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, 533—540.

LAHIRI, P. (1990). Adjusted Bayes and empirical Bayes estimation in finite population sampling. *Sankhya B*, 42, 50—66.

LOUIS, T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393—398.

RAO, J.N.K. (2003). *Small Area Estimation*. Wiley.

SHEN, W. and LOUIS, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B*, 60, 455—471.

SPJØTVOLL, E. and THOMSEN, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, 2, 435—449.

ZHANG, L.-C. (2003). Simultaneous estimation of the mean of a binary variable from a large number of small areas. *Journal of Official Statistics*, 19, 253—263.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 667—684

SMALL AREA ESTIMATION OF DISABILITY IN AUSTRALIA

Daniel Elazar[1]

ABSTRACT

In Australia, as in many countries, there has been a rapidly growing demand from policy makers in both regional and national jurisdictions for social and economic small area data to satisfy expanding decision-making requirements. To date, the Australian Bureau of Statistics (ABS) has attempted to meet this demand using simple synthetic estimation, but occasionally using more sophisticated small area models. The increasing user demand for small area estimates, together with practical difficulties in increasing survey sample sizes, has motivated the need to identify ways of finding reliable and defensible methods for producing quality small area estimates.

A project has commenced at the ABS to produce a series of manuals on the theory, application and processes for producing small area statistics in the Australian context. As part of this project, an empirical study has commenced into alternative approaches for producing small area estimates of disability. This builds on ABS experience in producing synthetic estimates from previous surveys of disability. The empirical study will assist in writing the small area estimation manual by developing practical knowledge and understanding of available small area methods with the ABS. In addition to this, we hope the results will go some way towards satisfying user demand for small area data on this topic. The main purpose of this paper is to discuss intended approaches for an application of existing small area methods to the topic of disability. As such the paper does not introduce any new methodological approaches to small area estimation. This paper firstly discusses the context of small area estimation in Australia. We then canvas the nature of the statistical problem we face and the advantages and disadvantages of the response and auxiliary variable data we have available for modelling small area estimates of disability. The paper then details the various small area models we propose to use. We consider both hierarchical Bayes and frequentist methods for estimating the proposed small area models.

[1] Assistant Director, Analytical Services Branch, Methodology Division, Australian Bureau of Statistics, ABS House, 45 Benjamin Way, Belconnen ACT 2617; daniel.elazar@abs.gov.au

1. The Australian Context

Over the past fifteen years there has been a rapidly growing demand for small area estimates in Australia. There are a variety of socio-political factors that have also stimulated this growth. Firstly, this growth in demand has largely coincided with an increased emphasis on evidence-based decision-making by government. Government agencies are now subject to greater accountability in providing a more efficient, effective and coordinated approach to the delivery of program services to regions with greatest need. Secondly, local governments are taking a more pro-active role in the economic and social development of their jurisdictions. Thirdly, there has been an increased focus of government policy making, both at national and state levels, on addressing the increased levels of economic and social disadvantage faced by communities residing in outer regional and remote areas of Australia. While these areas are often geographically very large, the populations they contain are usually very small. A significant proportion of Australia’s indigenous people, who are the subject of a number of health, cultural, community development and housing programs, also reside in these remote areas. Fourthly, non-government organization service-providers increasingly require data for funding submissions and planning. Fifthly, there is a rapidly growing statistical sophistication among economists in the use of more complex models, that combine both micro and macro economic dynamics, in forecasting economic trends and relationships. In the statistical realm, a flourish of new small area estimation methods, combined with unprecedented increases in computing capabilities, have meant that small area problems that were once intractable are becoming feasible.

ABS survey collections are designed to produce reliable estimates only at broad geographic levels such as at national and state levels. Practical issues have meant that there is little prospect of increasing sample allocations at the regional level in order to produce small area estimates of useful quality. The most feasible option, therefore, for satisfying the demand for small area data is the development of appropriate small area techniques to make use of existing survey data sources, along with suitable auxiliary data sources, including those available from other government or private agencies.

A number of small area estimation projects have been undertaken in the ABS over the last decade. These include a study of the estimation of labour force status for chosen small area regions (Bell and Carolan, 1998) using a time series model that takes account of autocorrelation between sample overlap groups. Several projects have also been undertaken into producing small area estimates of disability using census data.

The ideal approach to small area estimation, in the context of a national statistical service, would be to meet the following goals: meet decision-making objectives of users; cost-effectiveness; produce output of sufficient reliability for intended uses; be appropriate to the context of the problem; involve defensible methodologies; and provide results that can be readily interpreted and explained to users. In practice it may be difficult to meet all of these goals simultaneously, so that judgement needs to be exercised in determining the relative priorities of these objectives.

In response to this need, the ABS is developing a manual (or more likely a suite of manuals) on small area estimation methods and processes that will present a consistent approach to be used in the ABS, as a step towards assuring a standard quality of small area output. The manuals will be tailored to areas of the ABS involved in the production of small area estimates, whether they are involved in servicing client requests, implementing complex methodologies or validating, clearing and releasing output data.

The goals of the Small Area Estimation Practice Manuals Project, as it is known, are to increase the ABS' capability in satisfying the growing demand for small area statistics and to standardize and focus the ABS' approach to meeting this need. The manuals will help bridge the gap between the theoretical knowledge of small area estimation techniques and the practical application of such techniques. The manuals are intended to not only be a practical and methodological resource on how to go about producing small area data, but also a repository for capturing the growing experience of small area methods and processes. Within the context of a national statistical office, these goals are constrained by cost considerations, ease of implementation and interpretability of models and output for both statisticians and clients.

Brackstone (2002) gives a very good account of the issues facing national statistical offices in producing small area statistics. Pfeffermann (2002a) gives a concise summary of the key issues being confronted in small area estimation. Pfeffermann (2002b) and Trewin (1999) give an outline of the future directions for the application of small area estimation methods in the production of official statistics. They conclude that the use of sophisticated model based methods is inevitable in producing reliable small area estimates.

2. Empirical Study of Disability

An empirical study into small area estimates of disability is currently being undertaken as part of the Small Area Estimation Practice Manuals Project. The main purpose of this empirical study is to develop within the ABS, knowledge and understanding of the implementation and effectiveness of small area techniques. The empirical study will assist in providing answers to the following questions:

1. How much gain in quality can be achieved from using more sophisticated small area techniques over simpler ones (for example, Poisson generalized linear mixed models over linear or synthetic methods)?
2. What contribution does the quality of auxiliary data sources have towards the overall quality of small area estimates and what is the minimum level of

quality for auxiliary data? By quality we mean not just the accuracy of the auxiliary data but also it's relevance to and correlation with the response variables.

3. What is the relative efficiency gain of using a unit (person) level model compared with a corresponding area level model?
4. At how fine a level can viable small area estimates be produced before suffering from model breakdown? By fine we mean either geographical size or the cross-classification of small area estimates by other variables such as severity of disability, age and sex.
5. What are the best approaches to validating output small area estimates in practice? How can user knowledge or preconceptions be best utilised in validating small area output? For example, disability administrators will be able to identify the districts in which the demand for services out-strips supply and whether the small area statistics reflect this. Disability administrators will also be useful in assessing the validity of small areas with extreme values.
6. What are the most efficient and appropriate measures of accuracy for modelled small area estimates (Trewin, 1999)? Are reliable and comprehensive measures available that are similar in effectiveness to those used for measuring sampling error and bias of direct sample estimates?

3. Data Available for the Empirical Study of Disability

3.1. Survey of Disability, Ageing and Carers

The main source of disability data to be used for this empirical study comes from the 1998 Survey of Disability, Ageing and Carers (SDAC), conducted by the ABS. This will be the source of response variable data in the modelling of small area estimates.

Data item concepts and definitions used in SDAC follow those of the World Health Organisation (WHO) and other relevant international health protocols, to ensure statistical consistency. These standards address the complex issues of what is a disability, how best to determine whether a person has a disability or not and how should disabilities be categorised by type and severity. Precise concepts and definitions are pivotal in obtaining accurate measurements of the incidence of disability by type and severity. According to the International Classification of Functioning, Disability and Health, "disability is an umbrella term for impairments, activity limitations and participation restrictions. It represents the negative aspects of the interaction between an individual (with a health condition) and that individual's contextual factors (environmental and personal factors)." (World Health Organisation, 2001). In SDAC, a person has a disability if s/he has a limitation, restriction or impairment, which has lasted or is likely to last at least six months and restricts a range of everyday activities (SDAC, 1998). A person has a *restriction* if s/he has difficulty performing, or can't perform, a particular activity, needs assistance from another person or uses aids. An *impairment* is defined as any loss or abnormality of psychological, physiological or anatomical structure. Examples are loss of sight or a limb, disfigurement or deformity, hallucinations, loss of consciousness or any other lack of function of body organs. It is important to note that in SDAC, disability covers all everyday activities regardless of whether they relate to employment, household or social activities.

For statistical publication purposes, impairments are grouped into broad types, these being physical, psychological/psychiatric, intellectual, sensory and head injury/brain damage. Users of small area disability data are mainly interested in current estimates of disability by impairment type rather than disability per se, because demand for and delivery of services often relates to the specific type of impairment (for example, persons with a visual impairment may require assistance in the use of visual aids which is different to a person with a psychiatric impairment who may require assistance in managing their medical condition.) For this reason, estimates classified by impairment type are seen as the main objective of our small area estimation.

In this empirical study, we focus only on primary impairment type in the case of multiple impairments, for reasons of simplicity. Each person with one or more restricting impairments is therefore categorised to one and only one impairment type. Additional strength can be given to small area estimates by taking advantage of correlations between different response variables such as impairment type. These correlations, observed in the survey data, are determined by both the stochastic process that governs the population distribution of impairment types and non-sampling survey errors in coding persons to an impairment type. For example, brain damage may be miscoded to the Intellectual impairment type and vice versa. Due to the difficulty in trying to separate out these two influences we will only take account of the overall observed correlations between response variables.

3.1.1. SDAC Sample Design

There are two components of the SDAC sample: the household component and the cared accommodation component. The cared accommodation component covers specific non-private dwellings such as homes for the aged and retirement homes. *Only the household component is considered for this study* as user interest is in the provision of disability services to private households. The sample from the household component was 37,000 persons from an in-scope population of around 18 million.

SDAC is a multi-stage household survey, which at first stage selects a sample of census collectors' districts (CD's) with probability proportional to size (PPS). Each selected CD is then formed into blocks of approximately equal size, based on permanent landmarks such as roads, footpaths and rivers. Within each selected CD, a block is then selected at the second stage, again PPS. In the third and final stage, a systematic sample of dwellings (referred to as a cluster) is selected throughout the selected block.

All in-scope persons (broadly, permanent residents aged over 15 not in the permanent defence forces) are surveyed from each selected dwelling. SDAC uses the "any responsible adult" (ARA) methodology for initial demographic details and disability status, whereby the person who answers the door is asked to respond for all other in-scope persons in the dwelling. Personal interview is used for persons with a disability, carers and people aged 65 years or more. Due to high enumeration costs, remote and very remote areas of Australia are out of scope of the survey, making it difficult to extrapolate results to these areas.

3.1.2. Choice of Target Small Areas

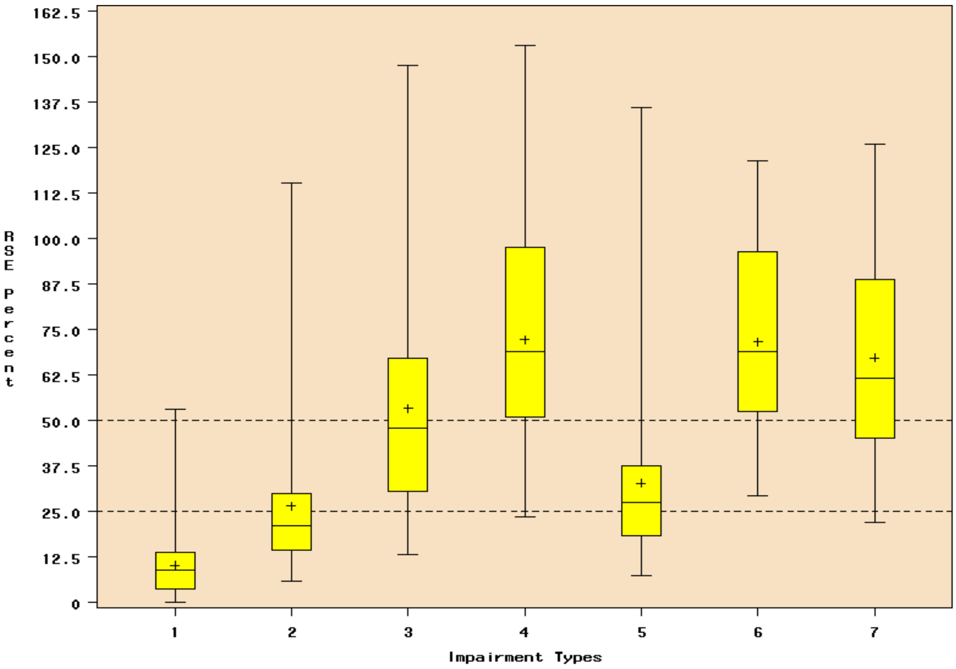
The data published from SDAC includes estimates for each impairment type by state, level of severity, age and sex for each state and territory of Australia. For some of the states and territories this data is subject to rather large sampling error. The sampling error in smaller geographic regions will therefore be too large for the direct estimates to be useful.

For this study, the target small area is the Statistical Sub-Division (SSD), which is a classification under the Australian Standard Geographical Classification (ASGC). Ideally we would like to use Disability and Health Services (DHS) regions, which are utilised by users of disability data for administrative purposes. However at the time of writing, definitions of these areas were still being sought. It is believed that DHS regions are similar to SSD’s.

There are 200 SSD’s covering all of Australia, of which 183 contain persons selected in SDAC. The population sizes of SSD’s vary considerably depending upon the size of the state or territory. The average population size across SSD’s is 100,000 persons, however these vary from between a few thousand to just over 800,000. SSD sample sizes vary from between 5 and 1700 person selections. In practice the smaller SSD’s may need to be collapsed to ensure a small area comprises a population of at least 20,000. This figure has been given as a minimum size rule of thumb for modelling small area estimates.

SSD’s have been chosen as the small area, firstly because there is an established user demand for data at a geographic level close to this. A second reason is that SSD’s are large enough that most have responses in the SDAC, which makes it feasible to apply the more complex methods of small area estimation. Previous work on small area estimation of disability used the Statistical Local Area (SLA) as the target small area, which is substantially smaller than SSD. We plan to repeat the current small area estimation study at the SLA level for comparison purposes with the previous work. It is still possible to provide synthetic predictions for small areas that are too small, however there would be insufficient data to perform an evaluation of the performance of those predictions.

Figure 1. below shows boxplots of the distribution of estimated relative standard errors (RSE’s) across SSD’s for the estimated number of people with various types of impairment. Horizontal grid lines have been drawn at 25 percent as an estimate with RSE greater than this is not considered sufficiently reliable according to ABS publication standards. The 50% grid reference indicates the threshold above which estimates are not statistically significant.



Impairment types:	1 = No impairment	2 = Any impairment	3 = Sensory
	4 = Intellectual	5 = Physical	6 = Psychiatric
	7 = Head injury/brain injury		

Figure 1. Boxplots of Small Area RSE's by Type of Impairment

From Figure 1, for the any impairment variable, just under 40% of small areas have an RSE of more than 25%. This is not too dissimilar to physical impairment, the most prevalent type, where just over 50% of small areas have an RSE of more than 25%. In the case of the sensory, intellectual, psychiatric and head injury/brain damage impairment types, very few SSD’s (if any) have RSE’s less than 25%. This strongly illustrates the need for small area estimation models to provide reliable estimates for all small areas.

Small areas with an undefined RSE due to a zero estimate (and consequently a zero standard error), have been excluded from the boxplots. Care needs to be taken in interpreting the estimated variances of small areas. A zero variance does not necessarily imply a high degree of accuracy for the estimate. Quite the opposite may be true where a small sample size, combined with the rarity of most impairment types, means that no persons with the impairment in question have been observed in the random sample within a small area. This problem will need to be rectified through smoothing before using these sampling variances as sampling error terms in the small area estimation models proposed in section 4.

3.2. Auxiliary Data

A number of auxiliary data sources have been obtained in order to help improve the efficiency of modelled small area estimates. All of these auxiliary data sources relate to formal care and assistance provided through a program funded by a government agency. SDAC records all persons with an impairment regardless of whether the person receives formal, informal or no support and assistance. A substantial proportion of people with a profound or severe impairment are informally cared for by a family member or friend, and therefore receive either no or infrequent formal care and assistance. It is also quite common for formal care services to be obtained from a privately run provider, especially when compensation monies have been paid and the recipient is therefore ineligible for government assistance. In addition to this there are those people with moderate or mild levels of severity that are unlikely to use any care or assistance and hence fall outside the coverage of government care programs. Finding auxiliary data that relates directly to informal care provision is by its very nature quite difficult. The main chance of success in obtaining additional strength from available auxiliary data, is if total disability levels (formal and informal) are directly related to the frequency of government assisted care. Preliminary data analyses indicate that this is not the case.

Due to a concern regarding the disclosure of personal information, government departmental privacy requirements have prevented us obtaining unit level auxiliary data. What we have obtained are counts of persons by impairment type at the small area level, in some cases classified by other variables such age and sex. Even if person level auxiliary data were obtained, the matching of person level data between the files would be problematic due to concerns over confidentiality.

Nonetheless, for some auxiliary items, there are corresponding variables on SDAC that could be used as substitutes for unit level auxiliary data, perhaps after some adjustment. While additional sample based auxiliary data may help improve the fit of the small area model to the data, it won’t be of much assistance when model based small estimates need to be predicted from non-sampled population units. For this reason we rely more heavily on non-sample based auxiliary data.

3.2.1. Data on Disability Service Provision

The first auxiliary data source we obtained were tables derived from the Commonwealth State/Territory Disability Agreement — Minimum Dataset

(CSTDA). This dataset contains information on persons receiving a disability service from a care provider funded through one or more state and territory or Commonwealth disability programs. Disability services include in-home support, accommodation, respite care, community support / access and employment. The CSTDA dataset was established through a multilateral government agreement in order to determine the allocation of government money for disability programs as well as coordinate activities between support agencies and focus service provision to regions of greatest need.

Even though the CSTDA dataset is at a fine level (small area by impairment type by severity by age by sex), the number of persons on this dataset is only 62,000 compared with an estimated 1.1 million people with a profound or severe restriction according to SDAC. This discrepancy appears to be largely due to the high levels of people with a severe or profound restriction receiving either informal or privately purchased care and assistance, and that the data are for persons receiving a support service on a single snapshot day. This means that the data is biased towards people with profound levels of certain impairment types, who require the most frequent assistance. In addition there are programs run by other government agencies that are not covered by the CSTDA. The CSTDA data targets people in the 0—64 age range, however it will include people aged 65 or over who have aged while continuing to receive a service. The CSTDA data also does not generally cover persons in most remote areas.

3.2.2. Disability Support Pension Data

Disability Support Pension (DSP) data was obtained at a small area level dissected only by disabling condition, as this was the closest variable available to impairment type. It is important to note that disabling condition is a different concept to restricting impairment and as a result, there is no well-defined correspondence between the two. Another conceptual difference between SDAC and the DSP is that people are assessed for entitlement to the disability pension on the basis of whether they can undertake employment activities. Ability to undertake domestic and social activities is not necessarily taken into consideration as it is in SDAC. DSP data does not cover the over 65 well, as people turning this age are moved onto the aged pension. In terms of total numbers, DSP has around 600,000 pension recipients compared with the 655,000 people with a profound or severe restriction on SDAC. There are also eligibility requirements such as means testing that applicants must satisfy before receiving the pension.

3.2.3. Population estimates by age and sex

As auxiliary variables, small area age by sex demographic counts are expected to be a good “safety net” to make up for any deficiencies, as explained above, in the explanatory power of the CSTDA and DSP data. Age-by-sex demographic counts are considered to be quite reasonable predictors of disability statuses in their own right, and especially for sub-populations such as the over

65's. This variable is also available at the person level from SDAC and could be used in a unit level model.

3.2.4. Remoteness

A measure of remoteness is also being included in the small area estimation models firstly as it is expected to be a potential indicator of the broad geographic distribution of disability. Remoteness categories are obtained from the Australian Standard Geographical Classification (ASGC) published by the ABS. There is anecdotal evidence that people incurring a serious disability in more remote areas of Australia tend to move to less remote areas where disability services are readily available or more extensive. This is not the case, however, for the Indigenous population, which mainly affects the Northern Territory. Users of disability small area data are likely to require estimates for remote areas, and as SDAC doesn't cover these areas, it seems sensible to build into the small areas models a functionality that would make this extrapolation possible. We have the option of including remoteness either as a parameter in the model or as a level for measuring random effects. We hope to determine from the empirical study, which option is more efficient.

3.2.5. Socio-economic Index

The ABS produced socio-economic indexes for areas based on the 1996 Population Census. These indexes can be readily derived for the small areas in this empirical study. The indexes for occupation and social disadvantage will be included in the model as there is considerable evidence of a relationship between health condition and socio-economic status (Mellor and Milyo, 2002 and Brodie-Reed et al., 2002). Social disadvantage is likely to be linked to lifestyle choices, which in turn predispose a person towards acquiring a disability. It's also plausible that some occupations are inherently more prone to industrial accidents than others.

4. Small Area Models

In this section we discuss the motivation behind choosing the range of small models we intend applying to the disability empirical study. One of the main objectives behind this empirical study is to apply a range of appropriate models, ranging from simple to more sophisticated, to assess which performs best in terms of accuracy, ease of implementation, production costs and model interpretability. In a production context, the choice of model will depend intrinsically upon client quality requirements and the availability and quality of auxiliary data. The understanding gained from this empirical study will be a starting point for applications to other small area problems.

The models we will be using for this empirical study will all be extensions of basic the Fay-Herriot model. Models of the Fay-Herriot form are widely used in the small area estimation literature. One reason for using Fay-Herriot models is that they incorporate, as a special case, synthetic estimation, which has been used in previous small area estimation work on disability at the ABS. It is therefore convenient that synthetic model estimates can be easily obtained from the Fay-Herriot model by removing the random effects term (Rao, 2003). Another attractive feature is that the Best Linear Unbiased Predictor (BLUP) under the Fay-Herriot model can be shown to take the form of a composite estimator (Pfeffermann, 2002a). The composite estimator is a weighted average of the direct survey estimator and a synthetic estimate based on a generalised linear model fitted to the observed data at a broader area.

In all the small area models proposed we consider the multivariate case wherever possible, subject to model identifiability. There are a number of response variables we wish to predict at the small area level, these mainly being impairment type (intellectual, physical, psychological/psychiatric, sensory, and head injury/brain damage) by level of severity (profound, severe, moderate, mild). It is important to take account of the variance covariance structure of these categories in order to improve model efficiency via multivariate models. A secondary reason is that the model fitting process is simpler than fitting separate models for each category.

There are three key dimensions to the small area models we wish to consider for the disability empirical study, which together form a framework. These dimensions are linear versus nonlinear, area versus unit level and mixed effects versus synthetic. We wish to consider a range of models across these dimensions to identify the best trade-off in competing models between simplicity, reliability, accuracy and interpretability. Previous work done in producing small area estimates of disability at the ABS, have used unit level models so one of the aims of the empirical study is to identify how much gain or loss of efficiency is afforded by unit as opposed to area level models.

We present below a range of models that cover the more promising elements of this framework but not necessarily all.

4.1. Model A: Linear, Area Level, Multivariate Fay-Herriot Model

The first small area model we consider is a linear, multivariate Fay-Herriot model at small area level, with mixed effects, as referred to in Rao, 2003. In a theoretical sense, a linear model is not appropriate when modeling rare count data, however as mentioned earlier, Model A will be applied as a benchmark for comparisons with other more sophisticated models. In the practical context of producing official statistics, a simpler model would be more appealing if we find it yields estimates almost as reliable as those of the more complex models.

Model A takes the following form. Let $\hat{\boldsymbol{\theta}}_{\mathbf{i}} = (\hat{\theta}_{i1}, ..., \hat{\theta}_{ir})^T$ be an $r \times 1$ vector of SDAC direct estimates of r impairment types by severity variables, for small area i . We assume that

$$\hat{\boldsymbol{\theta}}_{\mathbf{i}} = \boldsymbol{\theta}_{\mathbf{i}} + \mathbf{e}_{\mathbf{i}}, \quad i = 1,......,m \tag{1}$$

where $\boldsymbol{\theta}_{\mathbf{i}}$ is an $r \times 1$ vector of actual unknown counts of impairment types by severity, for small area i, and $\mathbf{e}_{\mathbf{i}} = (e_{i1},.......,e_{ir})^T$ is a $r \times 1$ vector of r sampling error terms for small area level i. The $\mathbf{e}_{\mathbf{i}}$ are assumed to be independently distributed with r-variate normal $N_r(\mathbf{0}, \boldsymbol{\Psi}_{\mathbf{i}})$, where the $\boldsymbol{\Psi}_{\mathbf{i}}$ are known variance covariance matrices (conditional on $\boldsymbol{\theta}_{\mathbf{i}}$) which can be calculated from the survey data.

$\boldsymbol{\theta}_{\mathbf{i}}$ is then modelled as

$$\boldsymbol{\theta}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{i}}, \quad i = 1,......,m \tag{2}$$

where $\mathbf{X}_{\mathbf{i}}$ is an $r \times rp$ matrix of p auxiliary variables, $\boldsymbol{\beta}$ is a $rp \times 1$ vector of regression coefficients and $\mathbf{v}_{\mathbf{i}}$ is the random effect term at small area level, independently distributed with multivariate normal $N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}})$.

Combining (1) and (2), Model A can be expressed as

$$\hat{\boldsymbol{\theta}}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{i}} + \mathbf{e}_{\mathbf{i}}, \quad i = 1,......,m \tag{3}$$

Model A incorporates both random effects $\mathbf{v}_{\mathbf{i}}$ and sampling effects $\mathbf{e}_{\mathbf{i}}$, each at the small area level. This is a worthwhile model feature as sampling error is known to vary considerably between different small areas, due to the variation in small area sample sizes and the level of clustering of disability in the small area population. In addition to this the actual unknown counts, $\boldsymbol{\theta}_{\mathbf{i}}$ are likely to vary between small areas even after taking account of the explanatory variables, thereby making the inclusion of the random effects term a desirable choice. Nevertheless, tests (such as that of Hausman) can be applied to determine the necessity of the random effects term, (Cameron & Trevedi (1998)).

Rao, 2003, points out that in situations where the small area direct survey estimates $\hat{\boldsymbol{\theta}}_{\mathbf{i}}$ are a non-linear function of survey estimates of total and the sample size for small area i is small, the assumption that $E_p(\mathbf{e}_{\mathbf{i}} \,|\, \boldsymbol{\theta}_{\mathbf{i}}) = \mathbf{0}$ in sampling model (1) will be invalid. This is the situation in our case as SDAC uses post-stratified ratio estimators of disability. Rao, 2003, gives a method for dealing with this problem, by replacing model (1) with a model for each estimate of total in the non-linear estimator, and then using hierarchical Bayes (HB) methods to deal with the mismatch between the sampling model (1) and linking model (2). An alternative, simpler approach might be to use the jackknife approach to correct $\hat{\boldsymbol{\theta}}_{\mathbf{i}}$ for most of the bias.

Model A can easily be extended to the unit level using unit level auxiliary data obtained from SDAC and small area level auxiliary data as contextual effects.

4.2. Model B: Poisson, Area Level, Multivariate Fay-Herriot Model

The main deficiency with Model A is that it involves fitting a linear model to rare count data, when a generalised linear model with an underlying Poisson distribution on the response variable would be more appropriate. The Poisson model is commonly regarded as the "benchmark model for count data" (Cameron & Trivedi, 1998). Like Model A, Model B is multivariate, is at the small area level and incorporates both fixed and random effects terms.

This model is taken from the generalised exponential family of models discussed by Ghosh et al. (1998). Using the same notation as in Model A, assume that the elements $\hat{\theta}_{ir}$ of $\hat{\boldsymbol{\theta}}_{\mathbf{i}}$ are distributed $Poisson(\mu_{ir})$ with mean parameters μ_{ir} obeying

$$\boldsymbol{\gamma}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{i}} + \mathbf{e}_{\mathbf{i}}$$

where $\boldsymbol{\gamma}_{\mathbf{i}} = (\ln(\mu_{i1}),.....,\ln(\mu_{ir}))^T$ and all other parameters are as specified for Model A.

Over dispersion, where the variance is greater than the mean, is a common problem when fitting a Poisson model. However the random effects term should adjust for this problem (citation).

The nature of Model B (multivariate GLMM) means that frequentist methods will not be viable and model identifiability will rely heavily on the success of the Hierarchical Bayes approach, for which we will be using the WinBUGS software. The SAS procedure *NLMIXED* won't be the primary software option as it can only be applied in the univariate case, however it may be useful for generating starting values and conditional priors for HB simulation runs and checking the convergence of conditional posteriors.

4.3. Model C: Logistic, Combined Unit/Area Level Fay-Herriot Model

Models A and B are at the area level. It would however be interesting to compare the efficiency of small area estimates produced from these models with that of a combined unit/area model. Due to privacy considerations, it has not been possible to acquire unit level auxiliary data from other agencies. Hence it is not possible to use a full unit level model unless we are willing to simulate auxiliary values at the unit level. Model C, therefore, makes use of the auxiliary and response variable data at the finest level of detail available: SDAC person level disability data for the response variable and small area level count data for auxiliary variables. Thus each person in a given small area will be associated with the Disability Support Pension (DSP) count which corresponds with their impairment type and the CSTDA count that corresponds with their impairment type, severity, age and sex.

A logistic transform with an underlying Bernoulli is used to take account of the binary nature of the disability response variables at person level. Like the previous models, Model C is multivariate with mixed effects.

We wish to model the r'th impairment type y_{ij}^{r} of person $j,\ j=1,......,n_i$ within small area i, $i=1,......,m$ and then use this model to predict impairment types for non-sampled units, thereby producing estimates of disability $\hat{\boldsymbol{\theta}}_{\mathbf{i}}$ (see section 4.1.3.1). We assume the y_{ij}^{r} to be independent $\text{Bernoulli}(p_{ij}^{r})$ variables with conditional probability density function:

$$f(y_{ij}^{r}=1\,|\,p_{ij}^{r})=p_{ij}^{r}$$

$$f(y_{ij}^{r}=0\,|\,p_{ij}^{r})=1-p_{ij}^{r}$$

Let $\boldsymbol{\lambda}_{\mathbf{ij}}=(\text{logit}(p_{ij}^{1}),.....,\text{logit}(p_{ij}^{r}))^{T}\quad i=1,.....,m\quad j=1,.....,n_i$, then $\boldsymbol{\lambda}_{\mathbf{ij}}$ can be modelled thus:

$$\boldsymbol{\lambda}_{\mathbf{ij}}=\mathbf{x}_{\mathbf{ij}}^{\mathbf{T}}\boldsymbol{\beta}+\mathbf{v}_{\mathbf{i}}+\mathbf{e}_{\mathbf{ij}},\quad i=1,.....,m\quad j=1,.....,n_i$$

where, as in previous models, the sampling errors $\mathbf{e}_{\mathbf{ij}}$ are independently distributed $N_r(\mathbf{0},\boldsymbol{\Psi}_{\mathbf{i}})$, $\boldsymbol{\Psi}_{\mathbf{i}}$ is known and the area level random effects $\mathbf{v}_{\mathbf{i}}$ are distributed $N_r(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{v}})$ independently of the $\mathbf{e}_{\mathbf{ij}}$.

Although the matrix of auxiliary variables, $\mathbf{x}_{\mathbf{ij}}$ is subscripted to indicate both person and small area level variables, the only person level variables it will include will be age and sex, as determined from the survey. All other covariates will be at the small area level.

4.3.1. Generating Small Area Estimates from Model C

Predictions of small area estimates need to be formed from the models that use unit level response variables. This can be done for each impairment type by summing unweighted response values for variable r, from the sample s_i in small area i and then adding to that the sum of the predicted proportions, $\tilde{p}_{ij}^{r}$ across the sample complement $s^{c}{}_{i}$. The $\tilde{p}_{ij}^{r}$ are predicted (Rao, 2003) by estimating $\boldsymbol{\beta}$ and generating a realisation of v_i^{r} from its underlying distribution. We then have:

$$\hat{y}_i^r = \sum_{j \in s_i} y_{ij}^r + \sum_{j \in s_i^c} \hat{p}_{ij}^r$$

where $\hat{y}_i^r$ is the predicted count estimate for the r 'th impairment type in the i 'th small area, y_{ij}^r is the sample response for the r 'th impairment type from the j 'th person in the in the i 'th small area.

An important issue is how do we ensure that the sum of the modelled count estimates across disability categories, $\sum_r \hat{y}_i^r$, agrees with the population benchmark for the i 'th small area.

5. Conclusions and Further Work

The disability empirical study will explore an assortment of small area estimation models to assess their relative performance against a range of criteria. These criteria are: the accuracy of modelled small area estimates, model robustness, ease of implementation in survey systems, production costs, model interpretability, and performance after validation. Knowledge gained from this empirical study will be used in producing a series of manuals for practitioners of official small area statistics.

A number of statistical issues need to be addressed in further work. One issue concerns how best to take account of design informativeness, (Pfeffermann and Sverchkov (1999), Pfeffermann et al. (2001) and Pfeffermann and Sverchkov (2003)). Another issue is that it would be quite desirable to incorporate a feature into the model that ensures additivity of predicted small area estimates to a broader region where sufficient sample sizes ensure reliable direct estimates (Pfeffermann and Bleuer (1993)). Models that take into account spatial autocorrelations of disability could also be explored akin to studies previously undertaken in the field of disease mapping (Wakefield and Elliot, (1999), Pascutto et al., (2000) and Elliot et al. (2000)). Pfeffermann (2002a), however discusses the results of a simulation study that shows that unless the correlations between small area random effects are very strong, the efficiency gains from a spatial model are not that appreciable.

There are a couple of changes in the data environment that, if implemented, would enhance the reliability of the small area model. The first is that including a disability related question in the next population census would provide an auxiliary variable that would strongly correlate with the response variable from SDAC. The second is that if the CSTDA data on disability service provision could be obtained that was longitudinal rather than single point in time, inherent biases towards those receiving services most frequently would be removed or at least greatly reduced.

For an official agency, it would be highly desirable to provide a reliable, quantitative measure of the accuracy of small area estimates, which would be an invaluable aid in informing users of the statistical reliability of such estimates (Trewin,1999). More systematic and defensible approaches to validating and ensuring the reliability of small area estimates, would also greatly aid official statistical agencies in embracing small area estimation methods.

Acknowledgement

I would like to acknowledge the cooperation and kind assistance from the Australian Institute of Health and Welfare in providing the ABS with tabulated data compiled from the Commonwealth State/Territory Disability Agreement — Minimum Datasets. I am also grateful to Centrelink and the Department of Family and Community Services for providing the Disability Support Pension data used in this study.

I would also like to thank Marion McEwin, Philip Bell and Ken Black for their comments on earlier drafts of the paper. Dr David Steel and Dr Borek Puza also gave invaluable suggestions on statistical approaches to undertaking the empirical study.

REFERENCES

AUSTRALIAN BUREAU OF STATISTICS (1998), Disability and Long Term Health Conditions, Australia, (Cat 4433.0).

BRACKSTONE, G. J. (2002), Strategies and Approaches for Small Area Statistics, *Survey Methodology*, 28(2), 117—123.

BRODIE-REED, I., BLACK, K., CHUBB, P., Ng, C. (2002), Socioeconomic Factors and Health (Data from the 2002 General Social Survey), Internal ABS Report.

CAMERON, A.C. and TRIVEDI, P.K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.

BELL, P. A. and CAROLAN, A. M. (1998), Trend Estimation for Small Areas from a Continuing Survey with Controlled Sample Overlap, Working Paper No. 98/1, ABS Cat No. 1351.0.

ELLIOT, P., WAKEFIELD, J. C., Best, N. G., and Briggs, D.J. (2000), *Spatial Epidemiology, Methods and Applications,* Oxford University Press.

GHOSH, M., NATARAJAN, K., STROUD, T.W.F. and CARLIN, B.P. (1998), Generalised Linear Models for Small-Area Estimation, *Journal of the American Statistical Association*, 93, 273—282.

MELLOR, J.M. and MILYO, J. (2002), Income Inequality and Health Status in the United States: Evidence from the Current Population Survey, *Journal of Human Resources*, 37(3), 510—539.

PASCUTTO, C., WAKEFIELD, J. C., BEST, N. G., RICHARDSON, S., BERNARDINELLI, L., STAINES, A. and ELLIOTT, P. (2000), Statistical Issues in the Analysis of Disease Mapping Data, *Statistics in Medicine*, 19, 2493—251.

PFEFFERMANN, D. and BLEUER, S.R. (1993), Robust Joint Modelling of Labour Force Series of Small Areas, *Survey Methodology*, 19, 149—163.

PFEFFERMANN, D. and SVERCHKOV, M. (1999), Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data, *Sankhya*, 61, 166—186.

PFEFFERMANN, D., MOURA, F. and SILVA, P. N. (2001) Multi-Level Modelling Under Informative Probability Sampling, 53rd session of the International Statistical Institute, Seoul, Korea, 2001.

PFEFFERMANN, D. (2002a), Small Area Estimation — New Developments and Directions, *International Statistical Review*, 70, 125—143.

PFEFFERMANN, D. (2002b), The Riga Conference: Summing Up Remarks, S*tatistics in Transition*, 4 (5), 799—801.

PFEFFERMANN, D. and SVERCHKOV, M. (2003), Fitting Generalised Linear Models under Informative Sampling, Chapter 12 of Analysis of Survey Data, Edited by R.L. Chambers and C.J. Skinner, Wiley, 175—195.

RAO, J.N.K. (2003), *Small Area Estimation*, Hoboken, NJ: Wiley.

TREWIN, D. (1999), Small Area Statistics Conference, *Survey Statistician*, 41, 8—9.

WAKEFIELD, J., and ELLIOTT, P. (1999), Issues in the Statistical Analysis of Small Area Health Data, S*tatistics in Medicine*, 18, 2377—2399.

World Health Organisation (2001*), International Classification of Functioning, Disability and Health*, Geneva <URL: http://www3.who.int/icf/icftemplate.cfm> last viewed 20 April 2004.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 685—695

APPLYING JACKKNIFE METHOD OF MEAN SQUARED PREDICTION ERROR ESTIMATION IN SAIPE

Tapabrata Maiti[1]

ABSTRACT

The Small Area Income and Poverty Estimation (SAIPE) project is an ongoing census bureau project to estimate numbers of poor school-age children by state, county and ultimately school district in the United States based upon Current Population Survey (CPS) and Internal Revenue Service (IRS) data, together with information from the latest decennial census. The current county-level methodology relies on a Fay-Herriot model fitted to log-counts (by county) of related school-age children in CPS-sampled households. Although census bureau produce transformed back bias corrected point estimates from a log-normal model, they don't produce measure of uncertainties of the estimates. The present paper discusses the measure of errors of the SAIPE estimates of county level child poverty rates.

Key words: Current Population Survey, empirical best predictor, jackknife, mean squared prediction error, mixed effects linear model, small area estimation.

1. Introduction

As summarized by Citro and Kalton (2000) and Bell (1997) the Small Area Income and Poverty Estimates (SAIPE) project at the Census Bureau has developed methods for estimating poverty and income statistics at the county, school districts, and state level using statistical models. The main objective of this program is to provide updated estimates of income and poverty statistics for the federal administration and federal funds allocation to local jurisdictions. On the basis of Census Bureau's estimates, the federal government allocated about $7 billion among the counties and states every year. In addition to these federal

[1] Department of Statistics, Iowa State University Ames, Iowa, USA. E-mail: taps@iastate.edu

programs, there are numerous local programs those use these income and poverty estimates for distributing their funds.

The SAIPE approach to county-level estimates was developed in response to legislation in 1994 (NRC Report of the National Academy of Sciences Panel of Estimates for Small Geographic Areas, Citro and Kalton 2000, p. 3) calling for the Census Bureau to supply 'updated estimates' of county-level child poverty for use in Title I allocations to counties in 1997-98 and 1998-99, and thereafter to provide estimates at school-district level. Prior to that time, decennial census data had been the source for such estimates (NRC Report 2000, p—16). "Updated estimates" were to be based on model using census data plus data from other sources. The 'other sources' which have been chosen for this purpose are the annual current population survey (CPS) and administrative-records data from IRS (income tax returns) and the Food Stamp program.

The CPS is the primary national survey measuring population and poverty each year, applying a rotating panel design to provide monthly data for clustered housing units sampled within a weighted probability sample of geographic units including about 1300 counties annually. It was chosen in the SAIPE program to provide the major indicators of national county level changes in child poverty, in the form of sample weighted estimates of numbers and proportion of poor children among children aged 5—17 related to primary householder (poor related school-age children). For improved stability of estimates, three years of CPS data (including that of the year before and the year after the income year of interest) are combined to provide the response variable. In order to describe the procedure more fully, let

$z_{iu} =$ CPS estimated total children in the 5—17 age-group for county *i* in the year *u*;

$q_{iu} =$ CPS estimated number of poor children in the 5—17 age-group for families in poverty for county *i* in the year *u*.

Then the CPS estimated poverty rate for county *i* in the year *u* is defined as

$$p_{iu} = q_{iu} / z_{iu} .$$

Suppose now a county i has CPS samples for all the three years t = 1, t and t+1. Then for this county, the direct estimate of the logarithm of the number of poor children in the 5—17 age group for the year t is defined as

$$y_{it}^{*} = \log\left\{\left(w_{i,t-1}^{*}\ p_{i,t-1} + w_{i,t}^{*}\ p_{i,t} + w_{i,t+1}^{*} p_{i,t+1}\right) \times \left(w_{i,t-1}^{*} z_{i,t-1} + w_{i,t}^{*} z_{i,t} + w_{i,t+1}^{*} z_{i,t+1}\right)\right\} \quad (1.1)$$

Here the weights $w_{i,u}^{*}\left(u = t-1, t, t+v\right)$ are proportional to the number of interviewed housing units in county *i* containing at least one child aged 5—17 in the year *u*. For counties with CPS samples in only one or two of the three years, the formula (1) is modified by taking values only for that year or a two-year average analogous to (1).

The objective of the SAIPE county model is to use decennial-census and administrative predictor variables to express the similarity of child-poverty data across counties, thereby 'borrowing strength' (Ghosh and Rao, 1994) from observed data to compensate for the absence of many counties from 3-year CPS samples and for the smallness of samples in many other countries. Since the outcome of modeling is to estimate numbers of related poor school-age children in all counties, SAIPE is constrained to use only administrative-record predictor variables, which are available in appropriately aggregated form for all counties nationally. Since the IRS and CPS data, along with other national records, are by law confidential at the individual level, the constraint of uniform national coverage is coupled to further constraints regarding strong agency controls on the manner of release of data. The most useful variables, which have been found to meet these constraints, are the county numbers of child exemptions for families in poverty and of all child exemptions reported on tax returns, along with county numbers of households participating in the food stamp program.

Many exploratory analyses of SAIPE data with alternative models have been performed over the last decade (NRC Report 2000a, chapter 5), in order to choose the best available model specification and small-area predictors from the variables derived from IRS county aggregated data, CPS sample data, and a long-form county aggregates from the most recent decennial census. The modeling framework chosen is that of Fay and Herriot (1979), as described in next section.

In either the log-number or log-rate form of the SAIPE Fay-Herriot (FH) model, the response variable cannot be computed in a county with sampled children when the sample contains no related poor school-age children. The data from such sampled counties are dropped when estimating model parameters. Then the number of poor children or the rates are obtained by exponentiating the FH model based estimates. In actual application the estimates are refined (bias corrected) using the properties of log normal distribution. SAIPE does not produce measure of error of these estimates. The Prasad-Rao (1990) type of mean squared error estimation technique is not directly applicable for this log transformed data. In this article we propose a simple jackknife based method to produce valid measure of error.

The paper is organized as follows. Section 2 defines the model on which the existing SAIPE small area estimation (SAE) methods are based along with estimation techniques. In Section 3, we propose the new method for county poverty rates and their measure of errors. The data analysis result is presented in Section 4 and the Section 5 draws the conclusion.

2. County level small area model in SAIPE.

The mixed effect linear models used and seriously considered in SAIPE are all of the following general Fay-Herriot (1979) type model (FH) form. For each county indexed by $i = 1, \mathrm{K} , m$, assume that sample sizes n_i and p-dimensional vectors $\mathbf{x}_i$ of predictor variables are known, and that response-variables satisfying

$$y_i = \theta_i + e_i, e_i \sim N\left(0, \frac{v_e}{n_i}\right) \qquad (2.1)$$

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, u_i \sim N\left(0, \sigma_u^2\right) \qquad (2.2)$$

are observed (whenever $n_i > 0$), where $\boldsymbol{\beta} \in R^p$ is a vector of unknown fixed-effect coefficients, and u_i, e_i are respectively county based random effects and sampling errors, independent of each other within and across countries. Ordinarily, σ_u^2 is unknown and estimated, while v_e is known. In SAIPE it also makes sense, via the 'bivariate model' using auxiliary decennial census data, to treat σ_u^2 as known and v_e as unknown, but we do not explicitly treat this possibility here.

Small area estimates (SAE's) based on such FH models are statistics designed to estimate with small mean squared error (MSE) the parameters

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \quad i = 1, \mathrm{K} , m.$$

In the SAIPE log-rate FH models, y_i is the observed log child-poverty rate for the *i*-th county, with the rate itself defined by exponentiating:

$$\theta_i^* = \exp(\theta_i) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta} + u_i\right) \qquad (2.3)$$

In the FH model, the estimators we consider for θ_i based on the data $\{y_i, n_i : n_i > 0,\ 1 \le i \le m\}$ above are the empirical Bayes estimators (c.f, Rao, 2003, chapter 9)

$$\hat{\theta}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i \left(y_i - x_i^T \hat{\boldsymbol{\beta}}\right) \qquad (2.4)$$

where $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2\right)$ are the maximum likelihood estimators in the models (2.1) and (2.2) and $\hat{\gamma}_i = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \dfrac{v_e}{n_i}}$. The estimator $\hat{\theta}_i$ is also known as the empirical 'best predictor' (EBP) of θ_i because it is obtained from the conditional distribution of θ_i given y_i without assuming a prior distribution on the model parameters. The EBP $\hat{\theta}_i$ is identical to the empirical best linear unbiased predictor (EBLUP) under normality as described in Prasad and Rao (1990).

Direct transformation back estimators for poverty rates are given by

$$\hat{\theta}_i^* = \exp\left(\hat{\theta}_i\right) = g\left(\hat{\theta}_i\right), (say) \qquad . \tag{2.5}$$

Using a simple Taylor expansion, one can approximate the mean squared prediction error of $\hat{\theta}_i^*$ as

$$MSPE\left(\hat{\theta}_i^*\right) \approx \{g'(\theta_i)\}^2 MSPE\left(\hat{\theta}_i\right) \tag{2.6}$$

where $MSPE(\theta_i)$ is the mean squared prediction error of $\hat{\theta}_i$ and is defined as $E\left(\hat{\theta}_i - \theta_i\right)^2$ and $g'(.)$ is the first derivative of $g(.)$. An estimator of $MSPE\left(\hat{\theta}_i^*\right)$ is then obtained as

$$mspe\left(\hat{\theta}_i^*\right) = \left\{g'\left(\hat{\theta}_i\right)\right\}^2 mspe\left(\hat{\theta}_i\right) \tag{2.7}$$

where $g'\left(\hat{\theta}_i\right)$ is the value of $g'(\theta_i)$ evaluated at $\theta_i = \hat{\theta}_i$ and $mspe\left(\hat{\theta}\right)$ is a suitable estimate of $MSPE\left(\hat{\theta}_i\right)$. The formula (2.7) was used by Jiang, Lahiri, Wan and Wu (2001) to data from the U.S. National Heath Interview Survey (NHIS).

The estimator (2.7) actually under estimate $MSPE\left(\hat{\theta}_i^*\right)$ as shown in the simulation study by Maiti and Slud (2002) and also commented by Rao (2003, P. 133). In fact, the Taylor approximation is not justifiable in this context because $\hat{\theta}_i - \theta_i$ is of order 0(1) for large m and hence the formula (2.7) is not second order correct.

3. Refined SAE and MSPE estimates for county level poverty rates

We will take the advantage of 'best predictor' and properties of log-normal distribution to refine the county poverty rates and their mean squared prediction error. The best predictor of θ_i^* is $E\left(\theta_i^* \middle| y_i, \boldsymbol{\beta}, \sigma_u^2\right) = H_i(\boldsymbol{\delta})$, say, where $\boldsymbol{\delta} = \left(\boldsymbol{\beta}^T, \sigma_u^2\right)^T$ is the vector of model parameters. Then the empirical best predictor for θ_i^* is $H_i\left(\hat{\boldsymbol{\delta}}\right)$ where $\hat{\boldsymbol{\delta}}$ is the consistent estimate of $\boldsymbol{\delta}$ obtained from the marginal distribution of y_i's. In particular, we have used maximum likelihood estimate of $\boldsymbol{\delta}$. In this application it is easy to check that

$$H_i(\boldsymbol{\delta}) = \exp\left\{\hat{\theta}_i^B + \frac{1}{2} g_i\left(\sigma_u^2\right)\right\}$$

where $\hat{\theta}_i^B = \mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right), \gamma_i = \dfrac{\sigma_u^2}{\sigma_u^2 + \dfrac{v_e}{n_i}}$ and

$g_i\left(\sigma_u^2\right) = \gamma_i \dfrac{v_e}{n_i}, \ i = 1, \mathrm{K}\ , m.$ Then the empirical best predictor is obtained as

$$H_i\left(\hat{\boldsymbol{\delta}}\right) = \exp\left\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}\right) + \frac{1}{2}\hat{\sigma}_u^2 \left(1 - \hat{\gamma}_i\right)\right\} \tag{3.1}$$

Unlike the previous method, this method estimates the target parameter directly by defining the function of model parameters.

For the estimation of mean squared prediction error, the Prasad-Rao (1990) type of formula is not applicable.

$$\begin{aligned} MSPE\left(\hat{\theta}_i^*\right) &= E\left(\hat{\theta}_i^* - \theta_i^*\right)^2 \\ &= E\left\{H_i\left(\hat{\boldsymbol{\delta}}\right) - \theta_i^*\right\}^2 \\ &= E\left\{H_i(\boldsymbol{\delta}) - \theta_i^*\right\}^2 + E\left\{H_i\left(\hat{\boldsymbol{\delta}}\right) - H_i(\boldsymbol{\delta})\right\}^2 \\ &= M_{1i}^*(\boldsymbol{\delta}) + M_{2i}^*(\boldsymbol{\delta}), \text{ say,} \end{aligned} \tag{3.2}$$

One can verify that $M_{1i}^*(\boldsymbol{\delta})$ is 0(1) term and $M_{2i}^*(\boldsymbol{\delta})$ is of $0\left(m^{-1}\right)$ term. Thus $M_{2i}^*(\boldsymbol{\delta})$ is estimated simply by $M_{2i}^*(\boldsymbol{\delta})$ to obtain an estimate of MSPE correct up to $0\left(m^{-1}\right)$. However, $M_{1i}^*(\boldsymbol{\delta})$as an estimate of $M_{1i}^*(\boldsymbol{\delta})$ would have bias of order $0\left(m^{-1}\right)$ for large *m*. Following the recommendation of Rao (2003), we apply jackknife method of estimation, recently proposed by Jiang, Lahiri and Wan (2002) to obtain appropriate estimate of $M_{1i}^*(\boldsymbol{\delta})$ and$M_{2i}^*(\boldsymbol{\delta})$. In particular, noting that

$$E\left(H_i(\boldsymbol{\delta}) - \theta_i^*\right)^2 = \exp\left[2\left\{\hat{\theta}_i^B + g_i\left(\sigma_u^2\right)\right\}\right] - \exp\left\{2\hat{\theta}_i^B + g_i\left(\sigma_u^2\right)\right\} = g_i^*\left(\sigma_u^2\right), \text{ say.}$$

Then the jackknife bias corrected estimate of $M_{1i}^*(\boldsymbol{\delta})$ is given by

$$\hat{M}_{1i}^*(\boldsymbol{\delta}) = g_i^*\left(\hat{\sigma}_u^2\right) - \frac{m-1}{m}\sum_{l=1}^{m} \left\{g_i^*\left(\hat{\sigma}_u^2(-l)\right) - g_i^*\left(\hat{\sigma}_u^2\right)\right\} \tag{3.3}$$

where, $\hat{\sigma}_u^2(-l)$ is obtained by deleting *l*-th observation from the data set. Instead of finding a closed form expression for $M_{2i}^*(\boldsymbol{\delta})$, we will apply jackknife method. The jackknife estimate of $M_{2i}^*(\boldsymbol{\delta})$ is given by

$$\hat{M}_{2i}^{*}(\boldsymbol{\delta}) = \frac{m-1}{m}\sum_{l=1}^{m}\left\{H_i\left(\hat{\boldsymbol{\delta}}(-l)\right) - H_i\left(\hat{\boldsymbol{\delta}}\right)\right\}^2 \tag{3.4}$$

where $\hat{\boldsymbol{\delta}}(-l)$ is $\hat{\boldsymbol{\delta}}$ obtained by deleting *l*-th data point from the full data set. By combining (3.3) and (3.4), a second order correct MSPE estimate of $\hat{\theta}_i^*$ is

$$mspe^*\left(\hat{\theta}_i^*\right) = \hat{M}_{1i}^*(\boldsymbol{\delta}) + \hat{M}_{2i}^*(\boldsymbol{\delta}) \tag{3.5}$$

The $mspe\left(\hat{\theta}_i\right)$ in (2.7) can be obtained either by using Prasad-Rao (1990) technique based on Taylor expansion or jackknife technique outlined in this section. We have used both the methods and found similar result. For brevity and comparability we report only the jackknife $mspe\left(\hat{\theta}_i\right)$.

$$mspe\left(\hat{\theta}_i\right) = \hat{M}_{1i}(\boldsymbol{\delta}) + \hat{M}_{2i}(\boldsymbol{\delta}) \tag{3.6}$$

where $\hat{M}_{2i}(\boldsymbol{\delta}) = \frac{m-1}{m}\sum_{l=1}^{m}\left\{\hat{\theta}_i(-l) - \hat{\theta}_i\right\}^2$ and

$$\hat{M}_{1i}(\boldsymbol{\delta}) = g_i\left(\hat{\boldsymbol{\delta}}(-l)\right) - \frac{m-1}{m}\sum_{l=1}^{m}\left\{g_i\left(\hat{\boldsymbol{\delta}}(-l)\right) - g_i\left(\hat{\boldsymbol{\delta}}\right)\right\}$$

where $\hat{\theta}_i(-l)$ is obtained by deleting *l*-th data point from the full data set.

4. Revisit the county level SAIPE data

We now apply the methods described in Section 2 and 3 to estimate the proportion of poor school-age children (children in the age-group 5—17) for all the counties in the U.S. for the year 1989. There is no particular reason of choosing 1989 other than we can compare the point estimates with the corresponding 1990 Census estimates, usually taken as the “gold standard”. Note that the point estimation is a secondary issue in this article.

We fit the FH model (2.1) — (2.2) to the county level SAIPE data where the response variable is log rate estimated from CPS data as described in the introduction. The county level predictor variables are as follows:

x_{1i} = logarithm of IRS - estimated Child Poverty Rate

x_{2i} = logarithm of Food - Stamp Participation rate

x_{3i} = logarithm of IRS Child Tax - Exemptions divided by Population estimate

x_{4i} = logarithm of Poverty Rate for residents aged 5 - 17 from the latest decennial census.

The predictors $\mathbf{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i})^T$ are available for all the counties. Counties with CPS samples but no poor school-age children are excluded in fitting the county model because of log transformation. There are 231 such counties out of 1259 counties in 1990. There are some difficulties with the sampling variances and model variances. The form described in Section 2 is one of several choices. For our illustration we assume v_e known and equal to 2.5 and then estimate σ_u^2 using maximum likelihood method. Let's denote by c_i as the Census based poverty rate, $\hat{\theta}_{1i}^* = \exp(\hat{\theta}_i) =$ estimate obtained by direct transformation back, and $\hat{\theta}_{2i}^* = H_i(\hat{\boldsymbol{\delta}}) =$ bias corrected estimate as outlined in Section 3. Along with these point estimates we produce the following measure of mean squared prediction error.

$R_{1i}^2 =$ mean squared prediction error of $\hat{\theta}_{1i}^*$ as given by formula (2.7).

$R_{2i}^2 =$ Naive mean squared prediction error of $\hat{\theta}_{2i}^* = g_i^*(\hat{\sigma}_u^2)$

$R_{3i}^* =$ Mean squared prediction error estimate of $\hat{\theta}_{2i}$ as given by formula (3.5)

Note that R_2^2 is kind of estimated posterior variance under log transformation and R_3^2 is its bias corrected version.

Table 1 provides $c_i, \hat{\theta}_{1i}^*, \hat{\theta}_{2i}^*, R_{1i}, R_{2i},$ and R_{3i} for thirty randomly selected counties to see the differences among them. Table 2 provide summary measures of these statistics for all the counties.

From the above tables it is clear that the direct transformation back approach under estimate the measure of errors.

The two sets of estimates are compared against the 1990 decennial census estimates for 1989. Based on the recommendation of the panel of National Academy of Sciences, for any estimate $\mathbf{e} = (e_1, \mathrm{K}, e_m)^T$, we compute the following:

- average relative bias = $\dfrac{1}{m}\sum_{e=1}^{m} \dfrac{|c_i - e_i|}{c_i}$
- average squared relative bias = $\dfrac{1}{m}\sum_{i=1}^{m} \dfrac{(c_i - e_i)^2}{c_i^2}$
- average absolute bias = $\dfrac{1}{m}\sum_{i=1}^{m} |c_i - e_i|$
- average squared deviation = $\dfrac{1}{m}\sum_{i=1}^{m} (c_i - e_i)^2$

Table 1. Point estimate and their mean squared prediction error for county level poverty estimate

County	n_i	y_i	c_i	$\hat{\theta}^*_{1i}$	$\hat{\theta}^*_{2i}$	R_{1i}	R_{2i}	R_{3i}
1	48	0.1880	0.237	0.1890	0.1930	0.04050	0.04190	0.04190
2	35	0.2570	0.285	0.2500	0.2580	0.06150	0.06420	0.06430
3	648	0.1740	0.187	0.1740	0.1740	0.01080	0.01080	0.01080
4	166	0.3550	0.296	0.3460	0.3480	0.04160	0.04210	0.04210
5	58	0.2240	0.283	0.2170	0.2220	0.04280	0.04400	0.04410
6	133	0.4660	0.247	0.4510	0.4550	0.06040	0.06110	0.06120
7	46	0.0870	0.122	0.0909	0.0931	0.01990	0.02060	0.02060
8	36	0.0278	0.298	0.0373	0.0384	0.00907	0.00945	0.00946
9	14	0.0714	0.503	0.0961	0.1020	0.03360	0.03670	0.03680
10	46	0.4780	0.277	0.4730	0.4850	0.10300	0.10700	0.10700
11	35	0.4290	0.182	0.3930	0.4050	0.09640	0.10100	0.10100
12	24	0.0833	0.194	0.0976	0.1020	0.02800	0.02970	0.02970
13	26	0.1540	0.362	0.1600	0.1660	0.04440	0.04690	0.04700
14	48	0.4790	0.312	0.4550	0.4660	0.09760	0.10100	0.10100
15	15	0.4000	0.280	0.3570	0.3790	0.12200	0.13300	0.13300
16	29	0.0690	0.108	0.0842	0.0872	0.02240	0.02350	0.02350
17	25	0.1200	0.186	0.1210	0.1260	0.03400	0.03600	0.03610
18	56	0.3040	0.283	0.2880	0.2940	0.05770	0.05940	0.05950
19	20	0.3500	0.223	0.3340	0.3500	0.10300	0.11000	0.11000
20	52	0.2690	0.156	0.2590	0.2640	0.05350	0.05520	0.05530
21	11	0.2730	0.242	0.2400	0.2570	0.09050	0.10100	0.10100
22	31	0.1610	0.195	0.1730	0.1790	0.04460	0.04690	0.04690
23	32	0.0937	0.257	0.1110	0.1150	0.02830	0.02970	0.02970
24	45	0.0889	0.215	0.0989	0.1010	0.02180	0.02260	0.02260
25	16	0.5000	0.212	0.3650	0.3860	0.12200	0.13200	0.13200
26	15	0.1330	0.324	0.1580	0.1670	0.05390	0.05860	0.05870
27	26	0.1540	0.279	0.1780	0.1850	0.04950	0.05230	0.05230
28	28	0.3930	0.256	0.3480	0.3600	0.09360	0.09870	0.09890
29	39	0.2560	0.206	0.2580	0.2650	0.06050	0.06300	0.06300
30	62	0.0968	0.231	0.1020	0.1040	0.01950	0.02010	0.02010

Table 2. Summary statistics for all the counties

	n_i	y_i	c_i	$\hat{\theta}^*_{1i}$	$\hat{\theta}^*_{2i}$	R_{1i}	R_{2i}	R_{3i}
Min.	1.00	0.00694	0.0266	0.00743	0.00749	0.000959	0.000969	0.00097
1st Qu.	26.00	0.09090	0.1190	0.09605	0.09893	0.017680	0.018100	0.01810
Median	48.00	0.16700	0.1705	0.16550	0.17000	0.034500	0.035900	0.03595
Mean	81.45	0.20450	0.1903	0.18810	0.19420	0.044080	0.046850	0.04694
3rd Qu.	79.00	0.27300	0.2460	0.25600	0.26500	0.058900	0.062130	0.06223
Max.	3130.00	1.00000	0.6770	0.82400	0.86400	0.253000	0.271000	0.27200

Table 3. The deviation statistics of the point estimates

Estimates	Average relative bias	Average squared relative bias	Average absolute bias	Average squared deviation
direct	0.8122	1.9357	0.1228	0.0279
θ^*_{1i}	0.7090	1.3694	0.1071	0.0196
θ^*_{2i}	0.6867	1.2628	0.1049	0.0188

Table 3 indicates that the bias corrected estimates outperform with respect to all of the measures.

5. Concluding Remarks

Transformations are often needed in practice to fit the data well with convenient models. Estimation from small area models is proven to be challenging, particularly the mean squared error estimation. Mean squared prediction error estimation from normal theory based regression model is well established in the literature. For this reason and convenience SAIPE decided to apply normal theory based area level small area models on log responses. While getting the point estimates using inverse transformation is not difficult, the mean squared prediction error estimation is non trivial. In this article we have described how to apply recently proposed jackknife method of mean squared prediction error estimation in case of log transformed data

In the data analysis our objective was not a SAIPE production rather it's an experiment about the mean squared prediction error estimation of the county poverty rate estimates. There may be other methods as well need to be explored.

REFERENCES

BELL, A. (1997). Models for county and state poverty estimates. Preprint, Census Statistical Research Division.

CITRO, C. and KALTON, G. (Eds) (2000). Small-Area Estimates of School-Age Children in Poverty, *Evaluation of Current Methodology* (National Research Council), Washington DC: Nat. Acad. Press.

FAY, R. and HERRIOT, R. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American statistical association*, 74, 341—352.

GHOSH, M. and RAO, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 55—93.

JIANG, J., LAHIRI, P., and WAN, S-M. (2002). A unified jackknife theory.Annals of Statistics, 30.

JIANG, J., LAHIRI, P., WAN, S-M., and WU, C-H. (2001). Jackknifing the Fay-Herriot model with an example. Tech. Rep. Dept. of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln.

MAITI, T. and Slud, E.V. (2002). Comparison of Small Area Models in SAIPE. Tech Rep. Census Bureau.

PRASAD, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American statistical association*, 74, 341—352.

RAO, J.N.K. (2003). Small Area Estimation. Wiley.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 697—711

A GENERALIZED CLASS OF COMPOSITE ESTIMATORS WITH APPLICATION TO CROP ACREAGE ESTIMATION FOR SMALL DOMAINS

G.C. Tikkiwal and Alka Ghiya[1]

ABSTRACT

This paper defines a generalized class of composite estimators, using auxiliary information, for small domains under simple random sampling and stratified random sampling schemes. The proposed class of composite estimators has desirable consistency property, and it includes a number of direct, synthetic and composite estimators. Further, this paper demonstrates the use of the estimators belonging to the generalized class for estimating crop acreage for small domains and also compare their relative performance with the corresponding direct and synthetic estimators, through a simulation study. The study suggests the use of some composite estimators at the ILRC's (the small domains under consideration) level and thus up to the level of district under certain conditions.

Key words: Composite estimators, Synthetic Estimators, Small Domains, Inspector Land Revenue Circles (ILRCs), Timely Reporting Scheme(TRS), Absolute Relative Bias(ARB), Simulated relative standard error (Srse), Simulation-Cum-Regression (SICURE) model

1. Introduction

Gonzalez and Waksberg (1973) and Schaible, Brock, Casady and Schnack (1977) compares errors of synthetic and direct estimates for standard Metropolitan Statistical Areas and Counties of U.S.A. The authors of both the papers conclude that when in small domains sample sizes are relatively small the synthetic estimator outperforms the simple direct; whereas, when sample sizes are large the direct outperforms the synthetic. These results suggest that a weighted sum of these two estimators, known as composite estimator, can provide an alternative to choosing one over the other. Further, Singh, Gambino and Mantel (1993) classifies composite estimators for small domains into three different categories: Type A — a direct estimator combined with a synthetic estimator, Type B — a

[1] Departament of Mathematics and Statistics J.N.V. University, Jodhpur — 342011, India.

synthetic adjusted combined with a synthetic, and Type C — a direct combined with a synthetic adjusted. The authors compare the performance of Type A and Type C categories of such estimators through a simulation study. The study also considers some direct and synthetic estimators. The simulation study, based on an artificial populations, concludes that the composite estimators of Type A perform better than those of Type C. Therefore, we confine to Type A category of composite estimators.

In general, a composite estimator of Type — A category may be defined as follows

$$\bar{y}'_{c,a} = w_a \; \bar{y}'_a + (1 - w_a)\bar{y}''_a$$

Where $\bar{y}'_a$ is a direct estimator and $\bar{y}''_a$ is a synthetic estimator of $\bar{Y}_a$, the population mean of small area 'a' and w_a are suitably chosen weights. Here $\bar{y}'_{c,a}$ is a Model- dependent estimator as it s a combination of design based estimator $\bar{y}'_a$ and model-dependent design biased estimator $\bar{y}''_a$ [cf. Sarndal (1994)]. The optimal values of w^*_a of w_a may be obtained by minimizing the mean square error of $\bar{y}'_{c,a}$ with respect to w_a and it is given by

$$w^*_a = \frac{MSE(\bar{y}"_a) - E(\bar{y}'_a - \bar{Y}_a)(\bar{y}''_a - \bar{Y}_a)}{MSE(\bar{y}'_a) + MSE(\bar{y}''_a) - 2E(\bar{y}'_a - \bar{Y}_a)(\bar{y}''_a - \bar{Y}_a)}$$

Under the assumption that $E(\bar{y}'_a - \bar{Y}_a)(\bar{y}''_a - \bar{Y}_a)$ is small relative to MSE $(\bar{y}''_a)$, the w^*_a becomes more manageable. In this case w^*_a may be approximated by

$$\mathrm{w}^{**}_\mathrm{a} = \frac{1}{1 + \mathrm{R}_\mathrm{a}}$$

where $R_a = MSE(\bar{y}'_a)/MSE(\bar{y}''_a)$. The weights $\mathrm{w}^{**}_\mathrm{a}$ can be estimated by replacing the mean square errors of $\bar{y}'_a$ and $\bar{y}''_a$ by their usual estimates, but the resulting estimates can be very unstable. To overcome this problem, Schaible (1978) proposes an "average" weighting scheme based on several weighting variables under three different models. Among these the Model 1 is most realistic. This model allows the mean square error of each component estimator to vary across the small area, that is $MSE(\bar{y}'_a) = b'_a$, $MSE(\bar{y}''_a) = b''_a$ for all a = 1, … , A.

In this paper we define a generalized class of composite estimators, using auxiliary information, under simple random sampling and stratified random sampling schemes. The generalized class of composite estimators, among others, includes the estimators such as : simple direct, simple synthetic, simple ratio, ratio synthetic, composite estimators forming by taking weighted sum of various direct and synthetic estimators referred above. Further, we demonstrate the use of estimators belonging to the generalized class for estimating crop acreage for small domains and also compare their relative performance with the corresponding direct and synthetic estimators, through a simulation study. The study suggests the use of composite estimators $t_{5,a}$ and $t_{7,a}$ at the ILRCs (the small domains under consideration) level, and thus up to the level of district, when there are no considerable deviations from their corresponding assumptions. When this condition is not satisfied, we should look for alternative methods of estimation obtained using either the SICURE model given by Tikkiwal, B.D. (1993) or presented by Ghosh & Rao (1994).

2. Notations

Suppose that a finite population U = (1, ... ,i, ... , N) is divided into 'A' non overlapping small domains U_a of size N_a (a = 1, ... , A) for which estimates are required. We denote the characteristic under study by 'y'. We further assume that the auxiliary information is available and denote this by 'x'. A random sample s of size n is selected through Simple Random Sampling Without Replacement (SRSWOR) design from population U such that n_a units in the sample's' comes from small domain U_a (a=1, ... , A).

Consequently,

$$\sum_{a=1}^{A} N_a = N \quad \text{and} \quad \sum_{a=1}^{A} n_a = n$$

We denote the various population and sample means for characteristics Z = X,Y by

$\overline{Z}$ = mean of the population based on N observations.

$\overline{Z}_a$ = population mean of domain 'a' based on N_a observations.

$\overline{z}$ = mean of the sample 's' based on n observations.

$\overline{z}_a$ = sample mean of domain 'a' based on n_a observations.

Also, the various mean squares and coefficient of variations of the population 'U' for characteristics Z are denoted by

$$S_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(z_i - \overline{Z}\right)^2, \qquad C_z = \frac{S_z}{\overline{Z}}$$

The coefficient of covariance between X and Y is denoted by

$$C_{xy} = \frac{S_{xy}}{\overline{X}\ \overline{Y}}$$

Where;

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)$$

The corresponding various mean squares and coefficient of variations of small domains U_a are denoted by

$$S^2_{z_a} = \frac{1}{N_a - 1}\sum_{i=1}^{N_a}\left(z_{a_i} - \overline{Z}_a\right)^2, \quad C_{z_a} = \frac{S_{z_a}}{Z_a} \qquad and \ C_{x_a y_a} = \frac{S_{x_a y_a}}{\overline{X}_a\,\overline{Y}_a}$$

where,

$$S_{x_a y_a} = \frac{1}{N_a - 1}\sum_{i=1}^{N_a}\left(y_{a_i} - \overline{Y}_a\right)\left(x_{a_i} - \overline{X}_a\right)$$

and Z_{a_i} (a = 1, ... , A and i = 1, ... , N_a) denote the i-th observation of the small domain 'a' for the characteristic Z = X,Y.

3. Generalised Class of Composite Estimators

We now define a generalized class of composite estimators of population mean $\overline{Y_a}$ based on auxiliary variable 'x' under SRSWOR design, as described in previous section, as follows

$$\overline{y}_{c,a} = w_a \overline{y}_a \left(\frac{\overline{x}_a}{\overline{X}_a}\right)^{\beta_1} + \left(1 - w_a\right)\overline{y}\left(\frac{\overline{x}}{\overline{X}_a}\right)^{\beta_2} \tag{3.1}$$

where $\beta_1 \text{and} \beta_2$ are suitably chosen constants.

The estimator $\overline{y}_{c,a}$ is a weighted sum of the direct estimator.

$$\overline{y}_{d,a} = \overline{y}_a \left(\frac{\overline{x}_a}{\overline{X}_a}\right)^{\beta_1} \tag{3.2}$$

an estimator which is expected to perform well for fairly big range of values of β_1 [Srivastava(1967)], and the generalized synthetic estimator

$$\overline{y}_{syn,a} = \overline{y}\left(\frac{\overline{x}}{\overline{X}_a}\right)^{\beta_2} \tag{3.3}$$

given by Tikkiwal and Ghiya (2000).

The proposed generalized class of composite estimators has desirable consistency property when the following assumption is satisfied.

$$\overline{Y}_a\left(\overline{X}_a\right)^{\beta_2} = \overline{Y}\left(\overline{X}\right)^{\beta_2} \tag{3.4}$$

It is to be noted that the estimator $\overline{y}_{syn,a}$ may be heavily biased unless the above assumption is satisfied.

Table 3.1. Various Direct and Indirect Estimators as special cases of the Generalized class of Composite Estimators

No.	Estimator	W_a	$(1-W_a)$	β_1	β_2
1.	Simple Direct $(\overline{y}_a)$	1	0	0	-
2.	Simple Synthetic $(\overline{y})$	0	1	-	0
3.	Simple Ratio $((\overline{y}_a / \overline{x}_a)\overline{X}_a)$	1	0	-1	-
4.	Ratio Synthetic $((\overline{y} / \overline{x})\overline{X}_a)$	0	1	-	-1
5.	Simple Product $((\overline{x}_a / \overline{X}_a)\overline{y}_a)$	1	0	1	-
6.	Product Synthetic $((\overline{y} / \overline{X}_a)\overline{x})$	0	1	-	1
7.	Composite: combining simple direct with simple synthetic $w_a\overline{y}_a + (1-w_a)\overline{y}$	w_a	$(1-w_a)$	0	0
8.	Composite: combining simple direct with ratio synthetic $w_a\overline{y}_a + (1-w_a)\dfrac{\overline{y}}{\overline{x}}\overline{X}_a$	w_a	$(1-w_a)$	0	-1
9.	Composite: combining simple ratio with ratio synthetic $w_a\dfrac{\overline{y}_a}{\overline{x}_a}\overline{X}_a + (1-w_a)\dfrac{\overline{y}}{\overline{x}}\overline{X}_a$	w_a	$(1-w_a)$	-1	-1

The proposed generalized class of composite estimators includes a number of direct, synthetic and composite estimators as special cases. Here follows a list of such estimators with corresponding choice of values of the different constants.

4. Design Bias and Mean Square Error

The design bias of the composite estimator $\overline{y}_{c,a}$ is given by

$$B(\bar{y}_{c,a}) = E(\bar{y}_{c,a}) - \bar{Y}_a = w_a \; B(\bar{y}_{d,a}) + (1 - w_a) \; B(\bar{y}_{syn,a}) \tag{4.1}$$

and

$$MSE(\bar{y}_{c,a}) = w_a^2 \; MSE(\bar{y}_{d,a}) + (1 - w_a)^2 \; MSE(\bar{y}_{syn,a}) + 2w_a(1 - w_a)E(\bar{y}_{d,a} - \bar{Y}_a)(\bar{y}_{syn,a} - \bar{Y}_a)$$

where $\bar{y}_{d,a}$ and $\bar{y}_{syn,a}$ are defined in Eqs. (3.2) and (3.3) respectively.

Under the assumption that $E(\bar{y}_{d,a} - \bar{Y}_a)(\bar{y}_{syn,a} - \bar{Y}_a)$ is small relative to MSE $(\bar{y}_{syn,a})$, as discussed in Section 1, the

$$MSE(\bar{y}_{syn,a}) = \mathrm{w}_\mathrm{a}^2 \mathrm{R}_\mathrm{a} + (1 - \mathrm{w}_\mathrm{a})^2 \tag{4.2}$$

Where

$$R_a = MSE(\bar{y}_{d,a}) / MSE(\bar{y}_{syn,a})$$

Under the given assumptions, appropriate value of the optimal weight $\mathrm{w}_\mathrm{a}^{**}$ is given by

$$\mathrm{w}_\mathrm{a}^{**} = \frac{1}{1 + \mathrm{R}_\mathrm{a}} \tag{4.3}$$

and the expressions of $B(\bar{y}_{d,a})$ *and* $MSE(\bar{y}_{d,a})$ are given by

$$B(\bar{y}_{d,a}) = \left(\frac{N_a - n_a}{N_a n_a} \right) \; \bar{Y}_a \left[\frac{\beta_1(\beta_1 - 1)}{2} C_{x_a}^2 + \beta_1 C_{x_a \; y_a} \right] \tag{4.4}$$

and

$$MSE(\bar{y}_{d,a}) = \left(\frac{N_a - n_a}{N_a n_a} \right) \bar{Y}_a^2 \left[\beta_1^2 C_{x_a}^2 + C_{y_a}^2 + 2\beta_1 C_{x_a y_a} \right] \tag{4.5}$$

Which is minimum if $\beta_1 = -C_{x_a y_a} / C_{x_a}^2$ [cf. Srivastava (1967)]

In order to obtain the bias and mean square error of the estimator $\bar{y}_{syn,a}$, let

$$\bar{y} = \bar{Y}(1 + \in_1) \qquad ; \quad \bar{x} = \bar{X}(1 + \varepsilon_2)$$

So that, $\mathrm{E}(\in_1) = \mathrm{E}(\in_2) = 0$ and

$$E(\in_1^2) = \frac{N - n}{Nn} C_y^2, \qquad E(\in_2^2) = \frac{N - n}{Nn} C_x^2, \qquad E(\in_1 \in_2) = \frac{N - n}{Nn} C_{xy}$$

The $\bar{y}_{syn,a}$ can be expressed as

$$\bar{y}_{syn,a} = \overline{Y}\left(\frac{\overline{X}}{\overline{X}_a}\right)^{\beta_2}(1+\in_1)(1+\in_2)^{\beta_2}$$

Assuming $|\in_2| < 1$

$$\bar{y}_{syn,a} = \overline{Y}\left(\frac{\overline{X}}{\overline{X}_a}\right)^{\beta_2}\left[1+\beta_2 \in_2 + \frac{\beta_2(\beta_2-1)}{2}\in_2^2 + \in_1 + \beta_2 \in_1\in_2 + \ldots\right]$$

Assuming further that the contribution of terms involving powers in $\in_1$ and $\in_2$ higher than the second to the value of $E(\bar{y}_{syn,a})$ is negligible, we get

$$E(\bar{y}_{syn,a}) = \overline{Y}\left(\frac{\overline{X}}{\overline{X}_a}\right)^{\beta_2}\left[1+\frac{N-n}{Nn}\left(\frac{\beta_2(\beta_2-1)}{2}C_x^2 + \beta_2 C_{xy}\right)\right]$$

and the design bias of $\bar{\mathrm{y}}_{\mathrm{syn,a}}$ is given by

$$B(\bar{y}_{syn,a}) = \overline{Y}\left(\frac{\overline{X}}{\overline{X}_a}\right)^{\beta_2}\left[1+\frac{N-n}{Nn}\left(\frac{\beta_2(\beta_2-1)}{2}C_x^2 + \beta_2 C_{xy}\right)\right] - \overline{Y}_a \qquad (4.6)$$

The $MSE(\bar{y}_{syn,a})$ is given by

$$MSE(\bar{y}_{syn,a}) = E(\bar{y}_{syn,a} - \overline{Y}_a)^2$$

$$= \overline{Y}^2\left(\frac{\overline{X}}{\overline{X}_a}\right)^{2\beta_2}\left[1+\frac{N-n}{Nn}\left\{(2\beta_2^2-\beta_2)\,C_x^2 + C_y^2 + 4\beta_2 C_{xy}\right\}\right]$$

$$-2\overline{Y}_a\overline{Y}\left(\frac{\overline{X}}{\overline{X}_a}\right)^{\beta_2}\left[1+\frac{N-n}{Nn}\left(\frac{\beta_2(\beta_2-1)}{2}C_x^2 + \beta_2 C_{xy}\right)\right] + \overline{Y}_a^2 \qquad (4.7)$$

The suitable value of β_2 is the one for which $MSE(\bar{y}_{syn,a})$ is minimum, so minimizing the $MSE(\bar{y}_{syn,a})$ w.r.t. β_2, gives simplified expression of β_2, if $\overline{X}_a \,=\!\!\!\&\, \overline{X}$ as follows

$$\beta_2 = \frac{\overline{Y}\left(C_x^2 - 4C_{xy}\right) - 2\overline{Y}_a\left(\frac{C_x^2}{2} - C_{xy}\right)}{\left(4\overline{Y}C_x^2 - 2\overline{Y}_a C_x^2\right)} \tag{4.8}$$

5. Generalized Class of Composite Estimators Under Stratification

Suppose that the finite population U = (1, ... , i, ... ,N) is divided into 'A' non overlapping domains $U_{\cdot a}$, of size $N_{\cdot a}$ (a=1, ... , A), for which estimates are required as discussed in Section 2. The population is also divided along a second dimension into 'H' non-overlapping categories (called groups) $U_{h\cdot}$ of size $N_{h\cdot}$ (h=1, … , H). As a result, the population is cross classified into HA cells, U_{ha} of respective sizes N_{ha}. Consequently,

$$N = \sum_{h=1}^{H} N_{h.} = \sum_{a=1}^{A} N_{.a} = \sum_{h=1}^{H}\sum_{a=1}^{A} N_{ha} \tag{5.1}$$

We assume that N_{ha} are known from a previous census or other reliable sources. Further, we assume that simple random samples of predetermined size $n_{h.}$ (h=1, … , H) are selected from group h such that $\sum_{h=1}^{H} n_{h.} = n$. That is, n is the size of the random sample selected using stratified random sampling. Also let $n_{.a}$ and n_{ha} (a=1, ... , A; h = 1, ... , H) are the units of the sample that belongs to domain $U_{\cdot a}$ and cell(h,a). So $n_{\cdot a}$ and n_{ha} are random.

Denoting $y_{ha_i}\left(i = 1, \ldots, N_{ha}\right)$, the i-th observation of the characteristic under study of the cell (h,a), we define various population and sample means as follows, using capital letters for population means and small letters for sample means.

$\overline{Y}_a = \frac{1}{N_a}\sum_{h=1}^{H} N_{ha}\overline{Y}_{ha}$ Population mean of small area 'a'

where, $\overline{Y}_{ha} = \frac{1}{N_{ha}}\sum_{i=1}^{N_{ha}} y_{ha_i}$

$\overline{y}_{a.} = \frac{1}{n_{.a}}\sum_{h=1}^{H} n_{ha}\overline{y}_{ha}$, Sample mean for small area ‘a’.

Where, $\bar{y}_{ha} = \dfrac{1}{n_{ha}} \sum_{i=1}^{n_{ha}} y_{ha_i}$

$\bar{Y}_{h.} = \dfrac{1}{N_{h.}} \sum_{a=1}^{A} N_{ha} \bar{y}_{ha}$, Population mean of the h-th group

$\bar{y}_{h.} = \dfrac{1}{n_{h.}} \sum_{a=1}^{A} n_{ha} \bar{y}_{ha}$, Sample mean of the h-th group

Similar notations are used, for various means for auxiliary characteristic x, just replacing 'y' with 'x' symbol.
We then following (3.1) define generalized class of composite estimators under stratification as follows

$$\bar{y}_{s,c,a} = \lambda_a \sum_{h=1}^{H} W_{ha} \bar{y}_{ha} \left(\frac{\bar{x}_{ha}}{\bar{X}_{ha}} \right)^{\beta_1} + \left(1 - \lambda_a\right) \sum_{h=1}^{H} W_{ha} \bar{y}_{h.} \left(\frac{\bar{x}_{h.}}{\bar{X}_{ha}} \right)^{\beta_2} \tag{5.2}$$

Where, $W_{ha} = N_{ha} / N_a$ and $\lambda_a (a = 1,...., A)$ are suitably chosen weights.

Remarks 5.1. The expressions of Bias and Mean square Error can directly be obtained following the expressions (4.1) and (4.2) of previous sections.

Remarks 5.2. For $\lambda_a = 0$ and $\beta_2 = 0,$ the estimator (5.2) reduces to

$\bar{y}'_{S,syn,a} = \sum_{h=1}^{H} W_{ha} \bar{y}_{h.}$, which in turn gives

$\hat{T}_{S,syn,a} = \sum_{h=1}^{H} N_{ha} \bar{y}_{h.}$, the estimator of population total 'T_a' of small area 'a' discussed by Sarndal [1984, Eq. (3.1), p.625].

Remarks 5.3. For $\lambda_a = 0$ and $\beta_2 = -1,$ the generalized estimator (5.2) reduces to

$$\bar{y}''_{S,syn,a} = \sum_{h=1}^{H} W_{ha} \frac{\bar{y}_{h.}}{\bar{x}_{h.}} \bar{X}_{ha}$$

This estimator is currently in use to provide improved estimator of states income in USA (cf. Schaible(1996),28-57)

6. Crop Acreage Estimation For Small Domains — a Simulation Study

In this section we demonstrate the use of the generalized composite estimator $\bar{y}_{c,a}$ along with the various direct and indirect estimators belonging to the general class to obtain crop acreage estimates for small domains and also compare their relative performance through a simulation study. This we do by taking up the state of Rajasthan, one of the state in India, for our case study.

6.1. Timely Reporting Scheme

In order to improve timelines and quality of crop acreage statistics, a scheme known as Timely Reporting Scheme (TRS) has been in vogue since early seventies in most of the States of India. The TRS has the objective of providing quick and reliable estimates of crop acreage statistics and there-by productions of the principle crops during each agricultural season. Under the scheme the Patwari (Village Accountant) is required to collect acreage statistics on a priority basis in a 20 percent sample of villages, selected by stratified linear systematic sampling design taking Tehsil (a sub — division of the District) as a stratum. These statistics are further used to provide state level estimates using direct estimators viz., unbiased (based on sample mean) and ratio estimators.

The performance of both the estimators in the state of Rajasthan, like in other states, is satisfactory at state level, as the sampling error is within 5 percent. However, the sampling error of both the estimators increases considerably, when they are used for estimating acreage statistics of various principle crops even at district level, what to speak of levels lower than a district. For example, the sampling error of direct ratio estimator for Kharif crops (the crops sown in June-July and harvested in October-November every year) of Jodhpur district, a district of Rajasthan State, for the agricultural season 1991-92 varies approximately between 6 to 68 percent. Therefore, there is need to use indirect estimators at district and lower levels for decentralized planning and other purposes like crop insurance.

6.2. Details of the Simulation Study

For the collection of revenue and other administrative purposes, the State of Rajasthan, like most of the other states of India, is divided into a number of districts. Further, each district is divided into a number of Tehsils and each Tehsil is also divided into a number of Inspector Land Revenue Circles (ILRCs). Each ILRC consists of a number of villages. For the present study, we take ILRCs as small domains.

In the simulation study, we undertake the problem of crop acreage estimation for all Inspector Land Revenue Circles (ILRCs) of Jodhpur Tehsil of Rajasthan. They are seven in number. These ILRCs are small domains from the TRS point of view. The crop under consideration is Bajra (Indian corn or millet) for the agriculture season 1993-94. the bajra crop acreage for agriculture season 1992-93 is taken as the auxiliary characteristic x. The various information regarding the ILRCs of Jodhpur Tehsil are provided in Table 6.2.1.

Table 6.2.1. Total Area (Irrigated And Unirrigated) Under Bajra Crop In Inspector Land Revenue Circles (Ilrcs) Of Jodhpur Tehsil For Agricultural Season 1992—93 And 1993—94

S.No	ILRC of Jodhpur Tehsil	No. of villages in ILRC	Total area (Irr. + U.Irr.) under the crop Bajra in 1992—93	Total area(Irr.+U.Irr) under the crop Bajra in 1993—94
1	Jodhpur (1)	29	7799.5899	5696.5000
2	Keru (2)	44	21209.5880	15699.6656
3	Dhundhada(3)	32	19019.0288	16476.4863
4	Bisalpur (4)	30	15153.9248	14269.0000
5	Luni (5)	33	19570.1323	16821.4508
6	Dhava (6)	40	25940.0979	25075.5000
7	Jajawal Kalan (7)	44	18007.4120	15875.0000
	Total	**252**	**126699.7737**	**109913.6027**

We now give below the list of all those estimators, whose relative performance is to be assessed for estimating population total T_a of small domain a for a = 1,2, …7.

Direct Estimators

(1) Direct ratio estimator $t_{1,a} = N_a \left(\frac{\bar{y}_a}{\bar{x}_a} \right) \bar{X}_a$

(2) Direct general estimator $t_{2,a} = N_a \bar{y}_a \left(\frac{\bar{x}_a}{\bar{X}_a} \right)^{\beta_1}$

Indirect Estimators

(3) Ratio synthetic estimator $t_{3,a} = N_a \left(\frac{\bar{y}}{\bar{x}} \right) \bar{X}_a$

(4) Generalized synthetic estimator $t_{4,a} = N_a \bar{y} \left(\frac{\bar{x}}{\bar{X}_a} \right)^{\beta_2}$

(5) Generalized composite estimator $t_{5,a} = N_a \bar{y}_{c,a}$

(6) Composite estimator $t_{6,a} = w_a N_a \bar{y}_a + (1 - w_a) t_{3,a}$

(7) Composite estimator $t_{7,a} = w_a t_{1,a} + (1 - w_a) t_{3,a}$

Before simulation, we first examine the following assumption, given earlier in Section 3.1 with respect to the seven small domains under study

$$\bar{Y}_a\left(\bar{X}_a\right)^{\beta_2} \cong \bar{Y}\left(\bar{X}\right)^{\beta_2}$$

taking $\beta_2 = -1$ for estimator $t_{3,a}$, $t_{6,a}$ *and* $t_{7,a}$ and then calculating optimum value of β_2 given in (4.8) for estimators $t_{4,a}$ *and* $t_{5,a}$. The Tables 6.2.2 and 6.2.3 provide relative absolute difference between $\bar{Y}_a / \bar{X}_a$ and $\bar{Y} / \bar{X}$, and between $\bar{Y}_a\left(\bar{X}_a\right)^{\beta_2}$ and $\bar{Y}\left(\bar{X}\right)^{\beta_2}$ for all the ILRCs respectively. From the examination of these Tables we note that both the assumptions closely meet for ILRCs 3, 5, 7 and also for ILRC 6 in second case only (Table 6.2.3). The assumptions deviate moderately for ILRC 4, but deviate considerably for ILRCs 1 & 2.

Table 6.2.2. Absolute Difference under Synthetic Assumption of Ratio Synthetic Estimator for various ILRCs.

ILRC	$\left(\bar{Y}_a / \bar{X}_a\right)$	$\left(\bar{Y} / \bar{X}\right)$	**Relative Absolute difference (%)** $\left[\ \left\lvert\left(\bar{Y}_a / \bar{X}_a\right) - \left(\bar{Y} / \bar{X}\right)\right\rvert \div \left(\bar{Y}_a / \bar{X}_a\right)\ \right] \times 100$
(1)	.7303	.8675	18.77
(2)	.7402	.8675	17.19
(3)	.8663	.8675	0.13
(4)	.9416	.8675	7.86
(5)	.8595	.8675	0.91
(6)	.9666	.8675	10.25
(7)	.8815	.8675	1.58

Table 6.2.3. Absolute Difference under Synthetic Assumption of Generalized Synthetic Estimator for various ILRCs

ILRC	$\bar{Y}_a\left(\bar{X}_a\right)^{\beta_2}$	$\bar{Y}\left(\bar{X}\right)^{\beta_2}$	**Relative Absolute difference** $\left[\left\lvert\bar{Y}_a\left(\bar{X}_a\right)^{\beta_2} - \bar{Y}\left(\bar{X}\right)^{\beta_2}\right\rvert \div \left(\bar{Y}_a\left(\bar{X}_a\right)^{\beta_2}\right)\right] \times 100$
(1)	3.3115	4.6578	40.6453
(2)	2.1134	2.4947	80.5662
(3)	0.7758	0.7791	0.4266
(4)	1.2314	1.1343	7.8851
(5)	0.8136	0.8223	1.0700
(6)	2.4000	2.4441	1.8345
(7)	0.1425	0.1378	3.2418

Now taking villages as sampling units for simulation purposes and otherwise, 500 independent simple random samples for each size of 25, 50, 63, 76 and 88 are selected from the population of 252 villages of Jodhpur Tehsil. Then, to assess the relative performance of the estimators under consideration, their Absolute Relative Bias (ARB) and Simulated relative standard error (Srse) or simply coefficient of variation are calculated for each ILRC as follows :

$$ARB\left(t_{k,a}\right) = \frac{\left|\frac{1}{500}\sum_{s=1}^{500} t_{k,a}^{s} - T_a\right|}{T_a} \times 100 \qquad (6.2.1)$$

and

$$Srse\left(t_{k,a}\right) = \frac{\sqrt{ASE}\left(t_{k,a}\right)}{E\left(t_{k,a}\right)} \times 100 \qquad (6.2.2)$$

where

$$ASE\left(t_{k,a}\right) = \frac{1}{500}\sum_{s=1}^{500}\left(t_{k,a}^{s} - T_a\right)^2 \text{ and } E\left(t_{k,a}\right) = \frac{1}{500}\sum_{s=1}^{500} t_{k,a}^{s}$$

for k = 1,2, ... ,7 and a= 1,2 , ... , 7.

6.3. Results

We present the results of ARB and Srse in Table 6.3.1 only for n = 50 (a sample of 20 percent villages, as presently adopted in TRS) as the findings from other tables are similar.

Table 6.3.1. Simulated Relative Standard Errors and Absolute Relative Biases

Estimator	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$t_{1,a}$	37.27 (0.21)	17.46 (2.28)	8.51 (0.76)	16.29 (0.13)	12.73 (2.41)	12.28 (0.32)	15.29 (2.78)
$t_{2,a}$	18.55 (0.96)	18.32 (1.50)	6.56 (0.12)	15.43 (0.18)	11.27 (1.12)	13.68 (0.54)	11.34 (0.61)
$t_{3,a}$	19.11 (17.90)	20.67 (19.50)	5.71 (0.72)	10.11 (8.66)	5.71 (0.05)	12.14 (11.03)	5.85 (1.02)
$t_{4,a}$	40.54 (39.67)	21.00 (19.84)	5.96 (0.11)	10.17 (8.68)	5.95 (0.18)	8.43 (4.06)	7.16 (3.97)
$t_{5,a}$	16.87 (5.70)	13.80 (9.29)	4.41 (0.10)	8.49 (6.21)	6.17 (0.88)	6.29 (2.82)	6.05 (3.09)
$t_{6,a}$	17.02 (0.91)	17.14 (1.20)	5.07 (0.11)	9.95 (0.17)	8.03 (1.09)	11.42 (0.51)	5.61 (0.59)
$t_{7,a}$	16.48 (8.40)	13.48 (10.20)	4.78 (0.50)	10.15 (6.30)	5.01 (0.38)	8.14 (4.60)	5.45 (1.20)

Note : the figure shown in parentheses are the absolute relative biases in percentage.

For assessing relative performance of the various estimators, we have to adopt some rule of thumb. Here we adopt the rule that at the ILRC level, an estimator should not have Srse more than 10 percent and Bias more than 5 percent. We note from, the above table that none of the estimators satisfy the rule in ILRCs 1 and 2. This is happening because, in these circles, there is considerable deviation from the synthetic assumptions, as observed earlier. In ILRCs 4, where the synthetic assumptions deviate moderately, $t_{6,a}$ alone satisfy the rule. But in ILRC6, where also deviations are moderate, $t_{5,a}$ is best. The estimator $t_{5,a}$ is best in ILRC 4 also, provided we restrict ourselves only to Srse. In ILRC 3, where the synthetic assumptions closely meet, $t_{5,a}$ is best here also. However in ILRCs 5 to 7, where too he synthetic assumptions closely meet, $t_{7,a}$ alone is competitive with others.

From the above analysis and otherwise, we recommend the use of composite estimators $t_{5,a}$ and $t_{7,a}$ at the ILRC level, and thus up to the level of district in TRS, where there are no considerable deviations from synthetic assumptions. When this condition is not satisfied we should look for other types of estimators such as those through the SICURE Model [B.D. Tikkiwal, (1993)] or presented in Ghosh and Rao (1994) and assess there relative performance through studies of the kind, in series, over some years.

REFERENCE

GHOSH, M. and RAO, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 91, 55—93.

GONZALEZ, N.E. and WAKSBERG, J.(1973), estimation of the error of synthetic estimates. Paper presented at first meeting of international association of survey statisticians, Vienna, Austria, 18—25.

SARNDAL, C.E. (1984). Design — consistent versus model dependent estimator for small domains, J. Amer. Statist. Assoc., 79, 624—631.

SCHAIBLE, W.L. (1978). Choosing weights for composite estimators for small area statistics. Proceedings of the survey research methods section, Amer. Statist. Assoc., Washington. D.C., 741—746.

SCHAIBLE, W.L. (1996). Indirect Estimation of U.S. Federal programs. Research Monograph, Springer-verlag.

SCHAIBLE, W.L., BROCK, D.B., CASADY, R.J. and SCHNACK, G.A. (1977). An empirical comparison of the simple synthetic and composite estimators for small area statistics. Proceedings of Amer. Statist. Assoc., Social Statistics section 1017—1021.

SINGH, M.P., GAMBINO., J. and MANTEL, H. (1993). Issues and options in the provision of small area data. Proceedings of International scientific conference on small area statistics and survey design (held in September, 1992 at Warsaw, Poland), 37—75.

SRIVASTAVA, S.K. (1967) An estimator using auxiliary information in sample surveys. Calcutta statist. Assoc. Bull., 121—132.

TIKKIWAL, B.D. (1993). Modeling through survey data for small domains. Proceedings of International Scientific Conference on small Area Statistics and Survey Design (An invited paper), held in September 1992 at Warsaw, Poland.

TIKKIWAL, G.C. and GHIYA A. (2000). A generalized class of synthetic estimators with application of crop acreage estimation for small domains. Biom, J. 42, 7, 865—876.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 713—724

RATIO ESTIMATION FOR SMALL DOMAINS WITH SUBSAMPLING THE NON-RESPONDENTS: AN APPLICATION OF RAO STRATEGY

Godwin A. Udofia[1]

ABSTRACT

In this article, we consider modifications of some of the procedures for global ratio estimation in single-phase sampling with subsampling the non-respondents proposed by Rao (1986) to obtain an estimate of mean for a small domain that cuts across constituent strata of a population with unknown weights. The bias and mean-square error of each of the modified estimators are obtained for comparison. Unlike Rao (1986), the population mean of the auxiliary variable is assumed to be unknown before the start of the survey and hence double sampling is applied. Stratified simple random sampling is considered. Similar work on the ratio estimators proposed by Rao (1986) and extension to other sampling designs are the subject of an on-going research by the author.

Key words: Small domains; Auxiliary information; Biases; Mean-square-errors; Double Sampling.

1. Introduction

Rao (1986) presented certain sampling procedures for global ratio estimation of the mean of a characteristic of interest with subsampling the non-respondents and proposed certain ratio estimators suitable for different practical situations. Very often, such estimators are desired for small subpopulations also called domains of study by the U.N. Subcommission on Sampling (1950). The variance or mean-square-error of an estimator is usually increased when the estimation procedure is extended to small domains. It is of interest therefore to see how the sampling strategies proposed by Rao (1986) can be modified for small domains when the mean of the auxiliary variable is unknown.

As an attempt in this direction, two of the global ratio estimators, t_1 and t_2, proposed by Rao (1986) are modified for small domains that cut across

[1] Godwin A. Udofia, Department of Mathematics/Statistics, University of Calabar, Nigeria.

constituent strata of a reference population with unknown weights. The bias and mean-square-error of each of the proposed domain ratio estimators are obtained in section 3. Application of the procedure to double sampling for domain estimation is discussed in section 4.

2. Sampling in One Phase

Let $\pi = \{u_1, u_2, .., u_N\}$ denote a finite population the elements of which fall into L known strata with N_h elements in the hth stratum, h = 1, 2, …, L, $\sum_h N_h = N$. It is assumed that π can also be partitioned according to the distribution of variable Z into exhaustive set of M subpopulations or domains of study that is denoted by $\{D_j^* \; ; \; j = 1,2....,M\}$. As an example, buildings in a given territory can be grouped by number of outside doors and windows into strata and by number of sleeping rooms into domains of study for the purpose of estimating the population size. Each stratum consists of a substratum of N_{1h} respondents and a substratum of N_{2h} non-respondents, $N_{1h} + N_{2h} = N_h$, all h.

Let D_{hj}^* denote the part of domain j (D_j^*) in stratum h and N_{hj} the unknown number of elements in D_{hj}^*. Let y_{hij} denote the value of characteristic Y for element i in D_{hj}^*. The sampling procedure is as defined in subsection 2.1 below.

2.1. Subsampling the Non-respondents for Domain Estimation

A sample of size n is drawn from π by taking $n_h, \sum_h n_h = n,$ units by simple random sampling without replacement (SRSWOR) and independently from the hth stratum. Of the n_h units, n_{1h} respond and $n_{2h} = n_h - n_{1h}$ fail to respond. Also, out of the n_h units from stratum h, n_{hj} belong to D_{hj}^*, $n_j = \sum_h n_{hj}$. Out of the n_{1h} units that respond in stratum h, n_{1hj} belong to D_{hj}^* while $n_{2hj} = n_{hj} - n_{1hj}$ out of the n_{2h} non-respondents in the same stratum also fall in D_{hj}^*, $n_{hj} = n_{1hj} + n_{2hj}$. Following Hansen and Hurwitz (1946), a subsample of $m_h = n_{2h}/k$, $k > 1$, units is drawn from the n_{2h} non-respondents. Let m_{hj} denote the number of units in the subsample of m_h that belong to D_{hj}^*. Since N_{hj} is not known before the sample is drawn, n_{1hj}, n_{2hj} as well as m_{hj} are random variables.

2.2. Other Notations

The following notations are defined and used in the derivation of the results following Durbin (1958), Hartley (1959), Tin and Toe (1972) and Tripathi (1988). For any unit i in π,

$$y'_{hi} = y_{hij} \ if \ i \in D^*_{hj},$$

$$= 0 \text{ if } \text{ i} \notin D^*_{hj} \tag{2.1}$$

Then

$$\bar{Y}'_h = \left(\sum_{i=1}^{N_h} y'_{hi}\right) / N_h = \frac{1}{N_h}\sum_{i=1}^{N_{hj}} y_{hij} = \frac{N_{hj}}{N_h}\bar{Y}_{hj} \ where \ \bar{Y}_{hj} = \frac{1}{N_{hj}}\sum_{i=1}^{N_{hj}} y_{hij} \ ;$$

$$\bar{Y}'_{2h} = \left(\sum_{i=1}^{N_{2h}} y'_{hi}\right) / N_{2h} = \frac{1}{N_{2h}}\sum_{i=1}^{N_{2hj}} y_{hij} = \frac{N_{2hj}}{N_{2h}}\bar{Y}_{2hj} \ where \ \bar{Y}_{2hj} = \frac{1}{N_{2hj}}\sum_{i=1}^{N_{2hj}} y_{hij} \ ;$$

$$\bar{y}'_{1h} = \left(\sum_{i=1}^{n_{1h}} y'_{hi}\right) / n_{1h} = \frac{n_{1hj}}{n_{1h}}\bar{y}_{1hj} \ where \ \bar{y}_{1hj} = \frac{1}{n_{1hj}}\sum_{i=1}^{n_{1hj}} y_{hij} \ ;$$

$$\bar{y}'_{2h} = \left(\sum_{i=1}^{n_{2h}} y'_{hi}\right) / n_{2h} = \frac{n_{2hj}}{n_{2h}}\bar{y}_{2hj} \ where \ \bar{y}_{2hj} = \frac{1}{n_{2hj}}\sum_{i=1}^{n_{2hj}} y_{hij} \ ;$$

$$\bar{y}'_{m_h} = \left(\sum_{i=1}^{m_h} y'_{hi}\right) / m_h = \frac{m_{hj}}{m_{mh}}\bar{y}_{mhj} \ where \ \bar{y}_{mhj} = \frac{1}{m_{hj}}\sum_{i=1}^{m_{hj}} y_{hij};$$

$$S_{xy(hj)} = \frac{1}{N_{hj}-1}\sum_{i=1}^{N_{hj}} (x_{hij} - \bar{X}_{hj})(y_{hij} - \bar{Y}_{hj}); \tag{2.2}$$

and

$$S_{2xy(hj)} = \frac{1}{N_{2hj}-1}\sum_{i=1}^{N_{2hj}} (x_{hij} - \bar{X}_{2hj})(y_{hij} - \bar{Y}_{2hj}) \tag{2.3}$$

The order notations are defined as the need arises.

3. Ratio Estimators for Domain Mean

Since N$_{hj}$ is not known the mean, $\bar{Y}_j$, of characteristic Y for domain j is defined by using y'_{hi} in (2.1) as

$$\bar{Y}_j = \sum\nolimits_h W_h \bar{Y}'_h \ where \, W_h = N_h / N.$$

3.1. The Conventional Ratio Estimator For Domain Mean

In a conventional ratio estimation procedure with subsampling the non-respondents discussed by P. S.R.S. Rao (1986), it is assumed that in a random sample of size n, there are n_1 respondents and

$n_2 = n - n_1$ non-respondents on both variables X and Y. The ratio estimator of $\overline{Y}$ is given as

$$t_1 = \frac{\overline{y}^*}{\overline{x}^*}\overline{X} = r^*\overline{X};\ r^* = \overline{y}^*/\overline{x}^* \qquad (3.1)$$

where $\overline{X}$ is assumed to be known,

$\overline{y}^* = w_1\overline{y}_1 + w_2\overline{y}_{2m},\ \overline{x}^* = w_1\overline{x}_1 + w_2\overline{x}_{2m},\ w_1 = n_1/n,$

$w_2 = n_2/n$, $\overline{y}_1 = \left(\sum_{i=1}^{n_1} y_i\right)/n_1$ is the mean of the n_1 respondents, and

$\overline{y}_{2m} = \left(\sum_{i=1}^{m} y_i\right)/m$ is the

mean of a subsample of $m = n_2/k$, $k \geq 1$, non-respondents. For a given n, $\overline{y}_{2m}$ is unbiased for

$\overline{y}_2 = \left(\sum_{i=1}^{n_2} y_i\right)/n_2$ and hence $\overline{y}^*$ is unbiased for $\overline{y}_1 = w_1\overline{y}_1 + w_2\overline{y}_2$.

The bias of the ratio estimator in (3.1) is given by Rao (1986) as

$$B(t_1) = \frac{1-f}{n\overline{X}}\left(RS_x^2 - S_{xy}\right) + \frac{W_2(k-1)}{n\overline{X}}(RS_{2x}^2 - S_{2xy}) \qquad (3.2)$$

and its large sample mean-square error (MSE) is given by

$$\text{MSE}(t_1) = \frac{1-f}{n}\sum_{h=1}^{2}\frac{(NW_h - 1}{(N-1)}S_{gh}^2 + W_2\frac{(k-1)}{n}S_{2g}^2 \quad \text{b} \qquad (3.3)$$

where

$S_{gh}^2 = \sum_{i=1}^{N_h}(y_{hi} - Rx_{hi})^2/(N_h - 1)\ for\ h = 1, 2, S_{2g}^2 = \sum_{i=1}^{N_2}(y_i - RX_i)^2/(N_2 - 1)\ and\ W_2 = N_2/N$

for the N_2 non-respondents in the population.

An estimator of MSE(t_1) is obtained by replacing

$$S_{gh}^2 \ by \ s_{gh}^2 = \sum_{i=1}^{n_h} (y_i - r * x_i)^2 / (n_h - 1),$$

$$S_{2g}^2 \ by \ s_{2g}^2 = \sum_{i=1}^{m} (y_i - r * x_i)^2 / (m - 1),$$ and W_h by $w_h = n_h/n$.

A corresponding ratio estimator for domain j can be expressed in terms of (2.1) as

$$t_{1j} = \sum_h W_h \frac{\bar{y}_h^{*'}}{\bar{x}_h^{*'}} \bar{X}_h^{'} \tag{3.4}$$

where

$$\bar{y}_h^{*'} = w_{1h} \bar{y}_{1h}^{'} + w_{2h} \bar{y}_{2m_h}^{'} \tag{3.5}$$

is unbiased for $\bar{y}_h^{'} = w_{1h} \ \bar{y}_{1h}^{'} + w_{2h} \bar{y}_{2h}^{'} \ and \ \bar{x}_h^{*'} = w_{1h} \ \bar{x}_{1h}^{'} + w_{2h} \bar{x}_{2m_h}^{'}$

is also unbiased for $\bar{x}_h^{'} = w_{1h} \ \bar{x}_{1h}^{'} + w_{2h} \bar{x}_{2h}^{'}, \ w_{1h} = n_{1h} / n_h, w_{2h} = n_{2h} / n_h$

for the respondents group and the non-respondents group respectively. To a first approximation

$$t_{1j} - \bar{Y}_j = \sum_h W_h (\bar{y}_h^{*'} - R_h^{'} \bar{x}_h^{*'})(1 - \delta\bar{x}_h^{*'}) \tag{3.6}$$

where

$$\delta\bar{x}_h^{*'} = (\bar{x}_h^{*'} - \bar{X}_h^{'}) / \bar{X}_h^{'} \ and \ R_h^{'} = \bar{Y}_h^{'} / \bar{X}_h^{'} = \bar{Y}_{hj} / \bar{X}_{hj} = R_{hj}$$

The bias of t_{1j} is obtained from (3.6) as

$$B(t_{1j}) = E(t_{1j} - \bar{Y}_j) \approx \sum_h W_h \left[\frac{1 - f_h}{n_h \bar{X}_h^{'}} \left(R_h^{'} S_{x'_h}^2 - S_{x_h^{'} y_h^{'}} \right) + \frac{W_{2h}(k-1)}{n_h \bar{X}_h^{'}} \left(R_h^{'} S_{2x_h^{'}}^2 - S_{2x_h^{'} y_h^{'}} \right) \right] \tag{3.7}$$

which compares with (3.2)

To obtain an expression of (3.7) in terms of the domain information, all that is needed is to express the variances and covariances in (3.7) in terms of $y_{hi}^{'}$ and hence $x_{hi}^{'}$ as defined in (2.1).

From Durbin (1958) and Udofia (1992):

$$S^2_{x'_h} = \frac{N_{hj}-1}{N_h-1}S^2_{x(h_j)} + \frac{N_{hj}}{N_h-1}\left(1-\frac{N_{hj}}{N_h}\right)\left(\overline{X}_{hj}-\overline{X}_j\right)^2; \qquad (3.8)$$

$$S^2_{2x'_h} = \frac{N_{2hj}-1}{N_{2h}-1}S^2_{2x(h_j)} + \frac{N_{2hj}}{N_{2h}-1}\left(1-\frac{N_{2hj}}{N_{2h}}\right)\left(\overline{X}_{2hj}-\overline{X}_{2j}\right)^2; \qquad (3.9)$$

Also we obtain

$$S_{x'_h y'_h} = \frac{1}{N_h-1}\sum_{1=1}^{N_h}\left(x'_{hi}-\overline{X}'_h\right)\left(y'_{hi}-\overline{Y}'_h\right) \qquad (3.10)$$

Substitution for y'_{hi} and x'_{hi} yields the result

$$S_{x'_h y'_h} = \frac{N_{hj}-1}{N_h}S_{xy(hj)}$$

Substitution of (3.8) to (3.10) in (3.7) gives the result

$$B(t_{1j}) = \sum_h Wh\{ \frac{1-f_h}{n_h P_{hj}\overline{X}_{hj}}\left(\frac{NW_{hj}-1}{NW_h-1}\right)\left[R_{hj}S^2_{x(hj)} - S_{xy(hj)} + \frac{NW_{hj}}{NW_h-1}Q_{hj}R_{hj}\left(\overline{X}_{hj}-\overline{X}_j\right)^2\right] +$$

$$+\frac{W_{2h}(k-1)}{n_h P_{hj}\overline{X}_{hj}}\frac{N_2W_{2hj}-1}{N_2W_{2h}-1}\left[R_{hj}S^2_{2x(hj)} - S_{2xy(hj)} + \frac{N_2W_{2hj}}{(N_2W_{2h}-1)}Q_{2hj}R_{hj}(\overline{X}_{2hj}-\overline{X}_{2j})^2\right]\Bigg\} \qquad (3.11)$$

where $P_{hj} = \dfrac{N_{hj}}{N_h}, Q_{hj} = 1-\dfrac{N_{hj}}{N_h}$ *and* $Q_{2hj} = 1-\dfrac{N_{2hj}}{N_{2h}}$ *and* $W_{hj} = N_{hj}/N$

A large sample approximation of the mean-square error,
MSE(t_1) = $E(t_{1j}-\overline{Y}_j)^2$, of t_{1j} is obtained from (3.6) as MSE(t_{1j})

$$\approx \sum_h W_h^2\left\{\frac{1-f_h}{n_h}\left(\frac{NW_{hj}-1}{NW_h-1}\right)\left[S^2_{z(hj)} + \frac{NW_{hj}}{NW_{hj}-1}Q_{hj}\left((\overline{Y}_{hj}-\overline{Y}_j)^2 + R^2_{hj}(\overline{X}_{hj}-\overline{X}_j)^2\right)+\right.\right.$$

$$\left.\left.+\frac{W_2(k-1)}{n_h}\frac{N_2W_{2hj}-1}{N_2W_{2h}-1}\left[S^2_{2z(hj)} + \frac{N_2W_{2hj}}{N_2W_{2hj}-1}Q_{2hj}\left((\overline{Y}_{2hj}-\overline{Y}_{2j})^2 + R^2_{hj}(\overline{X}_{2hj}-\overline{X}_{2j})^2\right)+\right.\right\} \qquad (3.12)$$

where $z_i = y_{hij} - R_{hj} X_{hij}$ if i $\in D^*_{hj}$ and zero otherwise.

Thus $S^2_{z(hj)} = \sum_{i=1}^{N_{hj}}(y_{hij} - R_{hj}\ x_{hij})^2/(N_{hj}-1)$ since the mean of z_i over D^*_{hj} is zero.

An estimator of MSE(t_{ij}) is obtained by replacing
W_{hj} by $w_{hj} = n_{hj}/n$, W_h by $w_h = n_h/n$,

W_{2hj} by $w_{2hj} = n_{2hj}/n$, $\bar{Y}_{hj}$ by

$$\bar{y}_{hj} = \sum_{i=1}^{n_{hj}} y_{hij} / n_{hj}, \bar{Y}_j \ by \ \bar{y}_j = \sum\nolimits_h n_{hj} \ \bar{y}_{hj} / n_j, \bar{Y}_{2hj} \ by$$

$$\bar{y}_{2hj} = \sum_{i=1}^{n_{2hj}} y_{hij} / n_{2hj}, \bar{Y}_{2j} \ by \ \bar{y}_{2j} = \sum\nolimits_h n_{2hj} \ \bar{y}_{2hj} / n_{2j}, Q_{hj} \ by \ q_{hj} = 1 - \frac{n_{hj}}{n_h}, R_{hj} \ by \ r_{hj} = \bar{y}_{hj} / \bar{x}_{hj}$$

and $S^2_{z(hj)} \ by \ s^2_{z(hj)} = \sum_{i=1}^{n_{hj}} (y_{hij} - r_{hj} \ x_{hij})^2 / (n_{hj} - 1)$

3.2. An Alternative Ratio Estimator For Domain Mean

In a situation where there is no non-response on the auxiliary variable, Rao (1986) has proposed the global ratio estimator $t_2 = \frac{\bar{y}*}{\bar{x}}\bar{X}$ where $\bar{x}$ is the sample mean for the auxiliary variable, X. Examples of the type of auxiliary variable considered are given by Rao. The bias of this estimator has also been given by Rao (1986) as

$$B(t_2) = E(t_2 - \bar{Y}) \approx \frac{1-f}{n\bar{X}}(RS_x^2 - S_{xy}) \tag{3.13}$$

and the large sample approximation to its mean-square-error is also given as

$$MSE(t_2) \approx \frac{1-f}{n} \frac{\sum_h (NW_h - 1) S_{dh}^2}{N-1} + W_2 \frac{(k-1)}{n} S_{2y}^2 \tag{3.14}$$

The corresponding estimator for domain j, j = 1, 2, …, M, and its statistical properties will now be considered.
A ratio estimator of $\bar{Y}_j$ that corresponds to t_2 is given in terms of (2.1) as

$$t_{2j} = \sum_h W_h \frac{\bar{y}_h^{*'}}{\bar{x}_n^{'}} \bar{X}_h^{'} \tag{3.15}$$

We can derive expressions for the bias and MSE of t_{2j} from

$$t_{2j} - \bar{Y}_j = \sum_h W_h \left[(\bar{y}_h^{*'} - R_h^{'} \bar{x}_h^{'})(1 - \delta\bar{x}_h^{'}) \right] \tag{3.16}$$

where $\delta\bar{x}_h^{'} = (\bar{x}_h^{'} - \bar{X}_h^{'}) / \bar{X}_h^{'}$. By using (3.16), the bias of t_{2j} can be obtained as

$$B(t_{2j}) = E(t_{2j} - \overline{Y}_j) = \sum_h W_h \frac{1-f_h}{n_h \overline{X}_h}\left(R_h^{'} S_{x_h^{'}}^2 - S_{x_h^{'} y_h^{'}}\right) \qquad (3.17)$$

Substitution from (3.8) and (3.10) in (3.17) gives the result

$$B(t_{2j}) = \sum_h W_h \frac{1-f}{n_h P_{hj} \overline{X}_{hj}} \frac{(NW_{hj}-1)}{(NW_h - 1} \Big[R_{hj} S_{x(hj)}^2 -$$

$$S_{xy(hj)} + \frac{NW_{hj}}{(NW_{hj}-1)} Q_{hj} R_{hj} (\overline{X}_{hj} - \overline{X}_j)^2 \Bigg] \qquad (3.18)$$

From (3.16), the MSE of t$_{2j}$ can be obtained as

$$\text{MSE}(t_{2j}) = \sum_h W_h^2 \left[\frac{1-f_h}{n_h}(S_{y'_h}^2 + R_h^{'2} S_{x'_h}^2 - 2R_h^{'} S_{x_h^{'} y_h^{'}}) + \frac{W_{2h}(k-1)}{n_h} S_{2y'_h}^2\right] (3.19)$$

Substitution from (3.8) and (3.10) gives the result
MSE(t$_{2j}$)=

$$\sum_h W_h^2 \left\{ \frac{1-f_h}{n_h}\left(\frac{NW_{hj}-1}{NW_h - 1}\right)\left[S_{z(hj)}^2 + \frac{NW_{hj}}{(NW_{hj}-1)} Q_{hj}\left((\overline{Y}_{hj} - \overline{Y}_j)^2 + R_{hj}^2 (\overline{X}_{hj} - \overline{X}_j)^2\right)\right]\right.$$

$$\left. + \frac{W_{2h}(k-1)}{n_h}\left(\frac{N_2 W_{2hj} - 1}{N_2 W_{2h} - 1}\right)\left[S_{2y(hj)}^2 + \frac{N_2 W_{2hj}}{N_2 W_{2hj} - 1} Q_{2hj} (\overline{Y}_{2hj} - \overline{Y}_{2j})^2\right]\right\} (3.20)$$

where z$_i$ retains its earlier definition.

An estimator of MSE(t$_{2j}$) is obtained by replacing

$W_h, W_{hj}, W_{2hj}, Q_{hj}, \overline{Y}_{hj}, \overline{Y}_j, \overline{Y}_{2hj}, \overline{Y}_{2j},$

$R_{hj}, S_{z(hj),}^2 S_{2y(hj)}^2$ and Q_{2hj} *by* $w_h, w_{hj}, w_{2hj}, q_{hj}, \overline{y}_{hj}, \overline{y}_j, \overline{y}_{2hj}, \overline{y}_{2j}, r_{hj}, s_{z(hj)}^2$

$s_{2y(hj)}^2 = \sum_{i=1}^{n_{2hj}} (y_{hij} - \overline{y}_{2hj})^2 /(n_{2hj} - 1)$ *and* $q_{2hj} = n_{2hj} / n_{2h}$ respectively.

4. The Use of Double Sampling For Domain Estimation

In a situation where the mean of the auxiliary variable X for stratum h, h = 1, 2, …, L, is not known contrary to the assumption used by Rao (1986) for t$_1$ and t$_2$ and there is no nonresponse for X, the following sampling strategy is proposed. An initial sample $S(n_h^{'})$, of size $n_n^{'}$ is taken by SRSWOR and independently from stratum h, h = 1, 2, …, L, and the auxiliary variable X is measured on it under general specifications in the survey design or fixed survey rules also known as “essential conditions” and discussed in detail by Hansen and Hurwitz (1946) and by Hansen, Hurwitz and Madow (1953) vol. II.

Let $n_{nj}^{'}$ denote the number of units in $S(n_h^{'})$ that fall in D_{hj}^{*}. A second sample, $S(n_h^{*})$, of size n_h^{*}, $n_h^{*} < n_n^{'}$, is drawn from $S(n_h^{*})$ by SRSWOR and the study variable, Y, is measured on it under more expensive and more efficient essential conditions than the general essential conditions that can produce accurate information for the units of the subsample. Examples of such efficient essential conditions are best interviewers, probing, examination of relevant records, highly effective supervision of interviewers, coders and other staff involved in the data analysis. Let n_{hj}^{*}, denote the number of units in $S(n_h^{*})$ that fall in D_{hj}^{*}, n_{1h}^{*} the number of respondents in $S(n_h^{*})$ out of which n_{1hj}^{*} belong to D_{hj}^{*}, and $n_{2h}^{*} = n_h^{*} - n_{1h}^{*}$ the number of nonrespondents in $S(n_h^{*})$ out of which n_{2hj}^{*} belong to D_{hj}^{*}. A subsample of $m_h^{*} = n_{2h}^{*} / k, k \geq 1$, is drawn from the n_{2h}^{*} nonrespondents. Let m_{hj}^{*} of the m_h^{*} nonrespondents belong to D_{hj}^{*}. Let $x_{hi}^{'} = x_{hij}$ for all i $\in D_{hj}^{*}$ and zero otherwise. Let $y_{hj}^{'}$ retain its definition in (2.1). Since $\overline{X}_h^{'}$ is not known, we can use as its unbiased estimator the sample mean $\bar{x}_{n'_h}^{'} = \sum_{i=1}^{n_h^{'}} x_{hi}^{'} / n_h^{'}$ defined on the initial sample $S\left(n_h^{'}\right)$. A ratio estimator of $\overline{Y}_j$ is thus obtained as $t_{3j} = \sum_h \frac{\bar{y}'_{n_h^*}}{\bar{x}'_{n_h^*}} \bar{x}'_{n_h^{'}}$

where $\bar{x}'_{n_h^*} = \sum_{i=1}^{n_h^*} x_{hi}^{'} / n_h^{*}$ is the mean of $x_{hi}^{'}$ defined on $S\left(n_h^{*}\right)$,

$$\bar{y}_{n*_h}^{'} = w_{1h}^{*} \bar{y}_{1n*_h}^{'} + w_{2h}^{*} \bar{y}_{2m*_h}^{'} \tag{4.1}$$

$w_{1h}^{*} = n_{1h}^{*} / n_h^{*}$ *and* $w_{2h}^{*} = n_{2h}^{*} / n_h^{*}, \bar{y}_{1n_h^*} = \sum_{i=1}^{n_{1h}^*} y_{hi}^{'} / n_{1h}^{*}$ is the mean of $y_{hi}^{'}$ for the n_{1h}^{*} respondents and $\bar{y}_{2m*h}^{'} = \sum_{i=1}^{m_h^*} y_{hi}^{'} / m_h^{*}$ is the mean of $y_{hi}^{'}$ for the subsample of m_h^{*} nonrespondents. As it is well known, (4.1) is unbiased for

$\bar{y}_{n*_h}^{'} = w_{1h}^{*} y_{1n*_h}^{'} + w_{2h}^{*} y_{2n*_h}^{'}$ where

$\bar{y}'_{2n_h^*} = \left(\sum_{i=1}^{n_{2h}^*} y'_{hi} \right) / n_{2h}^*$ is the mean of y'_{hi} for the n_{2h}^* nonrespondents.

Let $\bar{y}'_{n_h^*} = \bar{Y}'_h + \bar{y}'_{n_h^*} - \bar{Y}'_h = \bar{Y}'_h \left(1 + \delta\bar{y}'_{n_h^*}\right)$, where $\delta\bar{y}'_{n_h^*} = \left(\bar{y}'_{n_h^*} - \bar{Y}'_h\right) / \bar{Y}'_h$. Similarly,

$\bar{x}'_{n'_h} = \bar{X}'_h \left(1 + \delta\bar{x}'_{n'_h}\right)$ where $\delta\bar{x}'_{n'_h} = \left(\bar{x}'_{n'_h} = \bar{X}'_h\right) / \bar{X}'_h$, and $\bar{x}'_{n'_h} = \bar{X}'_h \left(1 + \delta\bar{x}'_{n'_h}\right)$, where $\delta\bar{x}'_{n_h^{*'}} = \left(\bar{x}'_{n_h^*} = \bar{X}'_h\right) / \bar{X}'_h$. Then to a first approximation.

$$t_{3j} - \bar{Y}_j = \sum_h W_h \bar{Y}_h \left(\delta\bar{x}'_{n'_h} + \delta\bar{y}'_{n_h^*} + \delta\bar{y}'_{n'_h} - \delta\bar{x}'_{n'_h} - \delta\bar{x}'_{n_h^*}\right)\left(1 - \delta\bar{x}'_{n_h^*}\right) \tag{4.2}$$

and hence

$$B(t_{3j}) = E(t_{3j} - \bar{Y}_j) \approx \sum_h W_h \left(\frac{1}{n_h^*} - \frac{1}{n_h} \right) \frac{1}{\bar{X}'_h} \left(R'_h S^2_{x'_h} - S_{x'_h y'_h} \right) \tag{4.3}$$

Substitution from (3.8) and (3.10) in (4.3) gives the result

$$B(t_{3j}) = \sum_h W_h \frac{\left(1 - \frac{n_h^*}{n'_h}\right)}{n_{hj} P_{hj} \bar{X}_{hj}} \frac{(NW_{hj} - 1)}{NW_h - 1} \left[R_{hj} S^2_{x(hj)} - S_{xy(hj)} + \frac{NW_{hj}}{NW_{hj} - 1} Q_{hj} R_{hj} \left(\bar{X}_{hj} - \bar{X}_j\right)^2 \right]$$

By ignoring terms in the expansion of (4.2) with powers higher than one, it can be shown that the MSE of t_{3j} is

MSE(t_{3j}) =

$$\sum_h W_h^2 \left\{ \left(\frac{1}{n_h^*} - \frac{1}{N_h} \right) S^2_{y'_h} + \left(\frac{1}{n_h^*} - \frac{1}{n'_h} \right) \left(R'^2_h S^2_{x'_h} - 2R'_h S_{x'_h} y'_h \right) + \frac{w_{2h}(k-1)}{n_h} S_{2y'_h} \right\} \tag{4.4}$$

Substitution from (3.8) and (3.10) in (4.4) gives the result.

$$\text{MSE}(t_{3j}) = \sum_h W_h^2 \left\{ \frac{NW_{hj} - 1}{NW_h - 1} \left[\frac{1 - f_h}{n_h} \left(S^2_{y(hj)} + \frac{NW_{hj}}{NW_{hj} - 1} Q_{hj} (\bar{Y}_{hj} - \bar{Y}_j)^2 \right) + \right. \right.$$

$$\left. + \left(\frac{1}{n_h^*} - \frac{1}{n'_h} \right) R_{hj} \left(R_{hj} S^2_{x(hj)} - 2S_{xy(hj)} + \frac{NW_{hj}}{NW_{hj} - 1} Q_{hj} R_{hj} (\bar{X}_{hj} - \bar{X}_j)^2 \right] .$$

$$\left. + \frac{W_{2h}(k-1)}{n_h} \frac{(N_2 W_{2hj} - 1}{(N_2 W_{2h} - 1)} \left(S^2_{2y(hj)} + \frac{N_2 W_{2hj}}{(N_2 W_{2hj} - 1)} Q_{2hj} (\bar{Y}_{2hj} - \bar{Y}_{2j})^2 \right) \right\}$$

where $f_h = n_h^{'} / N_h$. An estimator of MSE(t_{3j}) is obtained by replacing W_h, W_{2h}, W_{hj}, W_{2hj}, Q_{hj}, Q_{2hj}, $\overline{Y}_{hj}, \overline{Y}_j, \overline{Y}_{2hj}, \overline{Y}_{2j}, S^2_{y(hj)}, S^2_{2y(hj)}$ by w_h, w_{2h}, w_{hj}, w_{2hj}, q_{2hj}, $\overline{y}_{hj}, \overline{y}_j, \overline{y}_{2hj}, \overline{y}_{2j}$ respectively and $S_{xy(hj)}$ by $s_{xy(hj)}$ =

$$\sum_{i=1}^{n_{hj}} \left(x_{hij} - \overline{x}_{hj}\right)\left(y_{hij} - \overline{y}_{hj}\right) / \left(n_{hj} - 1\right).$$

Remark

The biases and mean-square-errors of t_1 and t_2, which are global estimators, depend on components of the total population variance whereas those of t_{1j}, t_{2j} and t_{3j} depend on components of variance for domain j that is of interest. The biases and mean-square-errors of t_{1j}, t_{2j} and t_{3j} also depend on the variability of the subpopulation of domain j in the different strata and are hence higher in magnitude than those of the corresponding global estimators. The high increase in the biases and mean-square-errors of t_{1j}, t_{2j}, t_{3j} is associated with the efficiency of the initial stratification which accounts for the large contribution from $(\overline{Y}_{hj} - \overline{Y}_j)^2$ and $(\overline{X}_{hj} - \overline{X}_j)^2$. The method adopted minimizes, to a certain extent, the problem of obtaining a reliable sampling frame which can be serious in many developing countries. The comparison between t_1 and t_2 made by Rao (1986) can also be extended to their corresponding domain estimators.

Acknowledgement

The author expresses his sincere gratitude to Professor P. S. R. S. Rao who suggested the topic in response to the author’s request in 1988 for materials to support his Ph.D thesis on domain estimation. The author also expresses his profound gratitude to Professor M. P. Singh who introduced Statistics in Transition to him.

REFERENCE

DURBIN, J. (1958). Sampling theory for estimators based on fewer individuals than the number selected. Bull. Int. Stat. Inst., 36, 3, 113—119.

HANSEN, M.H. and HURWITZ, W. N. (1946). The Problem of nonresponse in sample surveys. Journ. of Amer. Stat. Assoc., 41, 517—529.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). Sampling Survey Methods and Theory, Vol. II, John Wiley and Sons, Inc., New York, N.Y.

HARTLEY, H. O. (1959). Analytic Studies of Survey data. Institudo di Statisca, volume in honor of Corrado Gini, Rome.

RAO, Poduri S. R. S. (1986). Ratio Estimation with subsampling the nonrespondents. Survey Methodology, 12, 217—230.

TIN, M. and Than Toe (1972). Estimation for domains in multistage sampling. Journ. of Amer. Stat. Assoc., 67, 913—916.

UDOFIA, G. A. (1992). Contributions to the theory of response bias and domain estimation in double sampling. Unpublished Ph.D. thesis, University of Ibadan.

U.N.Statistical Office (1950**)**. The Preparation of sample survey reports. Stat. Papers Series C. No. 1.

YATES, F. (1953). Sampling methods for censuses and surveys. Charles Griffin and Co. London. Third edition. 1960.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 725—754

CONSIDERATION ON OPTIMAL SAMPLE DESIGN FOR SMALL AREA ESTIMATION[1]

Grażyna Dehnel, Elżbieta Gołata and Tomasz Klimanek[2]

ABSRACT[3]

The paper focuses on small area estimation with two-stage sampling, with special emphasis on choices that need to be made about the levels of stratification and clustering. Alternative ways of clustering, more and less detailed, are considered and an attempt is made to choose a method which is most suitable for small area estimation (given a fixed budget for the survey). The study tests the empirical impact of the number and size of clusters on the characteristics of direct, synthetic and composite EBLUP estimators. The optimal sample allocation for a two-stage-design, in terms of domains, is found to be very close to the optimal sample allocation from the population point of view. The gains in the small area estimation are compared with the losses in the precision of the population mean estimator. In general ,finer stratification is preferred. In particular, stratification that coincides with the domains of interest is the finest stratification that is practicable.

***Key words*:** Small area estimation, two-stage sampling, optimal sample allocation.

1. Introduction

The present study deals with selecting sampling designs that are efficient in small area estimation. The problem is a difficult one to be solved due to many optimisation problems which arise in survey design, and need to be considered. According to Rao J.N.K., (2003) the most important design issues for small domain estimation are the following: the number of strata, the construction of strata, the optimal allocation of a sample, selection probabilities. This list can be

[1] The research was conducted within the EURARA project no. IST-2000-26290 entitled *Enhancing Small Area Estimation Techniques to Meet European Needs,* a part of 5th framework programme for research, technological development and demonstration of EU. Its main co-ordinator is Office for National Statistics, UK.

[2] g.dehnel@ae.poznan.pl, elzbieta.golata@ae.poznan.pl, t.klimanek@ae.poznan.pl, The Poznan University of Economics, Poznan, Poland.

[3] We would like to thank the Editor and the Referees for their insightful comments and suggestions.

enlarged by adding the problem of defining the optimisation criteria, possibilities of obtaining strongly correlated auxiliary information, the choice of estimators taking into account their efficiency under specific sampling designs.

The ideal goal is to find an "optimal" design that minimizes the MSE of a direct estimator subject to a given cost. This goal is seldom achieved in practice due to operational constraints and other factors. As a result, a "compromise" design that is close to the optimal one is adopted. In practice, it is not possible to anticipate and make plans for all small areas. Because some domains have no representation in the sample, indirect estimators will always be needed. Given the growing demand for reliable small area statistics, it is important to consider design issues that have an impact on small area estimation, particularly in the context of planning and designing large-scale surveys. Although optimal sample design for whole-population statistics is well understood (Särndal et al 1992), there has been relatively little work on designs that optimise small area estimates.

Taking into account the main sample design issues as presented by Rao, in our research conducted within the EURAREA project we focused on empirical investigation of two of the above mentioned problems: sample allocation and clustering.

Rao (2003 p.22) gives examples of two-step compromise sample allocations which satisfy reliability requirements at small area level (as well as at large area level), using only direct estimates. These examples are: Canadian LFS (Singh et al., 1994), Canadian Community Health Survey (Béland, Bailie, Catlin, Singh, 2000), U.S. National Household Survey on Drug Abuse and the 2000 Danish Health and Morbidity Survey. The general idea is to allocate a sample in two steps: the first step involves getting reliable provincial estimates and allocating the remaining sample in the second step to produce the best possible estimates at domain level. By oversampling small areas, it is possible to significantly decrease the CV of direct estimates for these areas at the expense of a small increase in CV at the national level.

Singh, Gambino and Mantel (1994) described how the sampling design of the Canadian Labour Force Survey was adjusted to cater for small area statistics. Every month a sample of 59.000 households was drawn and these households remained in the survey for six months. The sample was allocated in two steps; first, 42.000 households in the sample were allocated with the aim to optimise the estimation of the regional and provincial parameters. The remainder, 17.000 households, was assigned so as to optimise estimation at small area (sub-provincial) level. Singh *et al.* (1994) documented the substantial gains in precision for the least populous areas, at the cost of a slight loss of efficiency for the largest areas at the same level as well as for the provinces and for the whole country.

The first part of our investigation concerned the distribution of the sample among different sized areas — assuming that all areas are included in the sample — subject to a constraint on the total sample size. We tried to check the empirical impact of the sample allocation on the characteristics of direct, synthetic and composite EBLUP estimators for both types of small area units considered in the EURAREA project: NUTS3 and NUTS4. The study revealed that the EBLUP estimation under a sampling design optimal for the composite estimator was the optimal arrangement for small area estimation, with the actual results being described in a separate report (Longford at al., 2004).

The following sections of the article present the results obtained in the second part of our investigation, which addresses the problem of clustering. Clustering is often used in order to reduce survey costs, but it also brings about a decrease in the "effective" sample size. This affects estimation for unplanned domains because it can lead to situations where some domains become sample rich while others may have no samples at all. Therefore Rao (2003 p. 22) suggests minimizing the clustering in the sample. He also underlines the importance of the choice of a sampling frame, sampling units, their sizes and the number of sampling stages. According to Rao (2003) another method of providing better sample size distribution for small areas is to replace a large stratum with many small strata. This approach results in unplanned small domains containing mostly complete strata (see also Marker 2001 p.183—184).

The present study tested the empirical impact of the number and size of clusters on the characteristics of direct, synthetic and composite EBLUP estimators. Two different situations can be considered:

(a) The clustering takes place at a lower level than the estimation.

This case is discussed in sections 1—4 and the results obtained show that the allocation that was optimal for small areas was also optimal (more or less) for global scale — for larger areas.

(b) The clustering takes place at a higher level than the estimation or the area level for which estimates are to be provided is also the PSU level.

This case is discussed under section 5. This type of clustering results there being no sample data at all for most target areas — a typical situation for small area statistics (this case is very important in England, Spain and Italy).

Empirical estimation was applied to small areas at NUTS3 and NUTS4 levels with respect to the population of Poland. The database, called POLDATA, was prepared specifically for the EURAREA project[1]. Its construction draws on three sources: the 1995 micro-census in Poland, the 1995 Polish Household Budget Survey and the Polish Local Data Bank. In Poland, there are 44 NUTS3-level and 373 NUTS4-level areas. The target variable is a household income. Its

[1] For the purpose of the EURAREA validation program, a special database has been set up. The Polish database — the so-called super-population labelled POLDATA provides real information about the target variables (income, household structure, unemployment rate) and represents as closely as possible the characteristic of Poland in 1995 with respect to the new administrative division of the country which was introduced in January 1999.

within-area means are estimated by fitting a two-level model with two covariates, age and sex.

Initially, the simulations were based on a population specially adjusted for this study. The problem was to obtain a population containing domains of equal size for reductions of weighting in the formulas to estimate variance at local level. Our population was constructed as follows: an identical number of units of lower territorial division was chosen per unit of a higher level of territorial aggregation. This number of units of lower per higher division was set against the median number observed in the POLDATA, that is:

	Min	Max	Mean	**Median**
Dwellings per census district	1	135	21.0	**22**
Census districts per commune (NUTS5)	1	279	11.0	**7**
NUTS5 per NUTS4	1	19	6.5	**6**
NUTS4 per NUTS3	1	16	8.6	**9**

Then, dwellings were randomly assigned to census districts, districts to communes and so on, obtaining:

Total number of dwellings: 365 904 = 22 × 7 × 6 × 9 × 44
Total number of census districts 365 904 / 22 = 16 632 = 7 × 6 × 9 × 44
Total number of communes 16 632 / 7 = 2 376 = 6 × 9 × 44
Total number of NUTS4 2 376 / 6 = 396 = 9 × 44
Number of NUTS3 396 / 9 = 44

The basic characteristics of the population in relation to income (the target variable) are as follows

Territorial unit	Number of dwellings per unit	Minimum value of the target variable	Maximum value of the target variable	Median income	Within-area variance	Between-area variance	Ratio of the between and within area variances $\varpi = \frac{\sigma_B^2}{\sigma_W^2}$
DISTRICT	22	186.2	1553.3	392.1	112 365	10 209	0.0909
NUTS 5	154	259.2	700.5	396.5	117 850	4 725	0.0401
NUTS 4	924	297.4	620.2	396.6	119 549	3 026	0.0253
NUTS 3	8316	344.8	485.3	402.7	121 541	1 033	0.0085

Source: Own calculations made within the EURAREA project based on POLDATA database, April 2003

The mean income in the population was equal to 407 PLN and the variance: 122 574.5. The standard deviation amounted to 350, which constitutes over 86% of the mean.

The approach, based on a specially adjusted population, raised doubts as to whether the results obtained for the artificial population are also valid for a real population. Consequently, another section (section 5) was added to the paper containing results of an analogous study conducted on real populations. In this part, the overall resources (funds) are represented by a fixed sample size. It was assumed that the constraint on the total sample size could be treated equivalently to the overall cost.

In the study we focused on the following problems:

1. How to incorporate costs in the decision?
2. Which clustering to choose? What kind of PSU (more or less detailed division) is more desirable for small area estimation?
3. How do the changes in inclusion probabilities influence estimation precision from the global and local point of view?
4. Are the regularities observed for the direct estimator also valid for the basic types of indirect estimators under the two-stage design?
5. What are the properties of indirect estimators under the two-stage design in a real population?

In order to fulfil the task, two-stage samples were drawn of a different number of PSU and a number of dwellings at the second stage which was adjusted in the following way. The overall sample size remained unchanged (in order to increase the probability of selecting PSU (p_{1i}), the probability of selecting dwellings in the selected PSU was automatically decreased ($p_{i|j}$)). A set of different probability combinations was applied. For each of the distinguished combinations of inclusion probabilities 100 samples were drawn and then processed with the ONS software for standard estimators[1]. The results obtained were analysed empirically.

The optimal clustering for a direct estimator was compared from the local and global point of view. At national level, the object was to find a combination giving the smallest empirical variance whereas at the local level the criterion function was defined as the weighted total of empirical variances

$$S_V = \sum_d N_d^q \operatorname{var}(\hat{\bar{Y}}_d). \tag{1}$$

As the criterion based on empirical variances is correct for direct estimators, but contains only part of the error in the case of the EBLUP and synthetic estimators, the empirical MSE was used as well (section 5). So apart from $S_V = \sum_d N_d^q \operatorname{var}(\hat{\bar{Y}}_d)$, it was necessary to use the criterion function defined as the weighted total of empirical MSE:

$$S_{MSE} = \sum_d N_d^q MSE(\hat{\bar{Y}}_d). \tag{2}$$

[1] All the estimators were calculated using a special software code in SAS prepared by the British Office for National Statistics (ONS).

2. Incorporating costs in the decision

Funding is often the main factor in choosing a sample design. To obtain an optimal allocation of the sample, one may either try to minimise the variance (or MSE) of an estimator for fixed resources or conversely one may tend to minimise the costs for fixed variance. There is a trade-off between the design chosen and costs of the survey. The costs of sampling represented by the constant and flexible costs (of selecting a unit per stratum, or PSU, constructing a frame, listing costs per element in a stratum or in a sampled cluster, obtaining the desired information etc ...) are crucial in optimal allocation procedures.

For the purposes of the study the assumption was that the budget for the survey was to be equal to 100 000 PLN. In the analysis, the sampling costs for different definitions of PSU were set to be as presented in Table 1. In the case of a two-stage design, the problem is to determine the sample size according to different definitions of the primary sampling unit (PSU) and inclusion probabilities, with respect to fixed resources. We assumed that the constant costs would amount to $C_0 = 6872$. Consequently, the amount of 93 128 PLN was divided between the sampling of primary and secondary sampling units according to different inclusion probabilities.

Table 1. Different types of clustering (PSU) and assumed sampling costs

Type of clusters	Number of clusters in the population	Cost of a survey PLN
NUTS3	44	22
NUTS4	396	20
NUTS5 — B	**2376**	C_{1B} **18**
Census district — A	**16632**	C_{1A} **16**
Dwellings	365904	2

Remark, types of clusters examined in the research are shaded

Source: Own calculations made within the EURAREA project based on POLDATA database

The costs function may be written as:

$$\hat{C} = C_0 + mC_1 + C_2 \sum_{g=1}^{m} n_{(g)} \tag{3}$$

where:

m — number of PSUs (NUTS3 or NUTS4);

$n_{(g)}$ — number of SSUs (dwellings) in the g-th PSU;

C_1 — cost of sampling the PSU (for districts $C_{1A} = 16\ PLN$, and for NUTS5 $C_{1B} = 18\ PLN$);

C_2 — cost of sampling units at the second stage (dwellings — in our case). It is assumed that this cost is not influenced by the type of PSU and is settled to be equal to 2 PLN;
C_0 — constant cost, which in a two-stage sampling was assumed to be equal to 6872 PLN.

3. Which clustering to choose? Checking what kind of PSU (more or less detailed division) is more desirable for small area estimation

In practice, there is likely to be only a small number of alternative ways of clustering in a sampling design. In fact, it was necessary to choose between two ways and determine which one would be preferable for small area estimation, given a fixed budget for the survey. Clustering involves a partition of the country into primary sampling units (PSUs). Suppose ***clustering A*** is more detailed than ***clustering B***, that is, one cluster or a group of clusters in ***A*** form clusters in ***B***. We assume that ***B*** may be the NUTS5 units of the country, and ***A*** — census districts, such that each commune — NUTS5 unit consists of a set of census districts.

The costs of sampling might be split into costs per subject and costs per cluster. Let us assume that sampling in each cluster involves set-up costs of C_1 ($C_{1\mathrm{A}}$ with ***clustering A*** and $C_{1\mathrm{B}}$ with ***clustering B***), and costs of contacting and interviewing C_2 which in our case are defined per dwelling (it is assumed that these do not depend on clustering). Thus, the cost of a survey with n_{I} clusters and n subjects is $C_1 n_{\mathrm{I}} + C_2 n$. With more detailed clustering, the set-up costs are probably lower, $C_{1\mathrm{A}} < C_{1\mathrm{B}}$, but pro-rated to subjects, they are probably higher — $C_{1\mathrm{A}} n_{\mathrm{IA}} > C_{1\mathrm{B}} n_{\mathrm{IB}}$ if there are more sampled clusters in ***clustering A*** than in ***clustering B***. The within-cluster sample sizes (their average is n / n_{I}) will tend to be smaller with the more detailed ***clustering A***.

Examining the impact of the way in which the population was divided into clusters on estimation efficiency, in global and local terms, we assumed that the resources are fixed. Optionally, we determined the inclusion probabilities for PSU and their aggregation level: districts or NUTS5. Having settled this, we obtained the probability of selecting units at the second stage with respect to constant costs. It was difficult to obtain an ideal solution, because the fractions did not give integer numbers of PSU or within-cluster sample size to be selected. It seemed appropriate to round up those numbers in such a way that the total sampling costs were not exceeded. Consequently, in some cases the overall sample size was decreased, which also resulted in appropriate changes in estimation precision. Inclusion probabilities and the sample sizes for different strategies applied to two-stage sampling with respect to fixed resources are presented in Table 2.

- For the two types of PSUs we increased the probability of selecting PSU (p_{1i}) and automatically decreased the probability of selecting dwellings in the selected PSU ($p_{j|i}$) with respect to fixed resources.
- For the same probability p_{1i} of inclusion PSU (and respectively adjusted $p_{i|j}$) the size of the sample selected is greater for bigger clusters — NUTS5 than in the case of smaller clusters — census districts.
- For smaller clusters the dispersion in the sample size selected for the same range of p_{1i} is much bigger than in the case of NUTS5. For inclusion probabilities of selecting PSU from an interval (0.13—0.2) the sample size obtained for clusters defined as census districts changes from 5.45% to 7.95%, while for NUTS5 as clusters, it ranges from 11.55% to 11.93%.

Table 2. Inclusion probabilities and the sample sizes for different strategies applied to two-stage sampling with respect to fixed resources

Inclusion probabilities p_{ij}	Probability of selecting		Number of PSU selected m	Within-cluster sample size	Sample size n	Costs used	Money left
	PSU P_{1i}	Dwellings in selected PSU $P_{j\|i}$					
Clustering A		**Census district**					
0.0909	0.100	0.909	1663	20	33260	93128	0
0.0755	**0.128**	**0.591**	**2124**	**13**	**27612**	**89208**	**-3920**
0.0795	0.125	0.636	2079	14	29106	91476	-1652
0.0682	0.150	0.455	2494	10	24940	89784	-3344
0.0636	0.175	0.364	2910	8	23280	93120	-8
0.0545	0.200	0.273	3326	6	19956	93128	0
Clustering B		**Commune- NUTS5**					
0.1193	0.125	0.955	297	147	43659	92664	-464
0.1177	0.150	0.786	356	121	43076	92560	-568
0.1168	0.175	0.669	415	103	42745	92960	-168
0.1155	0.200	0.578	475	89	42275	93100	-28
0.1138	0.225	0.506	534	78	41652	92916	-212
0.1120	0.250	0.448	594	69	40986	92664	-464
0.1106	0.275	0.403	653	62	40486	92726	-402

Inclusion probabilities p_{ij}	Probability of selecting		Number of PSU selected m	Within-cluster sample size	Sample size n	Costs used	Money left
	PSU P_{1i}	Dwellings in selected PSU $P_{j\|i}$					
0.1090	0.300	0.364	712	56	39872	92560	-568
0.1076	0.325	0.331	772	51	39372	92640	-488
0.1067	0.350	0.305	831	47	39057	93072	-56
0.1047	0.375	0.279	891	43	38313	92664	-464
0.1039	0.400	0.260	950	40	38000	93100	-28
0.0974	0.500	0.195	1188	30	35640	92664	464
0.0896	0.600	0.149	1425	23	32775	91200	1928
0.0864	0.700	0.123	1663	19	31597	93128	0
0.0779	**0.800**	**0.097**	**1900**	**15**	**28500**	**91200**	**1928**
0.0701	0.900	0.078	2138	12	25656	89796	3332

Remark — denotes an optimal allocation in order to minimize the sampling variance at population level,

Source: Own calculations made within the EURAREA project based on POLDATA database

It is worth noting that for smaller probabilities p_{1i} at the first stage and greater probabilities $p_{j|i}$ at the second stage, the sample size increases for the same resources used. Clearly, when the costs of selecting PSU are reduced (smaller probabilities of inclusion), bigger resources can be spent on selecting dwellings, which results in a bigger sample size. Nevertheless, this solution does not seem to be the most desirable, not only for small area estimation. As for the lower number of PSU selected, there might be no representation for a large number of domains.

In order to answer what kind of PSU (more or less detailed division) and which way of allocation is more desirable for small area estimation, we started with the solution offered by traditional survey sampling (Bracha, 1996 p.146—150). The traditional survey sampling approach was applied in order to determine which combination of inclusion probabilities would be preferable for estimation on a population scale. The optimal sample allocation for two-stage sampling involves searching for such fractions at both stages:

$$f_1 = \frac{m}{M} \text{ and } \tag{4}$$

$f_2 = \frac{n_i}{N_i}$, in order to minimize the variance of the estimator at population level (5)

$$D^2(y) = \frac{1}{N^2}\frac{M^2}{m}\left[(1-f_1)S_1^2 + \frac{1-f_2}{f_2}\overline{N}S_2^2\right], \tag{6}$$

where:

$$S_1^2 = \frac{1}{m-1}\sum_{g=1}^{m}(y_g - \overline{y})^2 \tag{7}$$

$$S_2^2 = \frac{1}{N}\sum_{i=1}^{M} N_i S_{2i}^2, \tag{8}$$

where:

$$S_{2i}^2 = \frac{1}{N_i - 1}\sum_{i=1}^{N_i}(Y_{ij} - \overline{\overline{Y}}_i)^2 \tag{9}$$

and

$$\overline{\overline{Y}} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} Y_{ij} \tag{10}$$

Assuming that the expression (12) is greater than 0, there exists an optimal solution that can be given in the following way:

$$f_{2\,opt} = \sqrt{\frac{C_1}{C_2\kappa}} \text{ and where:} \tag{11}$$

$$\kappa = \frac{S_1^2}{S_2^2} - \overline{N} \tag{12}$$

$m_{opt} = (\hat{C} - C_0)(C_1 + C_2\overline{N}f_{2\,opt})^{-1}$ for the two types of clustering (***A*** and ***B***, more and less detailed) and dwellings at the second stage we obtained: (13)

A) For clusters defined as census district, (more detailed clustering denoted as ***clustering A***)

$$f_{1A} = \frac{m}{M} = 0.1277 \quad \text{and} \quad f_{2A} = \frac{n_i}{N_i} = 0.5909.$$

B) For NUTS5 defined as clusters (less detailed division denoted as ***clustering B***)

$$f_{1B} = \frac{m}{M} = 0.771 \text{ and } \quad f_{2B} = \frac{n_i}{N_i} = 0.1040.$$

In both cases the optimal allocation results in a similar sample size equal to

$\pi_{ij\,A} = f_{1A} f_{2A} = 0.1277 \cdot 0.5909 = 0.755$ and

$\pi_{ij\,B} = f_{1B} f_{2B} = 0.771 \cdot 0.1040 = 0.8$.

This optimal solution[1], from the "global" point of view, is shadowed in Table 2.

4. How do the changes in inclusion probabilities influence estimation precision from the global and local point of view?

The question of optimisation criteria will always be debatable, because there is no obvious choice. The criteria applied most often for optimisation are costs and variance. In this case, optimisation would mean searching for an allocation of the sample that minimises the variance at a given cost. In other words, optimising a criterion for small area estimation involves choosing a design and an estimation technique that minimises a measure being a function of estimation precision for all domains (at fixed resources). This means determining how well on average the areas are estimated: the total of area specific variances, the total of area specific values of MSE, the total of area specific values of absolute relative bias etc.

Other topics which need further consideration, are:

- the importance of the objectives of the survey,
- the influence of auxiliary variables on the optimisation problem,
- the trade-off between national- and small area level properties.

In our previous study[2] we applied the criterion function defined as the weighted total of sampling variances $S_V = \sum_d N_d^q \operatorname{var}(\hat{\bar{Y}}_d^{DIRECT})$, for the powers $q \in (0,2)$ to measure efficiency for domains. We started with a naive approach searching for such an allocation of the sample which would minimise the criterion function defined for a direct estimator (the bias of the estimator is expected to be 0, so MSE is equal to the variance, see annex 1).

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} y_{id} \tag{14}$$

[1] The solution presented in the table is not exactly the optimal one, but very close to it. This way of presentation results from the approach of determining the inclusion probabilities for PSU, applied in searching for their impact on the estimation efficiency.

[2] See: Gołata E., Klimanek T., *Small-area estimation with complex sampling design; Initial results on optimal allocation of the overall sample size to the small areas*, October 2002, Internal Report prepared within Workpackage 4.2 of EURAREA project no. IST-2000-26290.

also investigated the impact of a change in the power *"q"* on the optimal sample allocation n_d^* in the case of the simple random sampling. (The simple total of sampling variances is smaller than the weighted sum, when the powers of the weight increase in value.) The increase in the criterion function was rapid for $q > 1$. Although it is debatable whether to weight the evaluation parameters for small domain estimation or not, the approach is usually recommended when the domains differ much [see *Standard Performance Criteria*, 2001]. Usually the weight is defined as the domain size in population N_d ($q = 1$ in the criterion function).

Typically, higher precision is desired for more populous areas, although the differences in precision should be reduced in relation to the population sizes. There is a trade-off between equal relative representation $RN_d = \frac{n_d^*}{N_d}$, equal estimation precision and the desire to estimate the area-level means with precision values that are related to the population sizes N_d. Keeping that in mind, the value of the power "*q*" promoting greater relative representation for smaller domains, was assumed: $q=0.5$. In the case considered in sections 1—4, neither the weight nor the power "*q*" is taken into account. They are of no importance for the adjusted population contains areas of equal size. In section 5 the value of $q=0.5$ was applied.

The efficiency of estimation in global and local scale was examined for 100 samples drawn from the population according to a two-stage design (different settings of clusters and inclusion probabilities, see Table 2). This yielded direct estimator of the mean income at population level and for domains, which in our experiment were defined as NUTS3. We applied the formulas for the mean and the sampling variance under two-stage sampling (with simple random sampling at each stage) for global and local scale as described by Särndal *at al* (1992 p. 137) and Bracha (1996 p.140—145 and 252—255).

By obtaining direct estimates and their precision under the discussed combination of clusters and changing inclusion probabilities we wanted to determine:

a) Which type of clustering is more efficient for small areas: more or less detailed — ***clustering A*** or ***clustering B***?
b) Is the improvement observed in estimation precision for small domains accompanied by a similar change in precision on a global scale? Is there any loss in estimation precision on a global scale observed along with improvement for local scale? If so, how large is it?
c) On the basis of empirical results obtained, what is the relation between estimation precision for global and domain scales and also for other performance criteria such as: bias, MSE, Relative Estimation Error REE.

Ad. a) Which type of clustering?

Measuring the estimation precision for domains with the above defined criterion function S_V, it can be noted that for more detailed ***clustering A***, it ranges from 10,000 up to 12,500 (see Table 3). While for less detailed ***clustering B*** and growing inclusion probabilities ***P_{1i}***, it decreases from almost 25,000 to 8,500. The estimation precision at population level is higher than for small domains, which results in the variance assuming values of 3—14. The range for ***clustering A*** is narrower for both functions. For ***clustering B*** the variance increases for decreasing inclusion probabilities ***P_{1i}***. For less detailed clustering, there are more possibilities of setting inclusion probabilities at both stages for given resources, so the relationship between precision at population and local levels is more explicit. According to the value obtained by the criterion function, one can note that it is smaller for less detailed clustering, which implies higher estimation precision — on average for all areas.

Table 3. Estimation precision at global and local scale in relation to inclusion probabilities and relative sizes of the within- and between-area variances, two-stage sampling and different types of clustering with respect to fixed resources

Probability of selecting PSU P_{1i}	Value of criterion function S_V	Sampling variance $\hat{V}(\hat{\bar{Y}})$	omega $\varpi = \frac{\sigma_B^2}{\sigma_W^2}$
Clustering A			
0.100	**10063**	6.11	0.0921
0.125	**11203**	**5.95**	**0.1125**
0.150	10794	6.00	0.1418
0.175	11646	7.33	0.1674
0.200	12444	8.22	0.2083
Clustering B			
0.125	24777	13.72	0.0403
0.150	21342	13.76	0.0421
0.175	18813	10.37	0.0426
0.200	16085	7.97	0.0444
0.225	14391	7.91	0.0465
0.250	13475	8.01	0.0479
0.275	12284	6.90	0.0498
0.300	11831	6.90	0.0514
0.325	10860	5.29	0.0531
0.350	10318	6.77	0.0544
0.375	9885	4.75	0.0567
0.400	9411	5.90	0.0581

Probability of selecting PSU P_{1i}	Value of criterion function S_V	Sampling variance $\hat{V}(\hat{\bar{Y}})$	omega $\varpi = \frac{\sigma_B^2}{\sigma_W^2}$
0.500	8140	4.21	0.0670
0.600	7916	4.45	0.0774
0.700	**7605**	4.26	0.0864
0.800	**8212**	**3.31**	**0.1000**
0.900	8402	4.52	0.1165

Source: Own calculations made within the EURAREA project based on POLDATA database

Ad. b) Is the improvement observed in estimation precision for small domains accompanied by a similar change in precision on a global scale?

A strong positive relation was observed between the criterion function for domains and the sampling variance at population level. The relation observed was stronger for ***clustering B*** $r\left(S_V, \hat{V}(\hat{\bar{Y}})\right) = 0.97$ than for ***clustering A*** $r\left(S_V, \hat{V}(\hat{\bar{Y}})\right) = 0.86$. This difference might depend not only on the size of PSU but also on the number of observations (combinations) considered (5 for ***clustering A*** in comparison with 17 for ***clustering B***). Thus, one can conclude that the method of sample allocation which results in reducing direct estimator variance at the national (population) level is also appropriate for optimal sample allocation at the domain level.

Ad. c) Relation between estimation precision for global and domain scale and for other performance criteria

Apart from the variances at domain and population level, we also calculated other empirical measures characterising estimation efficiency (absolute relative bias $A\hat{R}B_d$ and relative estimation error $R\hat{E}E(\hat{\bar{Y}}_d)$).

The results obtained for the distinguished set of possible sample allocations for two-stage design (with different clustering) are very similar to the ones obtained to evaluate estimation precision (see Table 4). It can be noted that both the smallest mean value of $\overline{A\hat{R}B_d}$ for domains and $\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$ were obtained for the sample allocation for which the smallest value of criterion function was observed.

In both cases concerning estimation precision and other measures of estimation efficiency, the optimal sample allocation for a two-stage-design, in terms of domains is very close to the optimal sample allocation from the population point of view. In both types of clustering taken into consideration, domains seem to require a slightly smaller probability of inclusion at the first stage combined with a small increase in probability at the second stage. Although the optimal allocation for domains is very close to the optimal allocation for the population, it might be interesting to measure the loss or gain in estimation precision due to different optimisation criteria (for population or for domains — see Table 5).

Table 4. Efficiency of estimation and bias at global and local scale, different strategies applied to two-stage sampling with respect to fixed resources

Probability of selecting PSU	Measures of estimation efficiency					
	Local scale			Global scale		
P_{1i}	$\overline{\hat{ARB}_d}$	$\overline{\hat{V}(\hat{\bar{Y}}_d)}$	$\overline{\hat{REE}(\hat{\bar{Y}}_d)}$	$\overline{\hat{ARB}}$	$\overline{\hat{V}(\hat{\bar{Y}})}$	$\overline{\hat{REE}(\hat{\bar{Y}})}$
Clustering A						
0.100	**2.91**	**229**	**3.69**	0.48	6.11	0.61
0.125	3.05	255	3.87	0.49	5.95	0.60
0.150	3.04	245	3.83	0.50	6.00	0.60
0.175	3.16	265	3.99	0.53	7.33	0.67
0.200	3.25	283	4.11	0.57	8.22	0.71
Clustering B						
0.125	4.65	563	6.54	0.72	13.72	0.91
0.150	4.17	485	5.28	0.71	13.76	0.91
0.175	3.89	428	4.95	0.65	10.37	0.79
0.200	3.61	366	4.61	0.54	7.97	0.69
0.225	3.46	327	4.40	0.56	7.91	0.70
0.250	3.30	306	4.25	0.54	8.01	0.70
0.275	3.22	279	4.07	0.53	6.90	0.65
0.300	3.16	269	3.97	0.50	6.90	0.65
0.325	3.00	247	3.81	0.46	5.29	0.57
0.350	2.92	235	3.70	0.53	6.77	0.64
0.375	2.87	225	3.63	0.44	4.75	0.54
0.400	2.82	214	3.56	0.48	5.90	0.60
0.500	2.64	185	3.33	0.41	4.21	0.52
0.600	2.58	180	3.26	0.45	4.45	0.52
0.700	**2.57**	**173**	**3.23**	0.41	4.26	0.51
0.800	2.64	187	3.33	**0.36**	**3.31**	**0.45**
0.900	2.67	191	3.39	0.39	4.52	0.53

Source: Own calculations made within the EURAREA project based on POLDATA database.

Table 5. Relation between measures of estimation efficiency and bias obtained under different sample allocation (population optimal / domain optimal), different types of clustering with respect to fixed resources

Type of clustering	Ratio of appropriate measures of estimation efficiency							
	$\frac{\overline{\left(\hat{ARB}_d\right)_G}}{\overline{\left(\hat{ARB}_d\right)_L}}$	$\frac{\overline{\hat{V}(\hat{\bar{Y}}_d)_G}}{\overline{\hat{V}(\hat{\bar{Y}}_d)_L}}$	$\frac{\overline{\hat{REE}(\hat{\bar{Y}}_d)_G}}{\overline{\hat{REE}(\hat{\bar{Y}}_d)_L}}$	$\frac{\overline{\hat{ARB}_G}}{\overline{\hat{ARB}_L}}$	$\frac{\overline{\hat{V}(\hat{\bar{Y}})_G}}{\overline{\hat{V}(\hat{\bar{Y}})_L}}$	$\frac{\overline{\hat{REE}(\hat{\bar{Y}})_G}}{\overline{\hat{REE}(\hat{\bar{Y}})_L}}$	$\frac{\omega_G}{\omega_L}$	$\frac{(S_V)_G}{(S_V)_L}$
Clustering A	1.0481	**1.1135**	1.0488	1.0208	0.9738	0.9836	1.2215	**1.1133**
Clustering B	1.0272	**1.0809**	1.0310	0.8780	0.7770	0.8824	1.1568	**1.0798**

Remark: subscript G stands for sampling design that proved to be optimal for *Global scale* and subscript L refers to *Local scale*

Source: Own calculations made within the EURAREA project based on POLDATA database

It can be concluded that a change in the sample allocation from optimal at population level, to "optimal" for domains might result in a gain in average estimation precision for domains. This might amount to about 11% of S_V for more detailed clustering (***clustering A***) and to almost 8% for less detailed (***clustering B***) and a loss of over 22% (***clustering B***) and 2.6% (***clustering A***) in variance $\hat{V}\left(\hat{\bar{Y}}\right)$ (estimation precision) at population level. This means that a more detailed clustering is more likely to gain precision for domains and less likely to lose precision at population level due to change in optimisation approach from global to local scale. It can be noticed that loss of 22% in estimation precision (variance) at global scale is equivalent to the increase in the value of $\overline{\hat{REE}(\hat{\bar{Y}})}$ from 0.45% to 0.51%. Other measures of estimation quality are not so sensitive (as variance) to the change in optimisation approach. In both types of clustering the change in optimality from global to local scale is connected with a decrease (16-22%) in the relative size of between-area variance.

5. Are the regularities observed for the direct estimator valid also for the basic types of indirect estimators under a two-stage design?

It might be interesting to learn about the results of applying indirect estimators to samples drawn according to two-stage sampling with different types of clustering. Many combinations of sample allocation were discussed, and the conclusion was that optimisation from the domain point of view is close to the one obtained for the population level. Therefore, it was deemed appropriate to apply the following indirect estimators (according to the EURAREA project) to global and local optimal allocation for both types of clustering and levels of territorial division NUTS3 and NUTS4:

- GREG with a standard linear regression model

$$\hat{\hat{Y}}_d^{GREG} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} y_{id} + \left(\overline{\mathbf{X}}_{.d} - \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} x_{id} \right)^T \hat{\boldsymbol{\beta}} \tag{15}$$

where:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in u_d} w_{id} x_{id} x_{id}^T \right)^{-1} \sum_{i \in u_d} w_{id} x_{id} y_{id} \tag{16}$$

- synthetic estimator considered under two different models:

a) a linear two-level model with individual data $y_{id} = x_{id}^T \boldsymbol{\beta} + u_d + e_{id}$

$$\hat{\hat{Y}}_d^{SYNTH} = \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}} \tag{17}$$

with $\overline{X}_{.d} = (\overline{X}_{.d,1}, \ldots, \overline{X}_{.d,p})^T$

b) a linear model with area-level covariates and a pooled sample estimate of within — area variance

$$\hat{\hat{Y}}_d^{SYNTH} = \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}\text{, with: } \overline{y}_{.d} = \overline{x}_{.d}^T \boldsymbol{\beta} + \xi_d\text{, } \xi_d \sim iid\ N(0, \sigma_u^2 + \frac{\sigma_e^2}{n_d})\text{,} \tag{18}$$

$$\hat{\sigma}_e^2 = \frac{1}{n - na} \sum_i \sum_d (y_{id} - \overline{y}_{.d})^2 \tag{19}$$

- EBLUP estimator using models:

a) a linear two-level model with individual data

$$\hat{\hat{Y}}_d^{EBLUP} = \gamma_d (\overline{y}_{.d} - \overline{\mathbf{x}}_{\mathbf{.d}}^{\mathbf{T}} \hat{\boldsymbol{\beta}}\ ) + \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}\ , \tag{20}$$

where:

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d}\ ; \tag{21}$$

b) a linear model with area-level covariates and a pooled sample estimate of within-area variance

$$\hat{\hat{Y}}_d^{EBLUP} = \gamma_d \hat{\hat{Y}}_d^{direct} + (1 - \gamma_d) \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}\ , \tag{22}$$

where:

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \tag{23}$$

The evaluation parameters for all areas at NUTS3 level are presented in Table 6 and for NUTS4 level in Table 7.

The results obtained in estimating income at NUTS 3 level seem to be quite good, as the mean value of REE ($\overline{\hat{REE}(\hat{\bar{Y}}_d)}$) does not exceed 7% for the worst estimator (in this case Synth_A). For both types of clustering: more and less detailed, the results obtained are similar, especially in relation to the relative measures. The differences observed are only slight. Somewhat smaller values are involved in ***clustering B*** rather than ***clustering A***.

Other known properties of indirect estimators are also valid. For example, the estimated bias of synthetic estimators is about 16-19 times as big as that for the direct estimator, but for EBLUP it is only 3 times as big. A notable variation of DIRECT and GREG estimators is also observed. Finally, the smallest mean REEs are observed for EBLUP estimators.

Table 6. Average values of evaluation parameters for local and global optimal sample allocation, indirect estimators, Income, NUTS 3 level, two-stage sampling, different types of clustering with respect to fixed resources

Estimator	$\overline{\hat{V}(\hat{\bar{Y}}_d)}$	$\overline{M\hat{S}E(\hat{\bar{Y}}_d)}$	$\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$	$\overline{A\hat{R}B_d}$	$\overline{\hat{V}(\hat{\bar{Y}}_d)}$	$\overline{M\hat{S}E(\hat{\bar{Y}}_d)}$	$\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$	$\overline{A\hat{R}B_d}$
Optimal sample allocation for domains					Optimal sample allocation for the whole population			
Clustering A								
Direct	247	249	0.0383	0.0031	255	258	0.0387	0.0036
Greg	248	251	0.0384	0.0031	256	259	0.0389	0.0036
Synth_A	7	1101	0.0670	0.0659	6	1100	0.0668	0.0659
Synth_B	18	791	0.0588	0.0569	17	791	0.0587	0.0567
EBLUP_A	183	212	0.0353	0.0109	191	220	0.0357	0.0111
EBLUP_B	165	202	0.0345	0.0124	175	207	0.0348	0.0118
Clustering B								
Direct	183	184	0.0330	0.0024	187	189	0.0333	0.0030
Greg	183	184	0.0330	0.0025	186	188	0.0332	0.0030
Synth_A	4	1098	0.0666	0.0658	3	1097	0.0664	0.0657
Synth_B	11	785	0.0583	0.0569	11	785	0.0583	0.0569
EBLUP_A	136	161	0.0308	0.0102	138	160	0.0305	0.0091
EBLUP_B	125	156	0.0303	0.0116	126	155	0.0301	0.0105

Source: Own calculations made within the EURAREA project based on POLDATA database

Table 7. Average values of evaluation parameters for local and global optimal sample allocation, indirect estimators, Income, NUTS 4 level, two-stage sampling, different types of clustering with respect to fixed resources

Estimator	$\overline{\hat{V}(\hat{\bar{Y}}_d)}$	$\overline{M\hat{S}E(\hat{\bar{Y}}_d)}$	$\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$	$\overline{A\hat{R}B_d}$	$\overline{\hat{V}(\hat{\bar{Y}}_d)}$	$\overline{M\hat{S}E(\hat{\bar{Y}}_d)}$	$\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$	$\overline{A\hat{R}B_d}$
Optimal sample allocation for domains					Optimal sample allocation for the whole population			
Clustering A								
Direct	2396	2415	0.1150	0.0096	2530	2549	0.1178	0.0097
Greg	2396	2422	0.1152	0.0096	2531	2558	0.1181	0.0096
Synth_A	5	3254	0.1110	0.1100	7	3253	0.1111	0.1100
Synth_B	17	1718	0.0810	0.0791	17	1718	0.0810	0.0791
EBLUP_A	906	1315	0.0847	0.0391	903	1336	0.0854	0.0404
EBLUP_B	**545**	963	0.0717	0.0391	537	973	0.0721	0.0401
Clustering B								
Direct	1667	1685	0.0971	0.0078	1641	1659	0.0964	0.0078
Greg	1661	1679	0.0970	0.0078	1635	1653	0.0963	0.0079
Synth_A	3	3246	0.1106	0.1097	4	3247	0.1108	0.1099
Synth_B	11	1712	0.0804	0.0790	12	1713	0.0806	0.0791
EBLUP_A	674	1088	0.0766	0.0391	674	1081	0.0766	0.0389
EBLUP_B	378	843	0.0660	0.0412	381	840	0.0661	0.0410

Source: Own calculations made within the EURAREA project based on POLDATA database

Comparing the estimated values of evaluation parameters for indirect estimators, one can determine what gain or loss in estimation precision is obtained while using the allocation optimal at local scale instead of allocation optimal at population scale (see Tab. 8). For the GREG estimator, the gain in precision amounts to more than 3% for ***clustering A*** and less than 2% for ***clustering B***. For synthetic estimators a loss in estimation precision is observed. It is very small for the estimator with area level covariates, and much more significant — amounting to about 12% for the model with individual level covariates. The most notable gain in estimation precision is expected for EBLUP estimators. For ***clustering A*** this gain amounts to 4.5% for EBLUP_A and 5.8% for EBLUP_B. For less detailed ***clustering B*** this gain amounts to about 1.2% for both types of EBLUP estimators.

In relation to changes in REE due to the type of allocation chosen, they are very slight. A small decrease is observed for GREG estimator (both types of clustering) and for EBLUP estimators in more detailed ***clustering A.*** For synthetic estimators and less detailed ***clustering B*** changes caused by the type of allocation are insignificant, amounting to less than 1% and in different directions.

Similarly to estimation for NUTS3 level, there seems to be no great differences in estimation precision due to the type of allocation chosen (***clustering***

A and ***B)***, at NUTS4 level of the territorial division. In estimating income at NUTS 4 level, the mean value of $\overline{R\hat{E}E(\hat{\bar{Y}}_d)}$ approaches 12% for the worst estimators (in this case also Synth_A and Greg). Again somewhat smaller values appear in the case of ***clustering B***. As for NUTS3, other properties of indirect estimators are valid. Bias of synthetic estimators is about 11 times as big as that for the direct estimator and again only 4 times as big for EBLUP.

For smaller units of territorial division — NUTS4 — the gain in estimation precision is bigger for synthetic estimators. Using the optimal allocation at local scale, a synthetic estimator with individual level covariates provides more stable estimates of about 28% for more detailed ***clustering A*** and of about 7.5% for less detailed ***clustering B***. The model with area-level covariates yields more precision especially for less detailed ***clustering B*** (about 15%). The gain in estimation precision for GREG amounts to almost 6% for more detailed ***clustering A***, while for less detailed a slight change in the opposite direction could be observed. The last remark holds also for EBLUP estimators, the changes being of less than 0.3% (with the exception of EBLUP_B, clustering A, in which case they are smaller than 1.4%).

Table 8. Gain and loss in estimation efficiency and bias of indirect estimators due to optimal sample allocation (population and domain scale), different types of clustering with respect to fixed resources

Estimator	Ratio of evaluation parameters estimated for population optimal to domain optimal sample allocation					
	$\frac{\left(\overline{A\hat{R}B_d}\right)_G}{\left(\overline{A\hat{R}B_d}\right)_L}$	$\frac{\overline{\hat{V}(\hat{\bar{Y}}_d)_G}}{\overline{\hat{V}(\hat{\bar{Y}}_d)_L}}$	$\frac{\overline{R\hat{E}E(\hat{\bar{Y}}_d)_G}}{\overline{R\hat{E}E(\hat{\bar{Y}}_d)_L}}$	$\frac{\left(\overline{A\hat{R}B_d}\right)_G}{\left(\overline{A\hat{R}B_d}\right)_L}$	$\frac{\overline{\hat{V}(\hat{\bar{Y}}_d)_G}}{\overline{\hat{V}(\hat{\bar{Y}}_d)_L}}$	$\frac{\overline{R\hat{E}E(\hat{\bar{Y}}_d)_G}}{\overline{R\hat{E}E(\hat{\bar{Y}}_d)_L}}$
NUTS3	**Clustering A**			**Clustering B**		
Greg	1.1649	1.0336	1.0112	1.2144	1.0174	1.0078
Synth_A	0.9994	0.8313	0.9979	0.9982	0.8858	0.9977
Synth_B	0.9978	0.9949	0.9975	1.0004	0.9889	1.0005
EBLUP_A	1.0234	1.0445	1.0128	0.8951	1.0122	0.9917
EBLUP_B	0.9491	1.0581	1.0068	0.9101	1.0122	0.9937
NUTS4	**Clustering A**			**Clustering B**		
Greg	0.9941	1.0564	1.0254	1.0144	0.9841	0.9925
Synth_A	0.9993	1.2863	1.0003	1.0014	1.0753	1.0016
Synth_B	0.9995	1.0119	1.0000	1.0011	1.1558	1.0027
EBLUP_A	1.0338	0.9970	1.0083	0.9954	1.0002	0.9993
EBLUP_B	1.0245	0.9863	1.0052	0.9948	1.0101	1.0017

Source: Own calculations conducted within the EURAREA project based on POLDATA database

6. What are the properties of indirect estimators under the two-stage design in a real population?

In this part, we explore the impact of the division into clusters (their number and size) in clustered sampling design on the synthetic and EBLUP estimators in POLDATA population. An added incentive for this is that estimation for many areas poorly represented in the sample (and areas not represented at all) relies almost solely on the synthetic estimator. Throughout, we represent the overall resources (funds) by a fixed sample size. It is assumed that the constraint on the total sample size could be treated equivalently to the overall cost.

As before, we conducted simulations in which we altered the distribution of the within-domain sample sizes, and evaluated the empirical values of MSEs of the small-area estimators and analysed the trade-off between the precisions for the small-area and national population-mean estimators. The clusters were formed by NUTS4-level units, and the target domains were also NUTS4 units. The sampling designs are described by the combinations of the numbers and sizes of clusters presented in Table 9.

Table 9. Combinations of the numbers and within-cluster sample sizes in two-stage sampling design with fixed overall sample size

Notation	Number of NUTS4 as PSU	Dwellings per PSU	Sample size	Cost of the survey
	M	***n_g***	***n***	***C***
PSU_10	10	552	5520	28 050
PSU_40	40	138	5520	29 400
PSU_80	80	69	5520	31 200
PSU_120	120	46	5520	33 000
PSU_184	184	30	5520	35 880
PSU_230	230	24	5520	37 950
PSU_276	276	20	5520	40 020

Note: The cost of a survey is calculated as: $\hat{C} = C_0 + mC_1 + C_2 \sum_{g=1}^{m} n_{(g)}$ with $C_0 = 0$, $C_1 = 45$ and $C_2 = 5$.

Source: Calculations based on POLDATA.

We evaluated the MSEs of the estimators of the small-area and national quantities, searching for combinations of estimators and designs that have highest (average) precisions. For direct estimators in the previous section, we used the weighted totals $S_V = \sum_d N_d^q \operatorname{var}(\hat{\bar{Y}}_d)$ but, recognizing that the synthetic and

EBLUP estimators are biased, the variances in S_V were replaced by MSEs; $S_{MSE} = \sum_d N_d^q MSE(\hat{\bar{Y}}_d)$.

7. Empirical MSEs of small-area estimators and clustering

The main conclusions from the comparison of the MSEs of the small area estimators can be summarized as follows: for each studied sampling design, EBLUP_B is the most efficient, Synth_A is the least efficient. EBLUP estimators are more efficient than their Synth counterparts, and the designs with more detailed clustering are more efficient. Details are given in Table 10. The results are quite unequivocal about the superiority of the EBLUP_B estimator. One must remember, however, that the studied sampling designs have equal overall sample sizes. The designs with greater number of clusters are more expensive, and in practice they may have to be implemented with smaller overall sample sizes. Thus, the assessment of the design PSU_276 by the average MSE is somewhat optimistic. By extrapolation, we might conclude that no clustering (stratified simple random sampling) is optimal, because it is the natural limit of clustered designs with infinitely refined clustering. However, it is not feasible to implement in practice or its cost is prohibitive.

Table 10. Average MSEs for all areas, Income, NUTS4, Poland, 1995

Type of domain / Estimator	Sampling design						
	PSU_10	**PSU_40**	**PSU_80**	**PSU_120**	**PSU_184**	**PSU_230**	**PSU_276**
Synth_A	2966	2822	2793	2777	2778	2756	2753
Synth_B	2550	2116	2027	1992	1985	1964	1964
EBLUP_A	2911	2601	2418	2299	2187	2101	2051
EBLUP_B	2529	1994	1826	1729	1670	1620	1586

Note: The cells of the row-wise minimums are shaded.

Source: Calculations based on POLDATA.

Table 11 lists the MSEs for the quartiles of domains - the domains are split into four groups according to their population sizes. **Q1** stands for first quarter (25%) of the areas that have the smallest population sizes, **Q2** for the next one quarter of the areas according to the population size, **Q3** for the areas between the median and upper quartile of the areas, and **Q4** for the remaining quarter of the areas that have the highest population sizes. Here the comparisons are much less clear-cut, although the designs with more detailed clustering have the smallest MSEs in many settings. (The row-wise minima are marked in the table by shading.) Nevertheless, whenever the design PSU_276 is not marked as optimal, it is not far behind the design with the smallest MSE. Table 12 summarizes the same estimators and designs by REE. Although different 'winners' are identified, the differences among the designs are small and a substantial part of them may be attributed to the limited number of simulations.

Table 11. Average MSEs for domains of different size, Income, NUTS4, Poland 1995

Type of domain / / Estimator	Sampling design						
	PSU_10	**PSU_40**	**PSU_80**	**PSU_120**	**PSU_184**	**PSU_230**	**PSU_276**
Q1							
Synth_A	3955	3767	3757	3725	3719	3713	3717
Synth_B	4555	3810	3751	3561	3545	3570	3548
EBLUP_A	3927	3479	3235	3008	2820	2719	2636
EBLUP_B	4538	3587	3342	2964	2813	2767	2635
Q2							
Synth_A	1626	1655	1547	1600	1618	1554	1532
Synth_B	1610	1295	1189	1219	1218	1174	1173
EBLUP_A	1585	1539	1397	1404	1404	1331	1301
EBLUP_B	1589	1220	1097	1113	1099	1048	1041
Q3							
Synth_A	2286	2215	2153	2166	2175	2134	2123
Synth_B	1818	1529	1421	1448	1451	1406	1401
EBLUP_A	2234	2042	1891	1823	1766	1673	1650
EBLUP_B	1795	1438	1291	1275	1256	1201	1198
Q4							
Synth_A	3988	3642	3706	3606	3591	3613	3629
Synth_B	2195	1812	1731	1723	1709	1688	1716
EBLUP_A	3885	3334	3140	2953	2753	2672	2611
EBLUP_B	2174	1715	1558	1551	1499	1451	1457

Note: The cells of the row-wise minimums are shaded.
Source: Calculations based on POLDATA.

Table 12. Average REEs for domains of different size, Income, NUTS4, Poland 1995

Type of domain / / Estimator	Sampling design						
	PSU_10	**PSU_40**	**PSU_80**	**PSU_120**	**PSU_184**	**PSU_230**	**PSU_276**
Q1							
Synth_A	11.61	11.28	10.97	11.03	11.04	10.87	10.82
Synth_B	12.59	10.87	10.44	10.17	10.16	10.09	10.03
EBLUP_A	11.55	10.93	10.54	10.53	10.48	10.35	10.24
EBLUP_B	12.54	10.58	10.04	9.76	9.65	9.61	9.51
Q2							
Synth_A	9.54	9.42	8.97	9.12	9.17	8.91	8.83
Synth_B	9.92	8.51	7.97	7.98	7.97	7.78	7.78
EBLUP_A	9.43	9.20	8.85	9.05	9.20	9.01	9.00
EBLUP_B	9.86	8.33	7.89	7.97	8.01	7.88	7.89
Q3							
Synth_A	10.26	10.01	9.66	9.75	9.78	9.57	9.52
Synth_B	9.94	8.82	8.27	8.34	8.34	8.12	8.08
EBLUP_A	10.15	9.73	9.36	9.43	9.55	9.32	9.34
EBLUP_B	9.88	8.60	8.06	8.14	8.20	8.02	8.05
Q4							
Synth_A	11.50	10.79	10.70	10.57	10.55	10.49	10.50
Synth_B	9.79	8.43	8.00	7.90	7.89	7.73	7.74
EBLUP_A	11.36	10.50	10.26	10.22	10.16	10.10	10.04
EBLUP_B	9.74	8.28	7.83	7.94	7.99	7.89	7.91

Note: The cells of the row-wise minimums are shaded.

Source: Calculations based on POLDATA.

The results of conducted simulations confirm that the MSE is strongly influenced by the type of clustering — refinement of the clustering is rewarded by a reduction of MSE. Allowing for more first-stage sampling units (PSU) and fewer second-stage sampling units means that a greater number of domains in represented in the sample, which yields a gain in estimation precision. EBLUP_B and Synth_B estimators provide smaller values of MSE (or REE) for all groups of domains. For the smallest domains (Q1) MSE of Synth_B is higher than the one of EBLUP_A.

Trade-off between the precision of national and small-area estimators

One must, therefore, conclude that EBLUP_B is the most efficient estimator for each studied sampling design. Designs with more clusters are beneficial for both small area estimation (with EBLUP_B) and for estimation of the national mean. The results are listed in Table 13.

For synthetic and EBLUP estimators there is almost ideal convergence for local and global scale. The bigger the number of clusters, the more efficient the estimators. For global scale, the best sampling pattern is 230 clusters. For local scale the best estimation precision (the lowest S_{MSE}) is obtained for the sampling pattern with 276 clusters.

Table 13. Estimation precision: global versus local scale, different clustering approaches, Income, NUTS 4, Poland 1995

Type of domain	Sampling design						
	PSU_10	**PSU_40**	**PSU_80**	**PSU_120**	**PSU_184**	**PSU_230**	**PSU_276**
Global scale							
V	195	69	43	33	28	18	19
MSE	297	98	97	65	54	59	67
Local scale **Criterion Function** S_{Var}							
Synth_A	8053	2879	1760	1311	1131	721	763
Synth_B	23 638	8219	4637	3438	3140	2315	2320
EBLUP_A	10 551	11 034	14 508	17 474	19 926	19 774	20 757
EBLUP_B	24 357	11 961	11 005	12 011	13 359	12 586	13 699
Criterion Function S_{MSE}							
Synth_A	121 761	115 197	114 464	113 423	113 446	112 712	112 652
Synth_B	89 422	74 431	70 871	70 472	70 153	69 144	69 555
EBLUP_A	119 069	105 888	98 577	93 696	88 917	85 417	83 522
EBLUP_B	88 573	70 096	63 860	61 895	59 953	58 031	57 607

Note: The cells of the row-wise minimums are shaded. The criterion functions

$S_V = \sum_d N_d^q \,\mathrm{var}(\hat{Y}_d)$ and $S_{MSE} = \sum_d N_d^q MSE(\hat{Y}_d)$ are divided by 10^4.

Source: Calculations based on POLDATA.

8. Conclusion

Results of simulation studies testing the properties of estimators for small areas cannot, in most cases, be generalised but reflect conditions of a particular situation. This was also the case with the present study, which is essentially experimental. The article has outlined the successive stages of optimal sample allocation, starting with the approach based on an artificial population and a sample represented in every domain in order to apply direct estimation. In the remaining sections, more complex solutions have been reviewed, which rely on indirect estimation allowing for specially adapted optimalisation criteria and incomplete sample representation across domains.

To answer what kind of PSU (more or less detailed division) and which way of allocation is more desirable for small area estimation, we started with the solution offered by traditional survey sampling with direct estimator. The optimal sample allocation at global (population) level was distinguished. Then, according to the criterion function defined as a weighted total of domain variances, an optimal allocation at local (domain) level was appointed.

Strong and positive relation was observed between criterion function for domains and the sampling variance at population level: ***clustering B*** $r\left(S_V,\hat{V}(\hat{\bar{Y}})\right)=0.97$ and ***clustering A*** $r\left(S_{\mathrm{V}},\hat{V}(\hat{\bar{Y}})\right)=0.86$. The relation between probability of selecting the PSU (or $\varpi=\frac{\sigma_B^2}{\sigma_W^2}$) and measures of estimation precision at global ($\hat{V}(\hat{\bar{Y}})$) and local (S_{V}) scale was significant. There was a strong evidence confirming that the bigger (closer to unity) the probability of selecting the PSU, the greater the between-area variance. This relation was almost ideal for both types of clustering ($r\left(p_{1i},\omega\right)=0.99$).

Correlation between the criterion function and other measures of estimation efficiency for direct estimator is very strong and positive (correlation coefficients equal almost to unity). The same conclusion also concerns the measures obtained for the population level. Increasing inclusion probabilities for PSU are in linear relation to the increasing value of ω representing relative size of between-area variance. And this increasing between-area variation is strongly and negatively correlated with the criterion function S_V. Hence, the smaller the dispersion between PSU, the bigger the criterion function.

It can be concluded that the change in the sample allocation from optimal at population level, to "optimal" for domains might result in a gain in estimation precision for domains and a loss in variance at population level.

More detailed clustering is more sensitive to gain precision for domains and less sensitive to lose precision at population level due to change in optimal sample allocation from global to local scale. Other measures of estimation quality are not that much sensitive (as variance) to change in optimisation approach. In both types of clustering the change in optimal from global to local scale is connected with a decrease in the relative size of between-area variance.

The optimal sample allocation for a two-stage design, in terms of domains is very close to the optimal sample allocation from the population point of view. In both types of clustering considered, domains seem to require a bit smaller probability of inclusion at the first stage in favour of a small enlargement of the probability at the second stage.

The optimal allocation for domains proved to be very close to optimal allocation for the population also in the last case considered in the study. On the basis of tests conducted on the real population it is evident that when a domain in not represented in the sample, the most effective estimations can be obtained using composite EBLUP estimators, in particular when applying an estimator based on the two-stage model with the area level covariates. Composite estimators were characterised by greater precision when more primary sampling units (PSUs) were used. This tendency was observed for all kinds of domains regardless of their size in the population but the gain in estimation precision obtained as a result of an increase in the number of PSUs did not entirely manifest itself in the group consisting of the smallest and biggest domains.

Sampling designs allowing for more detailed clustering proved to be optimal both for estimation across the population and across domains. The simulations described in this report document the importance that should be accorded to the sampling design for small area estimation. In general, detailed stratification is preferred; in most settings, the stratification that coincides with the domains of interest is the most desirable stratification that is practicable. If only clustering is feasible, more clusters are associated with more efficient small area estimation.

REFERENCES

BÉLAND, Y., BAILIE, L., CATLIN, G., SINGH, M.P., 2000, *An Improved Health Survey Program at Statistics Canada. Proceedings of the Section on Survey Research Methods,* American Statistical Association, Washington, D.C. pp. 671—676.

BRACHA, C., 1996, *Teoretyczne podstawy metody reprezentacyjnej*, (Theoretical basis of survey sampling), PWN, Warszawa.

CITRO, C., and KALTON, G., eds. (2000a). *Small-Area Income and Poverty Estimates. Priorities for 2000 and Beyond.* National Research Council, Washington, DC.

CITRO, C., and KALTON, G., eds. (2000b). *Small-Area Estimates of School-Age Children and Poverty. Evaluation of Current Methodology.* National Research Council, Washington, DC.

DOL, W., 1991, *Small area estimation: a synthesis between sampling theory and econometrics*, PhD thesis. Groningen. The Netherlands, p.45.

GOŁATA, E., KLIMANEK, T., *Small-area estimation with complex sampling design; Initial results on optimal allocation of the overall sample size to the small areas*, October 2002, Internal Report prepared within Workpackage 4.2 of EURAREA project no. IST-2000-26290.

KORDOS, J. (1999), *Problemy estymacji danych dla małych obszarów* (Issues Connected with Estimation for Small Areas, Wiadomości Statystyczne, Nr 1, pp.85—101.

LONGFORD, N. at al., 2004, *Initial Theory Report: Small-Area Estimation with Complex Sampling Design*, unpublished internal report prepared within the European Commission Information Society Technologies Programme - EURAREA, IST 2000-26290 coordinated by Office for National Statistics, London.

LONGFORD N.T.,2002*,* On optimal allocation of the overall sample size to the small areas*, unpublished internal report prepared within the European Commission Information Society Technologies Programme —- Eurarea no IST –2000-26290 coordinated by Office for National Statistics, London.*

MARKER, D.A., 2001, *Producing small area estimates from national surveys: method for minimizing use of indirect estimate,* Survey Methodology, December 2001, Vol. 27, 183—188, Statistics Canada.

RAO, J.N.K., 2003, *Small Area Statistics*, Wiley, New York.

SARNDAL, C-E., SWENSSON B., WRETMAN J., 1992, *Model assisted survey sampling*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

SINGH, M.P., GAMBINO, J. and Mantel, H.J. (1994), *Issues and strategies for small area data*, *S*urvey Methodology 20, 3—22.

Annex

1. The optimal direct allocation

The optimal direct allocation is the allocation of the sample sizes n_d, d=1, …, D, (n1 + n2 + … + nD = n), for which a summary of the sampling variances is minimized. We take a weighted total of the sampling variances for all domains as the criterion to evaluate the estimation precision. It can be expressed by: $S = \sum_d N_d^q \operatorname{var}(\hat{\mu}_d)$. We assume simple random sampling, in which all areas are represented in the sample, and start with the direct estimator. The sampling variance of the direct estimator $\hat{\mu}_d$ for area d is $\operatorname{var}(\hat{\mu}_d) = \frac{\sigma^2}{n_d}$, so the criterion function takes the form of: $S = \sum_d N_d^q \operatorname{var}(\hat{\mu}_d) = \sum_d N_d^q \frac{\sigma^2}{n_d}$. Thus, the optimal sample sizes are: ${}_{DA}n_d^* = n \times \frac{N_d^{q/2}}{N_1^{q/2} + N_2^{q/2} + \ldots + N_D^{q/2}}$.

2. The optimal composite allocation

The composite estimator assumes that the criterion function is constructed upon the variance specified for the (unconditional) composite estimator: $\tilde{\mu}_d = (1 - b_d)\hat{\mu}_d + b_d \hat{\mu}$. The coefficients b_d are set so that the mean squared error of $\tilde{\mu}_d$ is minimized. When the overall sample size is much greater than the sample size for area d, the sampling variation of $\hat{\mu}$ can be ignored. The expected mean squared error of $\hat{\mu}_d$ is then $\frac{\sigma_B^2}{1 + n_d \omega}$, where $\omega = \frac{\sigma_B^2}{\sigma_W^2}$ is the ratio of the between- and within-area variances. Minimizing the objective function $S = \sum_d N_d^q \frac{\sigma_B^2}{1 + n_d \omega}$ requires solving the equation

$$-\frac{\sigma_B^2}{(1 + n_d \omega)^2} N_d^q + \frac{\sigma_B^2}{(1 + n_D \omega)^2} N_D^q = 0$$

or, equivalently

$$\frac{1+n_d\omega}{1+n_D\omega}=\left(\frac{N_d}{N_D}\right)^{\frac{q}{2}}.$$

In general, this can be solved by the method of Lagrange multipliers. Assuming that ω is known, the solution of this equation can be expressed analytically as follows:

$$_{CA}n_d^*=\frac{(D+n\omega)N_d^{\frac{q}{2}}}{\omega\sum_d N_d^{\frac{q}{2}}}-\frac{1}{\omega}$$

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 755—776

PROBLEMS OF ESTIMATING UNEMPLOYMENT FOR SMALL DOMAINS IN POLAND

Elżbieta Gołata[1]

ABSRACT[2]

The paper presents results of some attempts to estimate unemployment for small domains in Poland. These are the results of the research undertaken within the EURAREA project (IST-2000-26290) compared with some own research. The properties of the estimators are discussed from the domain specific point of view and combining all areas

Key words: Small Area Estimation, unemployment, properties of indirect estimators.

1. Introduction

First attempts at applying various approaches to parameter estimation for small areas in Poland were undertaken about ten years ago, especially after the international conference on small area statistics held in Warsaw in 1992 (Kalton, Kordos, Platek, 1993). There were only a few attempts to apply small area estimation (SAE) methods to measure the extent of unemployment, poverty, household structure and in agriculture related surveys (Kordos, Paradysz, 2000). Further application and examination of "standard" indirect estimators properties were undertaken within the EURAREA project[3]. The standard estimators were defined in the project as: 'the techniques of domain estimation (synthetic estimators, GREGs and composite estimators) which entered into use in the United States and Canada in the 1980s, and have been the subject of steady theoretical refinement since' (EURAREA Documents, IST 2000-26290, Annex 1 — "Description of Work"p.4).

[1] elzbieta.golata@ae.poznan.pl. The Poznan University of Economics, Poznan, Poland.

[2] This paper is based on the presentation prepared for the 54 ISI Session in Berlin, August 2003.

[3] The EURAREA project no. IST-2000-26290 entitled *Enhancing Small Area Estimation Techniques to meet European needs* is part of 5th framework programme for research, technological development and demonstration of EU. Its main co-ordinator is ONS — Office for National Statistics, UK.

One of the main obstacles hampering the practical use of indirect estimators is their unknown effectiveness with respect to real data. Survey data seldom meet the assumptions adopted in the models. Hence, it is the objective of the EURAREA project to use real data in order to test estimators representing basic indirect estimation techniques as well as review and develop the theory in an attempt to accommodate it to the existing databases. The objectives of the paper are:

1. To test how the standard Small Area Estimators would perform on Polish database.
2. What properties of them could be distinguished as concerns:
   a. area specific properties
   b. evaluation for all areas
   c. properties of the estimators distribution
   d. general estimators characteristics
3. Consideration of region specific approach.
4. Influence of the model applied - correlation of the covariates.
5. To formulate suggestions for further research.

This study is limited to estimate the proportion of ILO unemployed at two different levels of territorial division: NUTS3 and NUTS4[1]. The evaluation is made using simulation experiments on population data. Another aim is to demonstrate the potential benefits of SAE techniques comparing their accuracy with the possible direct estimation, and a comparison between indirect estimators, looking at which do better in which circumstances, and why. In addition, this is a part of a broader study conducted in a comparable way in six different European countries[2], which will provide data to assess the effectiveness of SAE in the European context [see Heady P., 2003].

2. Assumptions of the simulation study

For the purpose of the EURAREA validation program, a special database has been set up. The Polish database — the so-called super-population labelled POLDATA — has been created on the basis of 3 data sources: the 1995 Micro-census, the 1995 Household Budget Survey and the Local Data Bank. POLDATA provides real information about the target variables[3] and represents as closely as

[1] In this study small areas: NUTS3 or NUTS4 correspond to sub-regional and local levels of a territorial division as set out for EU countries (Nomenclature of Territorial Units for Statistics, Decand G., 1996).

[2] There are six participant countries: Great Britain, Finland, Sweden, Italy, Spain and Poland.

[3] In the project three target variables are estimated: ILO unemployment, household composition and income. Two of those variables were available from Micro-census data. Income was imputed from the 1995 Household Budget Survey.

possible the characteristic of Poland in 1995 with respect to the new administration division of the country which was introduced in January 1999.

For the purposes of applying the standard estimators, the proportion of ILO unemployed (in the whole population over 15) in an area, and not the proportions of unemployed and economically active simultaneously, was estimated. Unemployment was estimated as binary not multivariate or Poisson variable. In choosing covariates, a set of variable categories was harmonized for the common model for all the countries. The intention was to include all variable categories which experience has shown to be effective. In the case of ILO unemployment the 'standard variables' are: age, sex, education, employment status and housing. The simulation study was conducted on samples drawn from the POLDATA according to a two-stage sampling with unequal probabilities. The estimators assigned in the project as 'standard' are as follows (Särndal et al.1992, Ghosh, Rao, 1994, Rao, 1999, Lehtonen, Veijanen, 1998, EURAREA Documents: *Standard Estimators*, 2001)[1] :

1. The direct estimator

$$\hat{\bar{Y}}^{(1)} = \hat{\bar{Y}}_d^{Direct} = \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id}y_{id} \; ; \qquad (1)$$

2. The GREG with a standard linear regression model

$$\hat{\bar{Y}}^{(2)} = \hat{\bar{Y}}_d^{Greg} = \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id}y_{id} + \left(\bar{\mathbf{X}}_{.d} - \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id}x_{id}\right)^T \hat{\boldsymbol{\beta}}\, , \qquad (2)$$

where: $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in u_d} w_{id}\, x_{id}\, x_{id}^T\right)^{-1} \sum_{i\in u_d} w_{id}\, x_{id}\, y_{id}$;

3. The synthetic estimator considered under three different models:

   a) a linear two level model with individual data $y_{id} = x_{id}^T\boldsymbol{\beta} + u_d + e_{id}$

$$\hat{\bar{Y}}^{(3)} = \hat{\bar{Y}}_d^{Synth\_a} = \bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}} \text{ with } \bar{X}_{.d} = (\bar{X}_{.d,1},...,\bar{X}_{.d,p})^T \qquad (3)$$

   b) a linear model with area-level covariates and a pooled sample estimate of within area variance

[1] The estimators discussed as 'standard' in the project were classified into four groups: (a) direct, (b) estimators considered under a design based framework — GREG, synthetic estimator, SPREE and sample-size dependant estimator, (c) estimators considered under a frequentist model based framework — regression synthetic estimator, EBLUP and (d) estimators considered under a Bayesian model based framework — Empirical Bayes (EB) and Hierarchical Bayes (HB) estimators. After a discussion, group (d) was excluded and the set of seven estimators was agreed upon. All the estimators were calculated using a special software code in SAS prepared by ONS.

$$\hat{\bar{Y}}^{(4)} = \hat{\bar{Y}}_d^{Synth\_b} = \bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}} \text{ ,with: } \bar{y}_{.d} = \bar{x}_{.d}^T\boldsymbol{\beta} + \xi_d \text{ , } \xi_d \sim iid\ N(0,\sigma_u^2 + \frac{\sigma_e^2}{n_d}) \text{ ,}$$

$$\hat{\sigma}_e^2 = \frac{1}{n-na}\sum_i\sum_d (y_{id} - \bar{y}_{.d})^2 \qquad (4)$$

c) a logistic model with area level covariates

$$\hat{\bar{Y}}^{(5)} = \hat{\bar{Y}}_d^{Synth\_c} = \log it^{-1}(\bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}})\text{ ;} \qquad (5)$$

4. The EBLUP estimator using models:
   a) a linear two level model with individual data

$$\hat{\bar{Y}}^{(6)} = \hat{\bar{Y}}_d^{Eblup\_a} = \gamma_d(\bar{y}_{.d} - \bar{\mathbf{x}}_{.d}^{\mathrm{T}}\hat{\boldsymbol{\beta}}\ ) + \bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}} \text{ ,} \qquad \text{where } \gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d}\text{;} \qquad (6)$$

   b) a linear model with area-level covariates and a pooled sample estimate of within-area variance

$$\hat{\bar{Y}}^{(7)} = \hat{\bar{Y}}_d^{Eblup\_b} = \gamma_d\hat{\bar{Y}}_d^{direct} + (1-\gamma_d)\bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}} \text{ , where } \gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}\text{ .} \qquad (7)$$

The ‘standard performance criteria’ are the criteria applied in the project to the estimates from all the simulated samples. The criteria to be used for each area are: absolute relative bias $A\hat{R}B = \frac{1}{K}\left|\sum_{k=1}^{K}\frac{(\hat{Y}_{d,k} - Y_d)}{Y_d}\right|$ and relative root average squared error $R\hat{E}E_{MSE} = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{Y}_{d,k} - Y_d\right)^2}}{\left|Y_d\right|}$ where: Y_d is the population value in small area *d* and $\hat{Y}_{d,k}$ is the estimate of Y_d from simulation *k*. The results for each area were also combined to produce an overall figure by taking the straightforward or population weighted average across areas (EURAREA Documents: *Standard performance criteria,* 2001, the performance criteria are listed in the Annex).

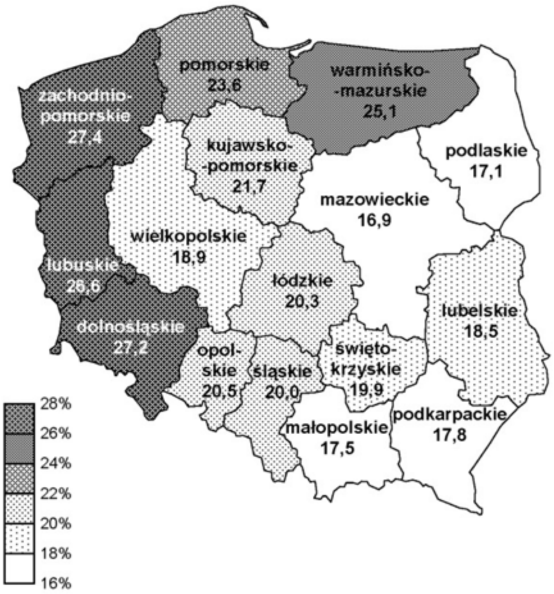
3. An analysis of estimator properties

Labour market problems excite special interest in the whole of Polish society. In the transformation process Polish labour market has changed from that of labour force shortage to one characterised by its overabundance. Unemployment, from the very beginning of 90^{thies} has started to assume alarming dimensions and is characterised by great territorial differentiations at national as well as regional level [see graph 1]. It is due to structural differences in economy and regional inequalities in the transformation process. The regularities observed at national level, in most cases, cannot be generalised and differ from region to region. This situation requires additional studies reflecting the regional specificities. For example in March 2003 the highest registered unemployment rate in Poland was observed in warmińsko-mazurskie voivodship — 29.4%, and the lowest in mazowieckie viovodship — 14.3% [voivotship refers to NUTS2 level according to Eurostat territorial division].

In Wielkopolska — one of the largest regions in Poland, unemployment was of about average country level — 18.9%, but its with-in-region differentiation was significant. In March 2003 the highest unemployment rate in Wielkopolska voivodship was observed in Zlotow county (NUTS4 level[1]) amounting to 22.44%. The county of Poznan was characterised by the lowest unemployment level equal to 2.6%. So the dispersion between the extreme values amounted to about 20 percentage points. The only available information for county level concerns registered unemployment. This paper presents some attempts made to estimate ILO unemployment rate for NUTS4 units of territorial division in Poland and in Wielkopolska region.

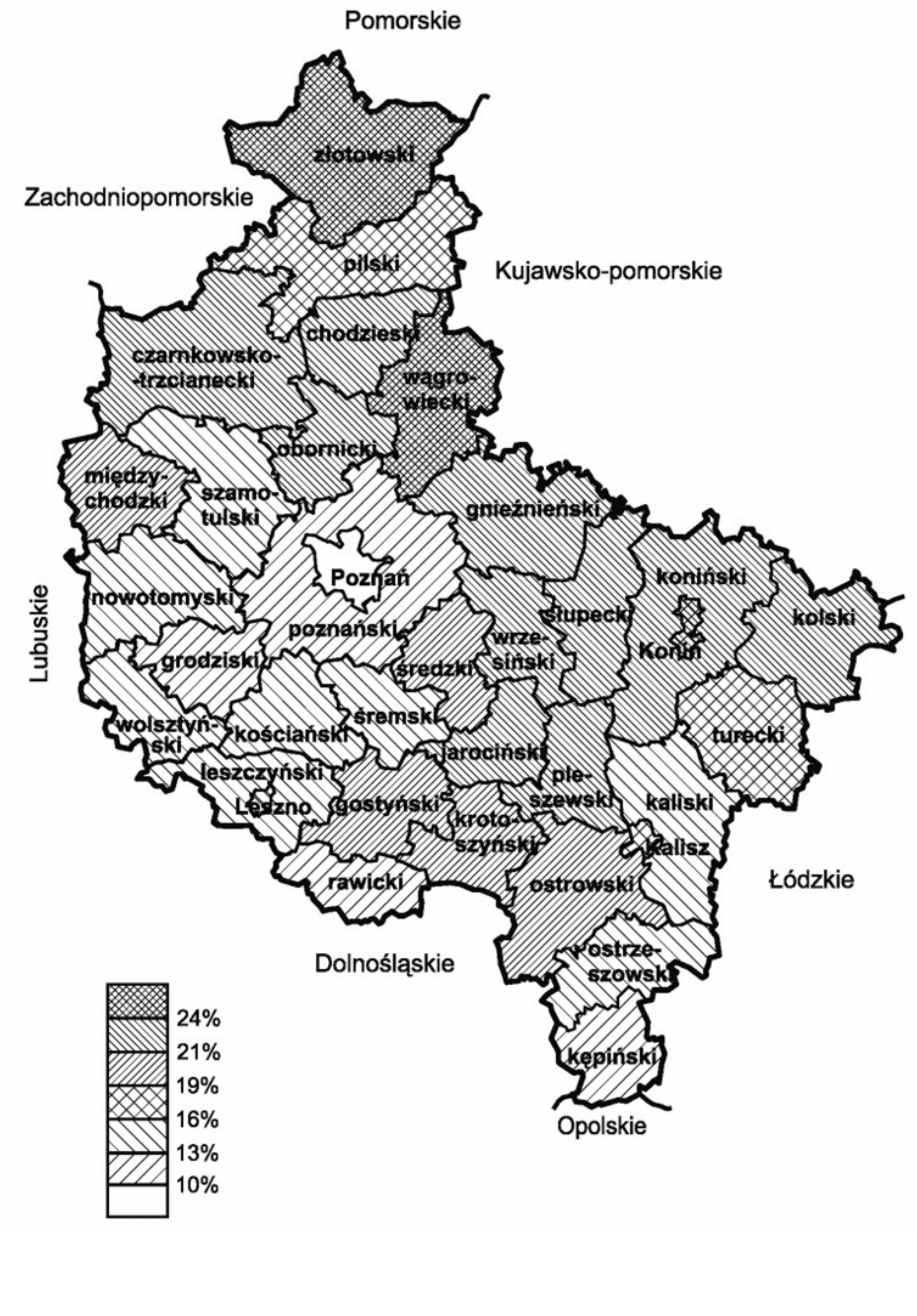
Graph 1.

A Territorial differentiation of ILO unemployment rate, NUTS2 Poland, March 2003



[1] NUTS3 level refers to a group of counties.

B. Territorial differentiation of registered unemployment rate, NUTS4 level in Wielkopolska region, March 2003



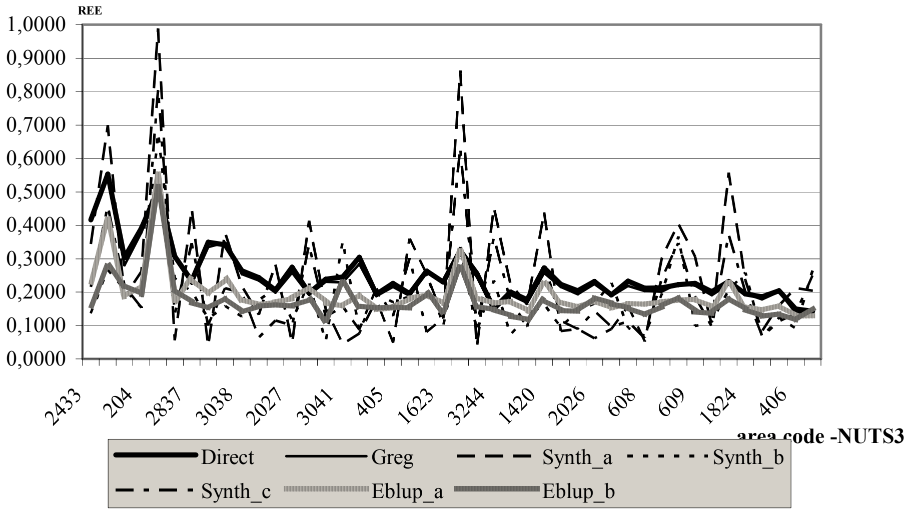
In traditional survey sampling, the proper sample size is one of the basic conditions of required precision. It occurs, that there is no relation between measures of precision applied for indirect estimates and the sample size for domain, whereas there is a strong relation observed for direct and GREG estimators. This relation is stronger for NUTS3 (the correlation coefficients assume the value of about — 0.7) than for NUTS4 (about — 0.5). The correlation obtained for EBLUP estimates is relatively stronger than for synthetic ones.

The area specific criteria are presented in the graph 2 [A refers to NUTS3 level and B refers to NUTS4 units]. Areas are ordered by their size in population. For NUTS3 Relative Estimation Error REE assumes the value of about 0.2 — on average. Bigger variation in relative estimation precision is observed for the synthetic estimators than for Eblups. There are some territorial units for which REE of synthetic estimators assumes extremely high values. They are characterised of exteremely low unemployment level [see graph 2 A: Warszawski — 1422] or very high unemployment [Koszalinski — 3244]. For NUTS4 level of territorial division REE assumes much higher values starting with 0.1; but in many cases it takes the value exceeding 1 for direct estimators [especially for extremely small units, see graph 2 B].

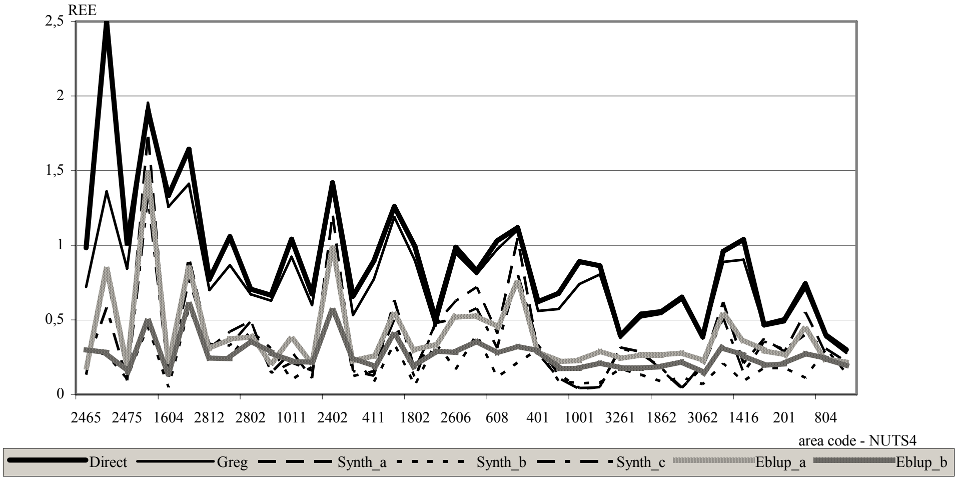
The overall evaluation of the estimators provides information on how well, on average, they estimate values in each area. Measures of performance estimated for all areas have to be taken into account and averaging has to be made across all localities. All the synthetic measures are higher for NUTS4 than for NUTS3 [see tab.1]. The difference between simple and weighted average of $A\hat{R}B$, $M\hat{S}E$ and $R\hat{E}E_{MSE}$ is insignificant, the weighted version is somewhat smaller. In spite of the way of calculating the mean, ranking of estimators obtained is exactly the same. But there is a difference concerning the "best" estimator depending on the level of territorial division. For NUTS3 the smallest value of $\overline{R\hat{E}E}_{MSE} = 0.1720$ is observed for *Eblup_b* estimator with area level covariates, while for NUTS4 it is the synthetic estimator $\overline{R\hat{E}E}_{MSE} = 0.2582$ *Synth_b* (also with area level covariates). It means that when estimating the proportion of ILO unemployed for NUTS3 using *Eblup_b*, the average mistake we make equals to about 17 % of the unknown estimated value. This average relative error for NUTS4 amounts to over 25% while applying *Synth_b*.

Graph 2.

A. Area specific criteria for the estimator in relation with the area size, REE, NUTS3, Unemployment, Poland, May 1995



B. Area specific criteria for the estimator in relation with the area size, REE, NUTS4, Unemployment , Poland, May 1995



Source: Own calculation based on POLDATA

The smallest bias ($\overline{A\hat{R}B}$) concerns the direct estimates. Its average relative value for NUTS3 level equals to about 3.5%, while for NUTS4 it assumes the value of about 8.3%. For synthetic estimates the average bias reaches the value of 37% for NUTS4 and less than 25% for NUTS3. As mentioned above, at NUTS4 level the smallest value of $\overline{R\hat{E}E}_{MSE} = 0.2582$ was observed for *Synth_b*. But a small difference with the value of $\overline{R\hat{E}E}_{MSE} = 0.2872$ obtained for *Eblup_b* should be underlined. When we notice that the average bias for *Eblup_b* estimator with area level covariates ($\overline{A\hat{R}B} = 0.1698$) is half of that for its synthetic counterpart $\hat{Y}^{(4)}$ ($\overline{A\hat{R}B} = 0.3706$), diversity of the indication of the 'best' estimator is more evident.

A summary characteristic of small area estimators may be presented by an analysis of empirical distributions of $R\hat{E}E_{MSE}$. The graphical visualisation of appropriate curves presenting the properties of the empirical distribution of $R\hat{E}E_{MSE}$ are on graph 3 A — NUTS3 level and B — NUTS4 level of territorial division.

For NUTS4 and synthetic estimators a group of domains can be distinguished for which the $R\hat{E}E_{MSE}$ obtains small values of less than 10%. Than two local maximums are observed: for $R\hat{E}E_{MSE} \approx 20\%$ and $R\hat{E}E_{MSE} \approx 35\%$. Domains for which $R\hat{E}E_{MSE}$ obtains bigger values are rare. The distributions for EBLUP estimators have their maximums for $R\hat{E}E_{MSE} \approx 30\%$. They are less leptokurtic than for synthetic estimators, but more homogeneous, with one peak, though strongly skewed to the right. The distributions for direct estimators obtain one maximum for $R\hat{E}E_{MSE} \approx 70\%$, are almost symmetric and of small kurtosis.

For NUTS3 level, the difference between maximums for different estimators is smaller. The distributions of $R\hat{E}E_{MSE}$ for direct estimators are more skewed to the right than for NUTS3. Most leptokurtic and of the smallest value for which the maximum is obtained, are the distributions for EBLUP estimators. They are also right skewed. The biggest difference concerns distributions obtained for $R\hat{E}E_{MSE}$ of synthetic estimators. They are more uniform, flattened, without distinct maximum. The frequencies are sinuous of a decreasing amplitude.

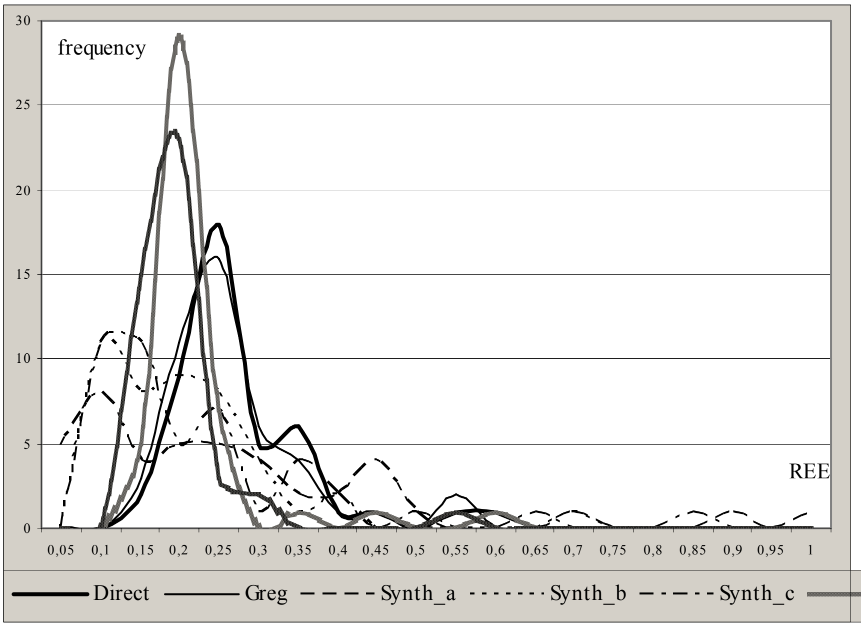
Table 1. Synthetic evaluation of the estimation quality of the percentage of ILO unemployed in NUTS3 and NUTS4 level, Poland, May 1995, simulation study within the EURAREA project

Estimator	Average Relative Bias $\overline{A\hat{R}B}$				Average Squared Error $\overline{M\hat{S}E}$				The Relative Estimation Error $\overline{R\hat{E}E}_{MSE}$			
	Simple Average	Population Weighted Average	min	max	Simple Average	Population Weighted Average	min	max	Simple Average	Population Weighted Average	min	max
						NUTS3						
Direct	0.0348	0.0356	-0.0082	0.0044	0.0004	0.0003	0.0001	0.0010	0.2534	0.2528	0.1411	0.5695
Greg	0.0375	0.0378	-0.0076	0.0022	0.0003	0.0003	0.0001	0.0010	0.2449	0.2446	0.1358	0.5743
Synth_a	0.2438	0.2245	-0.0663	0.0354	0.0006	0.0005	0.0000	0.0044	0.2512	0.2498	0.0355	1.0361
Synth_b	0.1434	0.1390	-0.0338	0.0282	0.0002	0.0002	0.0000	0.0012	0.1825	0.1814	0.0534	0.7434
Synth_c	0.1725	0.1597	-0.0529	0.0285	0.0004	0.0003	0.0000	0.0028	0.2047	0.2032	0.0621	0.8694
Eblup_a	0.0762	0.0593	-0.0263	0.0178	0.0002	0.0002	0.0001	0.0011	0.1950	0.1938	0.1281	0.6061
Eblup_b	0.0758	0.0660	-0.0192	0.0166	0.0002	0.0002	0.0001	0.0006	0.1738	0.1720	0.1145	0.5855
Min	***0.0348***	***0.0356***	***-0.0663***	***0.0022***	***0.0002***	***0.0002***	***0.0000***	***0.0006***	***0.1738***	***0.1720***	***0.0355***	***0.5695***
Max	**0.2438**	**0.2245**	**-0.0076**	**0.0354**	**0.0006**	**0.0005**	**0.0001**	**0.0044**	**0.2534**	**0.2528**	**0.1411**	**1.0361**
						NUTS4						
Direct	0.0829	0.0666	-0.0505	0.0491	0.0036	0.0030	0.0003	0.0142	0.8363	0.7676	0.3014	2.6590
Greg	0.0904	0.0672	-0.0336	0.0333	0.0029	0.0025	0.0003	0.0123	0.7572	0.7580	0.2236	2.6798
Synth_a	0.3706	0.3436	-0.0907	0.0536	0.0009	0.0008	0.0000	0.0082	0.3752	0.3756	0.0335	2.7014
Synth_b	0.2239	0.2186	-0.0740	0.0526	0.0005	0.0004	0.0000	0.0055	0.2579	0.2582	0.0434	1.4505
Synth_c	0.3185	0.2982	-0.0907	0.0451	0.0007	0.0007	0.0000	0.0082	0.3277	0.3280	0.0403	2.2775
Eblup_a	0.2417	0.1963	-0.0689	0.0428	0.0007	0.0007	0.0001	0.0055	0.3850	0.3854	0.0964	2.2639
Eblup_b	0.1698	0.1518	-0.0638	0.0500	0.0005	0.0004	0.0001	0.0046	0.2869	0.2872	0.1169	1.3489
Min	***0.0829***	***0.0666***	***-0.0907***	***0.0333***	***0.0005***	***0.0004***	***0.0000***	***0.0046***	***0.2579***	***0.2582***	***0.0335***	***1.3489***
Max	**0.3706**	**0.3436**	**-0.0336**	**0.0536**	**0.0036**	**0.0030**	**0.0003**	**0.0142**	**0.8363**	**0.7676**	**0.3014**	**2.7014**

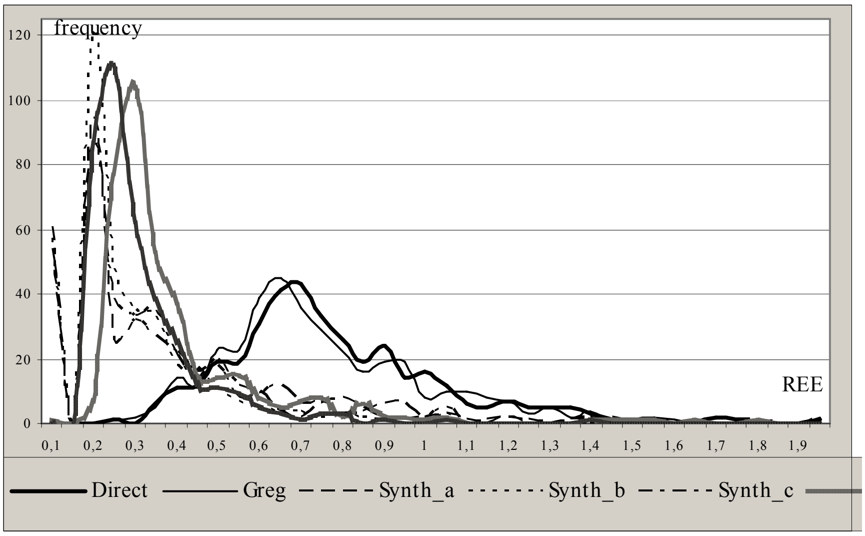
Source: Own calculations made on the simulations on POLDATA within the EURAREA project

Graph 3.

A. Distribution of REE for standard estimators, ILO Unemployment, NUTS3, Poland 1995



B. Distribution of REE for standard estimators, ILO Unemployment, NUTS4, Poland 1995



Source: Own calculations based on POLDATA

4. Extensions to estimation of ILO unemployment at NUTS4 Level

High values of relative estimation errors, which for NUTS4 level obtain the value from 25—40% seems unsatisfactory. Results obtained for NUTS3 are much better, as average value of $\hat{REE}_{MSE}$ does not exceed 25%. It seems that there is a possibility of improving the estimates by creating the model that would incorporate more correlated variables i.e. registered unemployment instead of the proportion of unemployed obtaining the unemployment claim. Some attempts in this direction were made for Wielkopolska — one of the largest regions in Poland which is characterised of an average level of unemployment rate and rather big territorial differentiations[1]. Differences in the quality of a model with unemployment benefit or registered unemployment may be characterised by the values of appropriate correlation coefficients. As presented in Table 2, the proportion of ILO unemployed at NUTS4 level is strongly and positively correlated with the proportion obtained from registration. The correlation coefficient assumes the value of $r = 0.73$, while using the data concerning unemployment benefit provides the correlation of $r = 0.43$. This relation is even weakening due to changes in the regulations. The proportion of registered unemployed in Wielkopolska region who obtain the claim equals to about 18%, while in 1995 it was just the opposite — about 75%.

The precision of a synthetic estimator depends on the strength of the relationship between the target variable and the covariates. The following graph presents how incorporating the registered unemployment rate to the model would improve the estimates [see graph 4 A and B — for *synth_b* or *eblup_b* estimators]. Change caused by inserting registered unemployment to the model instead of unemployment benefit resulted in increasing the coefficient of determination for area level usual regression model from $R^2 = 0.385$ to $R^2 = 0.715$.

This graphical way of presentation provides also information about the properties of the estimators as concerns the trade off between the bias and the variance for both types of models (with registered unemployment — RU and unemployment benefit UB). The horizontal axis represents the true value and the vertical axis gives the estimates of *Direct*, *Synth_b* and *Eblup_b* estimators. Graph 4 present also the relation of the estimates from a single sample (upper row) and mean over 1000 replicates to the true population values. In case of unbiased estimators of a small variance the dots on the graphs should provide a 45 degree line.

[1] This restriction is due to the difficulty in obtaining information about registered unemployment in 1995 for all gminas — the smallest units of territorial division. Such information is necessary to recalculate data and provide information about this covariate for the new administration division of the country.

Table 2. Correlation matrix between target variable and the covariates, NUTS4, Wielkopolska region, 1995

	ILO unemployment	Age	Registered unemployment	Number of rooms in the household's apartment	Sex	Education	Unemployment benefit
ILO unemployment	1						
Age	-0.2796	1					
Registered unemployment	0.7297	0.0172	1				
Number of rooms in the household's apartment	-0.3439	-0.0982	-0.2228	1			
Sex	-0.1745	0.5222	-0.3269	-0.2167	1		
Education	-0.3793	-0.3750	-0.4879	0.6507	-0.2882	1	
Unemployment benefit	0.4279	0.0077	0.6501	-0.4622	-0.4050	-0.4933	1

Source: Own calculations made on the simulations on POLDATA within the EURAREA project.

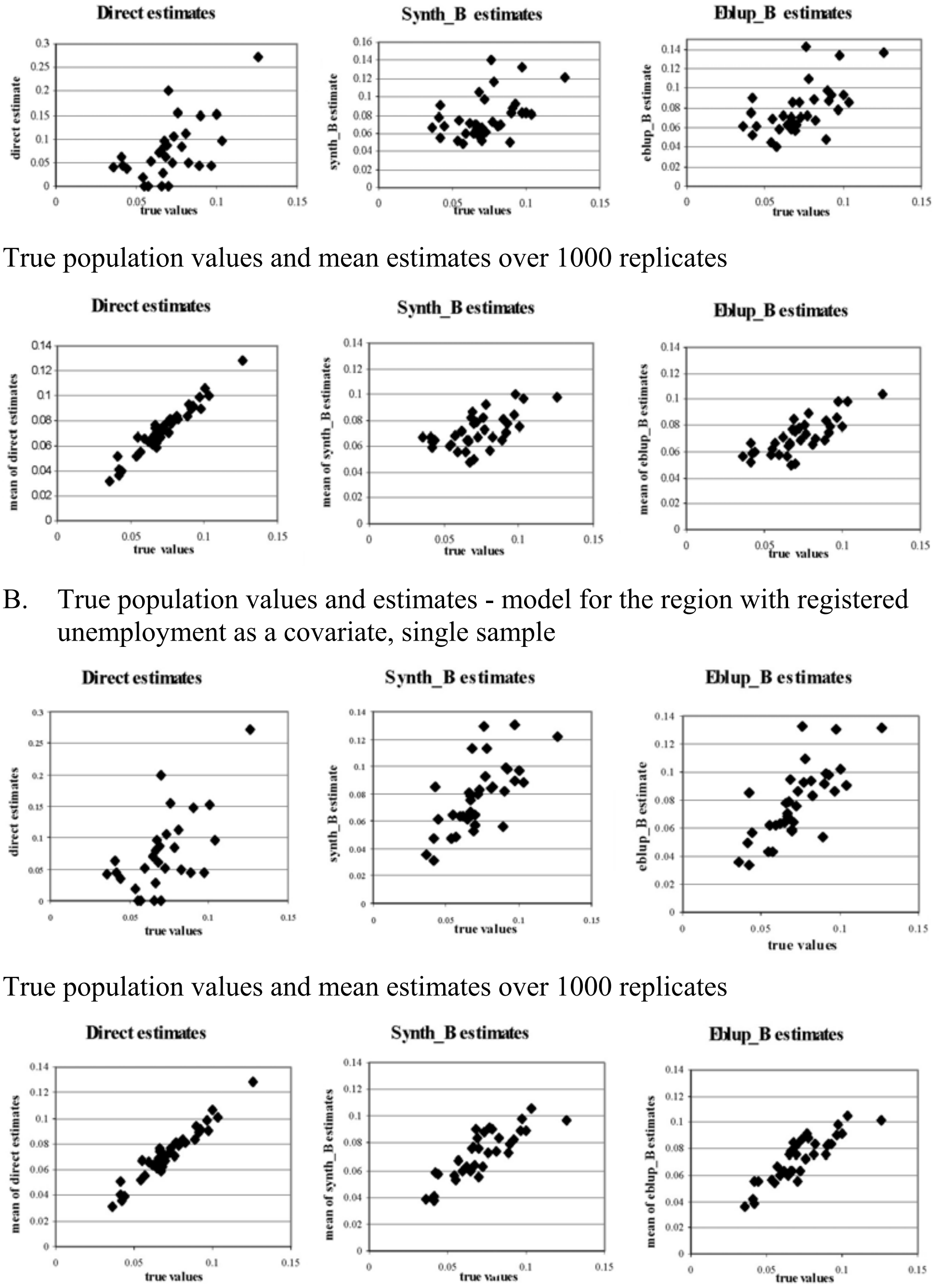
In case of *Direct* estimator, the values are scattered for the single sample, but the average over 1000 simulations lie very close to the 45 degree line — demonstrating that the estimates are unbiased and that large sample would reduce MSE to zero. The synthetic estimates may be are more close to the true value for a single sample (less scattered?). But due to errors the average values of the estimates from 1000 samples do not greatly improve the estimates derived from a single sample.

The precision of synthetic estimates is strongly influenced by the quality of the model applied.. How the strength of the relationship between the target variable and the covariates influence the estimation is presented in the middle of graph 4: in part A — referring to the model with unemployment benefit and in part B — referring to the model with registered unemployment as a covariate. It can be seen that although the points in part B are much scattered, they are less biased. And after 1000 simulations the "line" obtained much better fits to the 45 degree line than it was the case in part A.

The results obtained for the *Eblup_b* estimator do not differ much from the ones obtained for the *Synthetic_b*. The *Eblup_b* is a composite estimator being a weighted combination of the direct and synthetic ones. Taking some additional explanatory information from the direct component it reduces the bias of the synthetic estimator on the cost of some increase in variation. As a result, not only the estimates obtained from a single sample are more tightly grouped around the identity line, but a significant improvement can be observed over 1000 simulations. This can be seen on the right side of the graphs, especially in part B referring to the model with registered unemployment as a covariate.

Looking at the plots of single sample estimates against the true values in case of different estimators we obtained a picture scattered for direct estimates, while for synthetic estimates and for especially *Eblup_b*, the 45 degree line was visible. To test how the estimates represent the relationship between different areas — the territorial inequality of the region, we can use a cartographical form of presentation. But as maps are not always precise enough, for the aim of this study, a following comparison was prepared. The true population value of the proportion of ILO unemployed in Wielkopolska region was equal to 0.0722 (with standard deviation of 0.02). After 1000 simulations the expected value of the direct estimator amounted to 0.0725. Taking into consideration the indirect estimators, the eblup estimator provided best approximation. For the model with unemployment benefit, the expected value was equal to 0.0713 while for the model with registered unemployment the value of 0.0721 was obtained.

Graph 4.



A. True population values and estimates - model for the region, single sample

True population values and mean estimates over 1000 replicates

B. True population values and estimates - model for the region with registered unemployment as a covariate, single sample

True population values and mean estimates over 1000 replicates

Source: Own calculations made on the simulations on POLDATA within the EURAREA project.

Table 3. Synthetic evaluation of estimation quality of the proportion of ILO Unemployed, NUTS4, Wielkopolska region, 1995

Estimator	Model for the whole sample	Model for the region			Model for the region with registered unemployment as a covariate		
	Average value of Mean Squared Error	Average value of Mean Squared Error	Average Relative Estimation Error	Average Relative Bias	Average value of Mean Squared Error	Average Relative Estimation Error	Average Relative Bias
	$\overline{M\hat{S}E}$	$\overline{M\hat{S}E}$	$\overline{R\hat{E}E}_{MSE}$	$\overline{A\hat{R}B}$	$\overline{M\hat{S}E}$	$\overline{R\hat{E}E}_{MSE}$	$\overline{A\hat{R}B}$
Direct	0.0036	0.0028	0.6636	0.0662	0.0028	0.6636	0.0662
Greg	0.0029	0.0021	0.6434	0.0615	0.0021	0.6446	0.0563
Synth_A	0.0009	0.0004	0.2731	0.2291	0.0003	0.2146	0.1373
Synth_B	0.0005	0.0007	0.3794	0.2077	0.0005	0.3151	0.1222
Synth_C	0.0007	0.0005	0.2903	0.2081	0.0003	0.2398	0.1389
Eblup_A	0.0007	0.0005	0.3056	0.1550	0.0003	0.2601	0.1053
Eblup_B	0.0005	0.0007	0.3824	0.1643	0.0005	0.3284	0.1052

Source: Own calculation based on POLDATA

Other measures of estimation precision obtained for the region confirm the priority of eblup estimators [see tab.3]. But it should be added that the results obtained for synthetic estimators, especially in terms of MSE (or REE) are also good. Incorporating the registered unemployment into the model resulted in a decrease in the average value of MSE from 0.0007 to 0.0005 or from 0.0005 to 0.0003, depending on the estimator. However the relative contribution of bias and variance to the overall MSE depends on the estimator in question. The synthetic have high bias, but low variance. And the composite have a combination of both. It can be seen in the average value of relative bias: comparing the value of 12.2% for synthetic and 10.5% for eblup estimators. Comparison of the results obtained for the two models distinguished in the analysis provide another confirmation of the importance of the strength of relation between the target variable and the covariates. For the model with unemployment benefit the relative bias of synthetic estimators assumes the value exceeding 20% and about 15—16% for eblup estimators. While taking into account the percentage of registered unemployed decrease the relative bias to 12—13% and 10.5% for synthetic and eblup estimators respectively. The average value of REE still assumes quite big values, but taking into account the differentiation of the target variable and its low value in the population, it should be interpreted with great caution.

Concluding Remarks

One of the important benefits of the EURAREA project in Poland is entering into practical use of administration registers in a form that enables full integration of different databases on an individual level. It seems that future use of these data sets as sources of individual level covariates might well assist small area estimation. As no estimation for NUTS3 or NUTS4 level is made in Poland — any experience in this field is of great importance.

Measures of effectiveness obtained for small area estimation in comparison with direct estimates [see tab. 4] show impressive gain in estimation precision. For NUTS4 the best results obtained for *Synth_b* and *Eblup_b* inform that $\overline{M\hat{S}E}(\hat{\bar{Y}}^{(i)})$ is about 85% smaller than the variance of the direct estimator. The gain obtained for NUTS3 level is also great and it amounts to about 50% when applying *Eblup_b* estimator with area level covariates instead of direct one.

Table 4. Change in the estimation quality of the percentage of ILO unemployed obtained by different indirect estimators in comparison with direct estimator, Poland, May 1995, simulation study within the EURAREA project

Estimator *(i)*	$\frac{\left(\overline{\hat{V}(\hat{Y}_E^{(1)})} - \overline{\hat{V}(\hat{Y}_E^{(i)})}\right)}{\overline{\hat{V}(\hat{Y}_E^{(1)})}}$		$\frac{\left(\overline{M\hat{S}E(\hat{Y}_E^{(1)})} - \overline{M\hat{S}E(\hat{Y}_E^{(i)})}\right)}{\overline{M\hat{S}E(\hat{Y}_E^{(1)})}}$		$deff\,(\hat{Y}_E^{(i)}) = \frac{\overline{M\hat{S}E(\hat{Y}_E^{(i)})}}{\overline{\hat{V}(\hat{Y}_E^{(1)})}}$	
	NUTS3	NUTS4	NUTS3	NUTS4	NUTS3	NUTS4
2. *Greg*	0.0721	0.0617	0.0683	0.0557	0.9602	0.9574
3. *Synth_a*	0.9812	0.9977	-0.5524	0.7234	1.5998	0.2804
4. *Synth_b*	0.8734	0.9863	0.3430	**0.8455**	0.6771	**0.1567**
5. *Synth_c*	0.8909	0.9957	0.0084	0.7594	1.0218	0.2439
6. *Eblup_a*	0.5514	0.8810	0.3768	0.7644	0.6422	0.2389
7. *Eblup_b*	0.6649	0.9294	**0.5154**	**0.8452**	**0.4994**	**0.1570**

Remark: (i) – denotes the number of estimator for which the measure of effectiveness was calculated i=2, ... , 7

Source: Own calculations made on the simulations on POLDATA within the EURAREA project.

More detailed conclusions at this stage of the analysis are as follows:

1. Inconstancy in evaluation
   - The evaluation depends on the character of territorial units, their differentiation, level of aggregation, availability of auxiliary information etc.
   - Bias of synthetic estimators is evident, Synthetic estimators were seriously biased especially for regions of very high 2837 (Ełk), 2228 (Słupsk), 3244 (Koszalin) or very low unemployment 3042 (Poznań), 1422 (Warszawa) and 1217 (Kraków)
   - If REE would be treated as a synthetic evaluation parameter, the following estimation precision was obtained:

NUTS 3	0.1738 — *Eblup_b*	0.1825 — *Synth_b*
NUTS4	0.2579 — *Synth_b*	0.2869 — *Eblup_b*

   - Distributions of REE obtained for indirect estimators are heterogeneous. Bigger stability is observed for EBLUP estimators. Their characteristics: rightward skewness, one peak and big kurtosis are also more preferable.

2. Gain in effectiveness
   - However measures of effectiveness obtained for small area estimation in comparison with direct estimates show impressive gain in estimation precision.
   - At NUTS4 level, the best results obtained for SYNTH_B and EBLUP_B inform that MSE is about 85% smaller than the variance of the direct estimator.
   - Gain obtained for NUTS3 level is also great and amounts to about 50% when applying EBLUP estimator with area level covariates instead of direct one.
3. Suggestions for further research
   - Searching for strongly correlated covariates (also at area level)
   - Delimitation of similar regions form the point of view of the estimated variable (in example degree of social-economic development)
   - Constructing different models for delimited types of regions and providing the estimates separately

The results obtained are not stable for the area level and, as shown in the EURAREA project, for the target variable. This was indicated by the analysis of $\hat{REE}_{MSE}$ distributions. The biggest stability was observed for EBLUP estimators. Their characteristics: rightward skewness, one peak and big kurtosis are also more preferable. Further simulations are carried out in order to determine the optimal sample design. Results of simulation procedures to test the properties of estimators for small areas cannot, in most cases, be generalised but reflect conditions of a particular situation. This was also the case with the present study, which is essentially experimental.

REFERENCES

DECAND, G., 1996, *The Nomenclature of Territorial Units for Statistics (NUTS),* Baden/Vienna 1996.

EURAREA documents from years 2000-2003, ONS, EURAREA, IST 2000—26290.

GHOSH, M., Rao J.N.K., 1994, *Small Area Estimaton: An Appraisal,* „Statistical Science", Vol. 9, No. 1.

HEADY, P., 2003, *Early results from the EURAREA project*, Bulletin of the International Statistical Institute 54th Session, Proceedings, Berlin 2003.

KALTON, G., KORDOS, J., PLATEK, R., 1993, *Small Area Statistics and Survey Designs,* GUS, Warszawa.

KORDOS, J., PARADYSZ, J., 1999, *Some Experiments in Small Area Estimation in Poland,* [in:] *Small Area Estimation*, International Association of Survey Statisticians Satellite Conference Proceedings, Riga 20—21 August 1999, Latvia.

LEHTONEN, R., VEIJANEN, A., 1998, *Logistics Generalized Regression Estimators*, "Survey Methodology", June 1998, 24, 51—55.

RAO, J.N.K., 1999, *Some Recent Advances in Model-Based Small Area Estimation*, "Survey Methodology", December 1999, 25.

SäRNDAL, C.E., SWENSSON, B., WRETMAN, J., 1992, *Model Assisted Survey Sampling,* Springer — Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

ANNEX — Empirical measures of estimation precision

I. Domain specific:

1. The mean value obtained from 100 simulations (*p = 1,...,100)*:
   $$\hat{\bar{Y}}_d = \frac{1}{1000}\sum_{p=1}^{1000}\hat{Y}_{dp}$$
   was considered as the expected value
2. The empirical variance of the estimator was defined as:
   $$\hat{V}(\hat{Y}_d) = \frac{1}{999}\sum_{p=1}^{1000}(\hat{Y}_{dp} - \hat{\bar{Y}}_d)^2$$
3. The empirical value of MSE (mean square error) was calculated according to the following formula: $$M\hat{S}E(\hat{Y}_d) = \frac{1}{999}\sum_{p=1}^{1000}(\hat{Y}_{dp} - Y_d)^2$$
   where Y_d is the „true" value of the estimated variable in the population, in domain *d*
4. Standard estimation error was defined in two ways:
   On the base of the empirical variance: $_V\hat{S}(\hat{Y}_d) = \sqrt{\hat{V}(\hat{Y}_d)}$
   On the base of the empirical MSE: $_{MSE}\hat{S}(\hat{Y}_d) = \sqrt{M\hat{S}E(\hat{Y}_d)}$
5. Analogous procedure was applied to estimate the empirical value of the REE (relative estimation error)
   On the base of the empirical variance: $$_V R\hat{E}E(\hat{Y}_d) = \frac{\sqrt{\hat{V}(\hat{Y}_d)}}{\hat{\bar{Y}}_d}$$
   On the base of the empirical MSE: $$_{MSE} R\hat{E}E(\hat{Y}_d) = \frac{\sqrt{M\hat{S}E(\hat{Y}_d)}}{Y_d}$$

II. Synthetic measures

1. Average value of the empirical MSE:
   $$MSE = \overline{M\hat{S}E(\hat{Y})} = \frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{999}\sum_{p=1}^{1000}(\hat{Y}_{dp} - Y_d)^2\right] = \frac{1}{D}\sum_{d=1}^{D}M\hat{S}E(\hat{Y}_d)$$
2. Average value of the empirical variance:

$$Var = \overline{\hat{V}(\hat{Y}_d)} = \frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{999}\sum_{p=1}^{1090}(\hat{Y}_{dp} - \hat{\bar{Y}}_d)^2\right] = \frac{1}{D}\sum_{d=1}^{D}\hat{V}(\hat{Y}_d)$$

3. Average value of the standard estimation error defined in two ways: on the base of the empirical variance and on the base of the empirical MSE:

   $\sqrt{Var} = \overline{{}_V\hat{S}(\hat{Y}_d)} = \frac{1}{D}\sum_{d=1}^{D}{}_V\hat{S}(\hat{Y}_d)$ or

   $\sqrt{MSE} = \overline{{}_{MSE}\hat{S}(\hat{Y}_d)} = \frac{1}{D}\sum_{d=1}^{D}{}_{MSE}\hat{S}(\hat{Y}_d)$

4. Average value of the REE defined upon the empirical variance and upon the empirical MSE:

   $REE_{Var} = {}_V\overline{R\hat{E}E} = \frac{1}{D}\sum_{d=1}^{D}{}_V R\hat{E}E(\hat{Y}_d)$ or

   $REE_{MSE} = {}_{MSE}\overline{R\hat{E}E} = \frac{1}{D}\sum_{d=1}^{D}{}_{MSE} R\hat{E}E(\hat{Y}_d)$

5. Absolute relative bias: $A\hat{R}B(\hat{Y}_d) = \frac{\left|\hat{\bar{Y}}_d - Y_d\right|}{Y_d} = \frac{\left|\frac{1}{1000}\sum_{p=1}^{100}\hat{Y}_d - Y_d\right|}{Y_d}$

6. Absolute relative error: $A\hat{R}E(\hat{Y}_d) = \frac{\frac{1}{1000}\sum_{p=1}^{1000}\left|\hat{Y}_d - Y_d\right|}{Y_d}$

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 777—783

AN EFFICIENCY OF MODIFIED SYNTHETIC ESTIMATOR FOR THE POPULATION PROPORTION: A MONTE CARLO ANALYSIS

Tomasz Jurkiewicz[1], **Krzysztof Najman**[2]

ABSTRACT

The problem of insufficient number of sample observations representing a given population domain of interest (small area) can be solved by applying estimators, which will be able to combine sample information from the given domain with information about sample units representing other domains. Synthetic estimation technique assumes that the distribution of the variable of interest is identical in the given domain and in the entire population. This assumption, however, is rarely met, and as a result one obtains large estimation errors.

In this paper a two-stage estimation procedure is suggested. The first stage consists in applying some distance measures to identify the degree of similarity between the sample units from the investigated domain and sample units representing other domains. In second stage, those units, which turned out to be similar to units from domain of interest, are used to provide sample information with specially constructed weights.

Authors present results of the suggested procedure using Monte Carlo experiments based on data obtained form a continuing vocational training survey of enterprises.

Key words: synthetic estimation, small domain, distance measures.

1. Introduction

The process of economic and social development results i.a. in a growing demand for statistical information. Random sample surveys can be regarded as an effective way of satisfying that demand. However, because of various organisational and financial constraints those surveys may not able to supply credible data for a specific divisions of the population into smaller domains of

[1] University of Gdańsk, Gdańsk, Poland, e-mail: t.jurkiewicz@zr.univ.gda.pl
[2] University of Gdańsk, Gdańsk, Poland, e-mail: k.najman@zr.univ.gda.pl

interest. Insufficient number of observations representing a particular domain may be an obstacle in applying certain statistical techniques or may lead to considerable errors of estimation (Bracha, 1996). One possible way of solving this problem is constructing such estimators, which could use information about other components of a sample, namely those coming from outside the particular part of the population. Another possibility is to use additional information from outside the sample to estimate parameters of a defined subpopulation.

The "small domain" (small area) is defined as a domain of studies, for which (Jurkiewicz 2001):

- information is essential from the data user's point of view,
- it is not possible to obtain required information using the direct estimation method, because the size of the sample is too small, or when the information acquired with indirect methods is more credible.

There is no reason for which the scope of statistics of small areas should be confined to territorial (administration) units. From a methodological point of view it does not make any difference whether we consider a subpopulation of one territory or a subpopulation isolated according to any other method (Kordos, 1999).

The main aim of this paper is an attempt to evaluate efficiency of a modified synthetic estimator. The parallel aim of the study is to verify the modified synthetic estimator empirically on the basis of a sample survey called Continuing Vocational Training Survey (CVTS) conducted in Polish enterprises.

2. Estimators of small domains

The essence of indirect estimation consists in "borrowing the information" to strengthen the estimation in the domain being of interest to the statistician. In case of sampling surveys, it is possible to use the following sources of additional data (Kordos 1999, Domański, Pruska 2001):

- other domains in the sample;
- information about the number of particular strata and the number of domains in the studied population;
- information about the values of an additional variable in a sample;
- information about values of an additional variable in the studied population;
- other available data, e.g. data from studies of other periods.

The direct estimator of an unknown parameter ΘY_d in a small domain is the Horvitz-Thompson estimator, known as the expansion estimator. It uses only the data about randomly drawn components of a sample belonging to the small domain, that way is not a truly small domain estimator, but it is a datum for other estimators. The HT estimator is, however, unbiased, but because of the small size of the sample its variance is usually high. This estimator will have the following form for the proportion parameter:

$$ {}_{HT}p_d = \frac{k_d}{n_d} \tag{1} $$

where k_d and n_d stand for the number of elements distinguished in the domain d, and the size of the small domain d, respectively.

Synthetic estimation constitutes one of the first propositions of solving the principal problem of estimation for small domains, which stems from the insufficient size of a sample. Here, the assumption is made that the structure of the studied population in and outside the small domain is uniform, which allows to use the information from the whole sample to estimate characteristics for the domain of interest. This assumption may be weakened in some cases to require similarity of only certain parameters in the population and in the investigated domain. For the proportion, the estimator adopts the form of the following statistics:

$$ {}_{syn}p_d = \frac{k}{n} \tag{2} $$

where k and n denote the number of elements distinguished in the sample and the size of the whole sample, respectively.

While applying the synthetic estimation it is very important to pay careful attention to the problem of efficiency of the adopted model. The farther distance between the assumptions, which lie at the base of the estimation, and the reality considered, the more biased will be the estimators. It has to be borne in mind, that firstly, the bias may be of considerable size, and secondly, it is in no way taken into account in formulae for mean square errors and estimators of errors.

3. Modified Synthetic Estimator (MES)

The assumption about the compatibility of structures of the population and the domain is frequently not met, in particular in case of specific domains, which results in large estimation errors. To solve this problem one can suggest strengthening the estimation process by modifying the estimator with information from components similar to the studied one. The proposed procedure of estimation is carried out in two stages. The first step consists in establishing, which components are similar to the units from studied domain. Weights for additional information could be calculated in relation to the degree of similarity. Thus, data from similar components will imply a relatively high value of the weight, while data from distant components will have a relatively lower weight or will not be taken into account at all. The proportion estimator will adopt the following form:

$$_{MES}p_d = \frac{k_d + \sum_{i=1}^{n_{\sim d}} y_i w_i}{n_d + \sum_{i=1}^{n_{\sim d}} w_i} \tag{3}$$

where k_d – number of elements distinguished in the sample belonging to the domain,
n_d – size of the sample in the domain d,
w_i – weights for the components from outside the small domain,
y_i – values of the studied zero-one feature.

The establishment of the similarity of the studied feature to other features in the population may be carried out using one of the methods of multidimensional analysis[1]. It is worth to pay attention to advantages of the MES estimator, especially an opportunity of using information derived from outside the study. Namely, while establishing the similarity between domains it is possible to use data from completely different, e.g. earlier studies or the available information about the population. In such a case it is also possible to calculate the estimators of parameters for a domain, which is not represented in the sample.

A different possibility to use additional information about units from outside the small domain provides an evaluation of similarities between units. The first proposal is based on a k-means grouping method. Components belonging to the domain of study have to be classified into k centres. Weights for components from outside the small domain should be calculated proportionally to the distance from component to the nearest grouping centre. Although this method seems to be appropriate, the results received in earlier studies made by authors do not look encouraging.

The second proposal, which was applied in this paper, is based on individual distances between all units in the sample. The presumption was undertaken that the weight of component from outside domain of interest should be run on the distance to the nearest component from small domain. Euclidean measure of distance between components was used in this study. The weight w_i = 1 was assigned for two nearest components to each component from small domain. All others components have had weight equal zero. In consequence effective size of the small domain sample for MES estimator was 3 times higher than for HT.

[1] Some results of analysis of MES estimator were presented at 21st (2002) and 22nd (2003) Annual Conferences on Multivariate Statistical Analysis and will be published in Acta Universitatis Lodzensis, Folia Oeconomica in 2004.

4. A Random Sample Survey of Continuing Vocational Training

The study of the continuing vocational training was carried out in the task 1.4 of project for Ministry of Economy, Labour and Social Policy. The studied population consisted of enterprises which employed at least 10 persons and were registered in the REGON (oficial business register) in 2003. Some sectors were excluded from the population, such as public administration, health services and education. The size of the sample was calculated at the level of 15000 enterprises. A questionnaire construed for the sake of the study included 18 wide questions provides almost 600 variables. The sample received as a result of enquiry and interviews included 15012 components.

In the studied group of enterprises the number of companies from each province was from 433 (2.9%) to 1471 (9.6%). Those numbers should be sufficient for a credible description of the province as a whole, but is insufficient for more detailed study with the use of direct estimators. Thus the description of those domains could be based on other methods of estimation, giving more credible results. One of those possibilities is to consider any province as a small domain and to apply the methods of estimation used for small domains. On the other hand, sizes of samples usually are significantly smaller than 15 thousands individuals. In that case even the description of main domains could be based on methods giving more credible results. The main aim of the study was evaluation of modified synthetic estimator in such types of sample researches.

5. Evaluation of properties of the MES estimator

To evaluate the MES estimator the bootstrap method was used. At the beginning 47 variables was selected for evaluating similarities. In subsequent repetitions 1000 components were drawn independently at random, considering components that were found originally in the sample as the population in question. For each of 1000 simulation the euclidean distances between all components was counted. Subsequently, the values of expansion, synthetic and MES estimator ware calculated for 39 investigated variables for all 16 provinces.

To evaluate the properties of estimators of the ΘY_d parameter in this study, the mean bias of estimator in all *s* experiments was used, calculated according to the following formula:

$$BIAS_f = \frac{\sum_{i=1}^{s}(P_{f,i} - \Theta Y_d)}{s} \cdot 100 \tag{4}$$

where: $P_{f,i}$ is the value of the *f*-th estimator in the *i*-th experiment;
ΘY_d is the real value of proportion of the feature *Y* in domain *d*.

The second element of the evaluation was the (square) root of the mean square error, calculated according to the following formula:

$$sqr(MSE_f) = \sqrt{\frac{\sum_{i=1}^{s}(P_{f,i} - \Theta Y_d)^2}{s}} \cdot 100 \qquad (5)$$

The studied characteristics were the structural indices, that is why the bias and the mean error were expressed in percentage terms for the sake of transparency.

After the experiment the value of the third relative moment was calculated, that is the measures of the skewness of distribution of the acquired values of estimations and the Kołmogorow-Smirnow test for normality of the estimator distribution was applied.

6. Results of the study

Effective number of the small domain sample for MES estimator, measured as sum of weights, was three times higher than original, but from 3,5 to 12 times smaller than for synthetic estimator. In consequence, the variance of MES was smaller than the variance of expansion estimator, but was much higher than the variance of synthetic estimator. The bias of modified synthetic estimator in 70% cases was smaller than bias of synthetic estimator, the average bias of MES came 1,7% and for synthetic 2,5%.

The efficiency of MES estimator was higher than expansion in over 80% of cases, but only in 16% cases than synthetic. It could not be considered as very good result, but in dozen or so cases outcome of synthetic estimator was far from acceptable. Almost the same number of cases, when the expansion estimator was more efficient than synthetic, could be observed. The MES estimator was the most efficient one in up to 10% of cases.

Efficiency of estimator depends on two factors, bias and variance. The bias of synthetic estimator is equal to difference between values of investigated variable in small domain and population. The bias of MES estimator depends to a large extent on this difference. The efficiency of MES was higher when the difference was distinct, but not too high. When the bias was too significant mean square error of synthetic estimator was highest one, but MSE of expansion estimators becomes smallest one.

Slightly different results of efficiency of estimators could be observed when the size of sample from small domain was growing. For the experiment three small domains were consolidated into one and the simulation for such domain was carry out. The variance of the expansion estimator was in that case much smaller and the efficiency of synthetic and MES estimators become relatively worse.

The distribution of MES estimator wasn't as close to normal as distribution of synthetic estimator, but much closer than distribution of expansion estimator. Skewness of distribution of MES estimator was higher than distribution of synthetic estimator, but there wasn't any influence on efficiency.

7. Conclusions

Application of the modified synthetic estimator seems to be a reasonable alternative to the estimation of parameters of distributions in small domains, in particular in those domains, which significantly, though not too much, differ from the population. Its distribution has relatively lower variation than the distribution of the expansion one. Even if bias of the modified synthetic estimator may be quite considerable, in a vast majority of cases it is much smaller than the bias of the synthetic estimator. The distribution of the estimator in many cases may be considered as normal or close to normal.

The most important seems to be the proper choice of a set of variables to similarity investigation. In this study investigated variables and variables used to similarity analysis were slightly correlated. It could affect efficiency of modified synthetic estimator.

REFERENCES

BRACHA, C. (1996) *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

ROMAŃSKI, C., PRUSKA, K. (2001) *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

JURKIEWICZ, T. (2001) Efficiency of Small Domain Estimators for the Population Proportion: A Monte Carlo Analysis, *Statistics in Transition*, Vol. 5, No 2, pp. 237—248.

KORDOS, J. (1999) *Problemy estymacji dla małych obszarów* (Problems of estimation in small domains), Wiadomości Statystyczne, No. 1, pp. 85—101.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 785—796

APPLICATION OF THE HIERARCHICAL BAYES ESTIMATION TO THE POLISH LABOUR FORCE SURVEY

Jan Kubacki[1]

ABSTRACT

The author presents the application of hierarchical Bayes methods to the estimates of unemployment size for small areas applied to the Polish Labour Force Survey (PLFS). Constructed model includes the data obtained from published results of PLFS for regions in Poland and 2002 Census data. Also second model using PLFS data for counties in łódzkie region together with some administrative data was prepared. This model has two variants: first uses the PLFS data from 1999 year and second uses data from PLFS for 2002 year and data from 2002 Census. The evaluation of quality of these methods was presented with comparison to the earlier used methods (direct estimation).

Key words: Labour force survey, hierarchical Bayes estimation, Gibbs sampling, empirical Bayes estimation, small area estimation.

1. Introduction

During the last decade many efforts concerning the development of small area estimation was made. This is connected both with progress of theory and enhancement of the statistical methodology. The availability of the computer software and also possibility of using administrative data as a remedy for improving the quality of surveys has also the influence over the obtaining the reliable data at the local scale. Among variety of small area methods the hierarchical Bayes (HB) is one of the most promising small area technique. The hierarchical Bayes approach was used in many statistical problems, also in social statistics. For example the research of Datta, Lahiri, Maiti and Lu (1999) show the possibility of using the HB approach in estimation of unemployment at the local scale. Results prepared in this paper show that applying such methods can provide

[1] Central Statistical Computing Centre, Al. Niepodleglosci 208, 00-925 Warsaw, Poland.
E-mail: j.kubacki@stat.gov.pl

lower coefficient of variation of the estimates than using the well-known Fay-Herriot model (Fay, Herriot, 1979) applied by Bureau of Census.

In this paper, also the hierarchical Bayes approach was made. Here two unemployment size models are presented. First explains the region (voivodship) unemployment size obtained for PLFS in IV quarter of 2002 year. The second explains PLFS data obtained in IV quarter of 1999 year for lodzkie voivodship. This article is a continuation of previous efforts that concern the estimation of the size of unemployment for łódzkie voivodship using various techniques of small area estimation.

First results was presented at the IASS Satellite Conference on Small Area Estimation (Kubacki, 1999) when a proposal of using some of the estimation techniques employed in small area statistics was presented. These methods (including post-stratification estimator, synthetic estimator and GREG estimator) were used in determining the size of unemployment at the Local Labor Office (RUP) for the łódzkie voivodship. Using administrative data of the unemployment as an auxiliary variable was suggested. The quality of the estimation was evaluated using random groups technique.

The enhanced results of this paper was presented next year (Kubacki, 2000a,b) particularly at the conference: "Statystyka regionalna w służbie samorządu lokalnego i biznesu", Kiekrz, 5—7, June, 2000, where selected small area statistics method of unemployment estimation was presented. These methods were used for determination the size of the unemployment at the county level for łódzkie voivodship.

The comparison between some Bayes methods for estimation of the unemployment at the local level was presented at the 5th International Science Conference „Regional Statistics in Uniting Europe“ Łagów 2nd-5th September 2002. The measure of the efficiency of used Bayes techniques was the ratio between the variance obtained for estimation using empirical Bayes estimation and variance for estimation using "standard" techniques (such as direct estimation).

The application of hierarchical Bayes estimation presented here is possible partially due to availability of unemployment estimates and sampling error estimates. This estimates is used in construction of the hierarchical model, particularly in determining the priors used in simulation. The main assumption in this case is, that the distribution of the priors is normal and the variance of this distribution is the same as the sampling variance. This can be valid especially in situation, where estimates are more reliable (for regions — voivodships). However in instance of estimates that was done for Local Labour Offices (LLO — in other words — for counties or poviats) for some LLO's the LFS does not provide any data. In such case the approximated value of variance was used, which utilizes the assumption, that the variability for which there is no data such is similar to those, whose size is comparable (most often for rural counties). This can provide consistent estimation of variance for the LLO, for which no data is available.

2. About the Polish Labour Force Survey

Polish Labour Force Survey (PLFS) was originally designed as quarterly survey. Such scheme was used until 1999 year (Szarkowski and Witkowski, 1994). In that year partial redesign of the survey was made. The outline of the sampling plan is similar to that using before the 1999 year. Currently, the following principles was used (Kordos, Lednicki, Żyra, 2002):

The whole sample for each quarter has about 22 thousand households and it includes every person aged 15 and above, that belongs to the interviewed household. The rotation pattern, used for LFS, was not changed since the second quarter 1993, and can be summarized as follows:

- In every quarter 4 elementary samples are drawn, which, with respect of the continuously character of the survey is divided to 13 weekly elementary samples, and consist with 6110 or 6175 dwellings.
- Like in 1999 year, in every quarter, partial exchange of the elementary samples is performed. At each quarter four following samples is used: two elementary samples, that was employed during last quarter, one new introduced elementary sample and one elementary sample that was introduced the year before.
- Each elementary sample is selected independently, and each sample is used according to the following rule: two quarters in survey, two quarters pause and again two quarters in survey.

The sampling process, analogously like before 1999 year, is performed using two-stage sampling. The primary sampling units — in urban areas — are census regions, and in rural areas — the enumeration districts. The secondary sampling units are dwellings. During the first stage sampling, the stratification procedure is used, that utilizes the territorial division of the country and the urban — rural criterion. In case of strata, that contains villages and small towns — 8 dwellings are drawn from each primary sampling units, in case of medium-size strata 6—7 dwellings was drawn, and in case of large cities — 5 dwellings was drawn. At first sampling stage the Hartley-Rao sampling scheme was used, and at secondary sampling stage the simple random selection was used.

3. Techniques of estimation and model construction

Methods that uses hierarchical Bayes (HB) approach is based on assumption, that the prior distribution $f(\lambda)$ of model parameters λ is known and the posterior distribution $f(\mu|\lambda)$ of small area parameters μ (which are the target of such inference) given the data y is obtained. The Bayes theorem used here is based on the following reasoning:

Let us suppose, that the we must obtain the desired posterior density:

$$f(\mu \mid \mathbf{y}) = \int f(\mu, \lambda \mid \mathbf{y}) d\lambda \tag{1}$$

Using Bayes inference we have:

$$f(\mu, \lambda \mid \mathbf{y}) = \frac{f(\mathbf{y}, \mu \mid \lambda) f(\lambda)}{f_1(\mathbf{y})} \tag{2}$$

where *f₁(*y*)* is the marginal density of y and has the form:

$$f_1(\mathbf{y}) = \int f(\mathbf{y}, \boldsymbol{\mu} \mid \lambda) f(\lambda) d\boldsymbol{\mu} d\lambda \tag{3}$$

In simple case, the posterior density can be obtained analytically, what is involved with numerical integration of the marginal density (3). However, in composite situation, such integration becomes intractable. In recent years Markov chain Monte Carlo (MCMC) methods are often used in evaluating the target posterior density. These methods become more popular, mainly because of the availability of the statistical software and fast computers, which can perform massive computing. This technique uses the procedure called Gibbs sampling that was first formally introduced by Geman and Geman (1984) with connection of image processing techniques. This HB procedure allows obtain reliable estimates of both the parameter estimated and its sampling variance. In paper presented here some efforts was made to apply such methods to results of Polish Labour Force Survey.

In this paper two classes of models was used. First, that was used in country estimates can be described similarly as two level model which incorporates both priors on area estimates and priors on model parameters. General form of such model can be presented as follows:

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{Z}_i \boldsymbol{\alpha} + \mathbf{X}_i \mathbf{v}_i + \mathbf{e}_i, \text{ i=1,...,m} \tag{4}$$

$$y_i \mid \beta, \sigma_e^2 \overset{ind}{\sim} N(\mathbf{x}_{ij}^T \beta, \sigma_e^2) \tag{5}$$

$$\beta \mid \alpha, \Sigma_v \overset{ind}{\sim} N(\mathbf{Z}\alpha, \Sigma_v) \tag{6}$$

Such approach is the HB version of two level models, that is a special case of general linear mixed model with block diagonal covariance structure. In the case considered here, there is an assumption, that the priors of the estimates ($y_i \mid \beta, \sigma_e^2$) are obtained from survey sample and model priors ($\beta \mid \alpha, \Sigma_v$) can be evaluated by construction of model using direct estimates.

Because of the lack of availability of reliable variance estimates on the unit level and model priors, the second model has the similar form as the basic area level model, where only "flat" prior on β and ψ_i are given. Such model can be described as follows:

$$\hat{\theta}_i \mid \theta_i, \beta, \sigma_v^2 \overset{ind}{\sim} N(\theta_i, \psi_i) \tag{7}$$

$$\theta_i \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{z}_\mathbf{i}^\mathrm{T}\beta, b_i^2\sigma_v^2) \tag{8}$$

Similar methods was used by Datta et al. (1999), who present the application of hierarchical Bayes method using time series generalization of widely used cross-sectional model in small-area estimation.

In the case of estimates for regions (voivodships) the data from IV quarter of 2002 year was used (GUS, 2002). It also contains sampling variance estimates, which can be used in construction of the hierarchical model. This model contains 5 independent variables and explains the size of the unemployment at the regional scale using the following exploratory variables (the data comes from National Population and Housing Census 2002):

- Number of occupied dwellings — A
- Size of the working population — B
- Number of employed persons — C
- Number of unemployed persons — D
- Number of non-active persons — E

In the case of estimates for counties the model was similar, but because of the nature of data only three of this variables was used (i.e. number of occupied dwellings, number of employed persons and number of registered unemployed persons). Such selection was done partially because of the availability of data for areas corresponding to the county, and also due to the stepwise regression applied to the initial model having identical type of exploratory data as that for census data.

The counties estimates was prepared by means of specially designed software, that uses the microdata for obtaining both estimates and sampling variance. This software uses direct (Horvitz-Thompson) estimator in estimation of the unemployment size.

This estimator uses inclusion probabilities π_i constructed individually for each county that takes into consideration sampling scheme and non-response coefficient. The sampling variance was computed using random group techniques (Wolter, 1985). This method was similar to those used by Central Statistical Office in Polish LFS.

$$\hat{s}_{dg}^2 = \frac{1}{k(k-1)}\sum_{s=1}^{k}(y_{dgs} - y_{dg})^2$$

where *k* denotes number of random groups in sample, *s* stands for subsample.

As it is presented in Table 2, there are some counties for which there is no data at all. Using the model, it may be possible to obtain estimates of such data. However, as it will be shown below, such estimates are not always sensible.

The initial model for 2002 year can be presented in the consecutive form

$$y_i = \alpha_1 + \alpha_2 A_i + \alpha_3 B_i + \alpha_4 C_i + \alpha_5 D_i + \alpha_6 E_i$$

The specification for the region model is as follows. In both model, there is an assumption about normality of model parameters alpha[i] and estimated variables Y[p]. N is equal 16 for the whole country model and 21 for counties model.

```
  alpha[1] ~ dnorm(-11.40, 0.009)
  alpha[2] ~ dnorm(-0.009, 140.84)
  alpha[3] ~ dnorm(-0.45, 63.07)
  alpha[4] ~ dnorm(1.56, 32.28)
  alpha[5] ~ dnorm(0.29, 174.74)
  alpha[6] ~ dnorm(0.37, 49.67)
  for(p in 1 : N) {
  Y[p] ~ dnorm(mu[p], tau[p])
  mu[p] <- alpha[1] + alpha[2] * A[p] + alpha[3] * B[p] + alpha[4] * C[p] +
alpha[5] * D[p] + alpha[6] * E[p]
  }
```

The values of model parameters alpha[i] is obtained from traditional regression model. Both the mean and variance is obtained using this technique. The tau[p] is the sampling variance, that is estimated from the sample.

The specification for the model for counties is similar, but contains only values that are valid for that case (i.e. number of occupied dwellings, number of employed persons and number of registered unemployed persons). Here two cases of model was used. First has no priors on tau[p] — it is replaced by model variance (for all considered values) obtained from initial stepwise regression, and second uses empirical Bayes estimates as initial values for the MCMC simulation. There is also an assumption, that the model parameters can be used from regression model that uses Bayes estimates.

4. Results and discussion

Summary of the simulation results are presented below. These results are based on simulation made after the “burn-in” period, as it is suggested in Rao (2003). The summary of these results is presented in Table 1. The comparison of the CV before and after using MCMC simulation reveals, that there is significant benefit in using such methods. This is shown in Figure 1. However in the case of counties model, such benefit can be tentative, mainly due to the few values of the estimates for which no data is available. In this case, also the benefit in reduction of variance is observed, but it is not so evident as for the whole country model — see Figure 2. When the estimates are more reliable (such as those obtained by empirical Bayes procedure) both the model and the MCMC simulations become more stable. This can be caused by lower variance of the estimates (such as those acquired using empirical Bayes estimation) and more reliable auxiliary, for example data that comes from 2002 Census.

The year 2002 results have similar characteristics with results in paper that was presented by Datta et al. (1999). The comparison of the estimates obtained by MCMC method is consistent with initial results (i.e. The PLFS estimates for regions). This can be seen in Table 1. The benefit of using the HB procedure is evident for every region.

Table 1. Comparison of unemployment estimates from LFS (direct) and from hierarchical Bayes (HB) estimation by region in 4th quarter 2002

Region	Unemployment estimates from:		Coefficient Of variation for:		Reduction of st. dev.
	Direct est.	HB est.	Direct est.	HB est.	$s_{HB.} / s_{Dir}$
	'000		per cent		per cent
Dolnośląskie	344.0	327.3	6.0	2.6	44.3
Kujawsk.-pom	217.0	217.6	6.9	2.0	29.0
Lubelskie	185.5	184.3	7.4	3.0	40.8
Lubuskie	125.0	111.3	7.2	3.4	47.5
Łódzkie	265.0	251.8	5.7	2.8	48.7
Małopolskie	242.0	240.3	7.0	3.5	49.2
Mazowieckie	391.0	398.4	7.8	4.2	56.1
Opolskie	70.5	62.5	9.6	7.0	72.6
Podkarpackie	166.0	170.0	6.6	3.0	44.8
Podlaskie	92.0	88.4	10.9	4.5	41.3
Pomorskie	186.0	193.0	7.3	2.3	31.2
Śląskie	355.0	363.7	5.8	3.8	64.7
Świętokrzyskie	110.0	133.8	8.2	2.8	34.3
Warm.-mazurs.	150.0	164.3	7.3	2.9	39.7
Wielkopolskie	277.5	264.5	6.8	3.5	51.8
Zachodn.-pom.	197.0	195.6	6.3	2.7	43.3

Source: own calculations according to accepted models.

Detailed analysis of data using previous evaluation of applied model shows, that quality of such initial estimates is crucial for obtaining stable results of hierarchical Bayes estimation. The coefficient of determination (R^2) in case of 2002 estimation and 1999 estimation using Empirical Bayes is larger than 0.9, what causes (similar like in case of Empirical Bayes estimation) that the simulation results has regular characteristics. However in the case of 1999 data, when no reliable estimates are available, the simulation was made using the assumption, that the variance components are unknown — or unreliable. As it was stated earlier — only “flat” prior on β and ψ_i are given. Such assumption gives more consistent estimates, which can eliminate unreliable values of the HB estimates (for example — possible negative values for data, that should greater than zero, i.e. unemployment size).

Table 2. Comparison of results obtained from sample and from hierarchical Bayes (HB) for unemployment model applied to LLO’s for łódzkie voivodship (data for IV quarter of 1999)

County	Unemployment estimates from:		Coefficient Of variation for:		Reduction of st. dev.
	Direct est.	HB est.	Direct est.	HB est.	$s_{HB.} / s_{Dir}$
	‘000		Per cent		per cent
Bełchatów	23.7	23.9	30	9	29.7
Kutno	12.2	10.5	47	9	19.0
Łask	14.3	10.2	62	12	19.3
Łęczyca	1.5	1.8	57	44	77.6
Łowicz	-	2.3	-	34	-
Łódź	0.7	5.8	61	14	23.0
Łódź-Wschód	40.5	39.6	13	6	45.9
Opoczno	8.1	6.2	32	15	46.7
Pabianice	7.8	7.0	49	10	20.4
Pajęczno	10.3	2.9	25	26	106.0
Piotrków Tryb.	15.1	15.9	36	7	19.3
Poddębice	1.5	0.5	96	-	-
Radomsko	8.8	11.2	41	10	24.5
Rawa Mazow.	3.7	2.0	104	42	40.4
Sieradz	16.2	12.7	41	9	22.2
Skierniewice	4.3	4.9	162	22	13.5
Tomaszów Maz.	5.7	11.5	39	11	28.6
Wieluń	2.9	6.1	102	12	11.7
Wieruszów	1.5	1.2	57	70	123.4
Zduńska Wola	-	3.5	-	22	-
Zgierz	13.9	14.7	45	6	13.4

Source: own calculations according to accepted models.

Table 3. Comparison of results obtained from sample and empirical Baeys procedure (EB) and from hierarchical Bayes (HB) for unemployment model applied to LLO's for łódzkie voivodship (data for 1999 year — obtained using synthetic estimation)

County	Unemployment estimates from:		Coefficient of variation for:		Reduction of st. dev.
	EB est.	HB est.	EB est.	HB est.	$s_{HB.} / s_{EB}$
	'000		Per cent		per cent
Bełchatów	23,9	23,8	21,7	9,9	45,5
Kutno	10,9	9,7	24	6,5	26,8
Łask	10,6	9,2	35,4	7,9	22,3
Łęczyca	1,7	2,2	186,8	15,8	8,5
Łowicz	2,2	2,4	143,8	15,9	11
Łódź	5,7	5,9	55,1	7,1	12,9
Łódź-Wschód	39,9	37,7	9,1	7,2	78,4
Opoczno	7,3	5,7	27	9,8	36
Pabianice	7,3	7,1	30,3	7,4	24,4
Pajęczno	7,4	3,0	28,3	13,2	46,5
Piotrków Tryb.	15,6	15,0	22,6	5,3	23,5
Poddębice	1,3	1,1	44,7	34,2	76,5
Radomsko	10,1	10,1	25,3	7,3	28,6
Rawa Mazow.	2,6	2,4	50,6	15,5	30,5
Sieradz	13,2	11,9	30,8	6	19,5
Skierniewice	4,6	5,3	31	9,9	31,9
Tomaszów Maz.	7,6	10,3	23,7	7,7	32,3
Wieluń	4,4	6,0	27,5	7,1	25,7
Wieruszów	1,1	1,6	293,4	23	7,9
Zduńska Wola	3,3	3,8	94,8	9,6	10,1
Zgierz	14,5	13,7	21,9	5,4	24,7

Source: own calculations according to accepted models.

The comparison of data presented in Table 2 and Table 3 shows, that using more precise data allows to obtain more reliable estimates. Such consequence of using this kind of approach is consistent with assumption for model presented in table 2. Here the tau[p] was replaced by constant, that was obtained from the initial model error.

Figure 1. Standard deviation estimates before and after using hierarchical Bayes method applied to the country model.

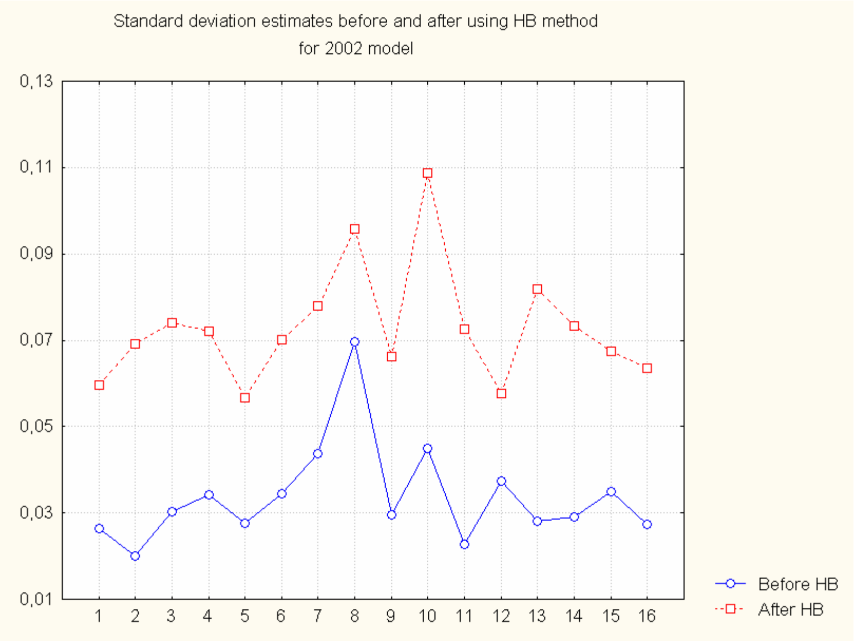
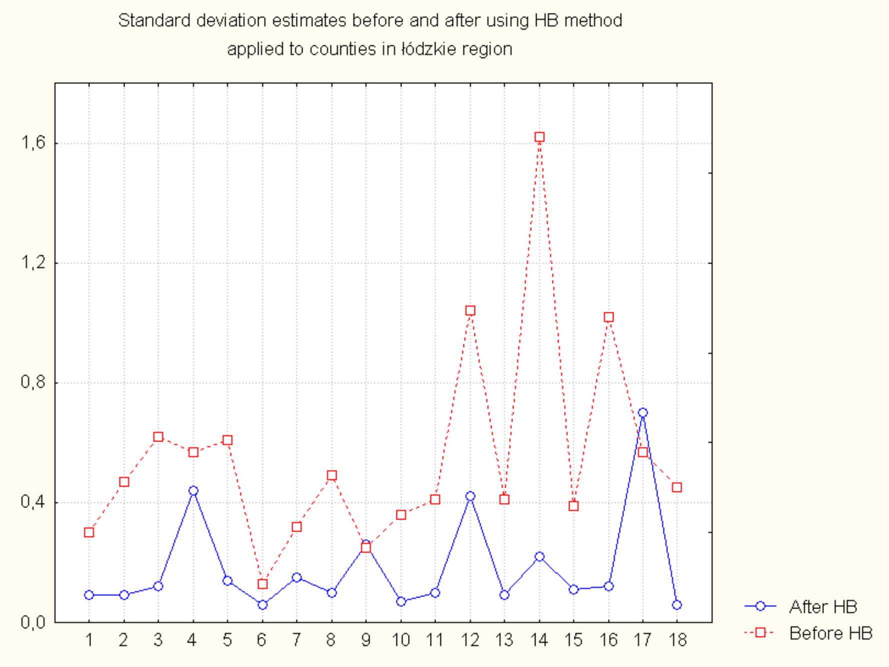


Figure 2. Coefficient of variation estimates before and after using hierarchical method applied to the county model.



5. Conclusions

The comparison of the model for LLO’s (counties) and regions reveals that the sample size and quality of initial model estimates has significant impact on both bias and precision of the estimates using hierarchical Bayes estimation. This is mainly caused by size of variance (related to the sample size) of the initial estimates that was incorporated in model and model quality as well. Further examination of HB procedure, where application of more precise estimates (synthetic estimates) and more reliable auxiliary data (from Census) was applied lead to the conclusion, that using better initial parameters has vital role is HB estimation. Further examination of using better initial estimates such as empirical Bayes estimators reveals more relationships between primary data and results. However such procedure can lead to difficulties, when the statistician wants to evaluate the empirical sense of obtained in such manner estimates and its variance. This is mainly due to the complex nature of obtaining such estimates. The analytical form of estimates like this may be very complex, and perhaps can obscure the real variability in population and caused by sampling scheme. So that using such estimates should be used with care.

Subsequent investigations of such treatment of statistical data of the nature should be done when other techniques of initial estimation are applied.

Acknowledgements

In this paper the simulation was done by WinBUGS 1.3 package. I am thankful to prof. Jan Kordos for helpful discussions and the whole WinBUGS team for possibility using this excellent tool.

REFERENCES

BRACHA, Cz., (2003) Estymacja danych z Badania Aktywności Ekonomicznej Ludności na poziomie powiatów dla lat 1995—2002 (Data Estimation from Polish Labour Force Survey for counties in 1995—2002.) GUS, Warszawa http://sp.stat.gov.pl/spis/estym_danych/index.htm

DATTA, G.S., LAHIRI P., MAITI T., LU K.L., (1999) Hierarchical Bayes estimation of Unemployment Rates for the States of the U.S., *Journal of the American Statistical Association*, 94, 1074—1082.

FAY, ROBERT E., HERRIOT ROGER A., (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association,* 74, 269—277.

GEMAN, S. and GEMAN D. (1984), Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721—741.

GUS (2002), Aktywność Ekonomiczna Ludności Polski IV kwartał 2002 r. (Labour Force Survey in Poland) IV quarter 2002.

KORDOS J., LEDNICKI B., ŹYRA M. (2002) The Household Sample Surveys in Poland, *Statistics in Transition*, 5, 4, 555—590.

KUBACKI, J. (1999) Evaluation of Some Small Area Methods for Polish Labour Force Survey in One Region of Poland, *Proceedings of the IASS Satelite Conference on Small Area Estimation*, Riga, Latvia, 245—249.

KUBACKI, J. (2000a) Some Small Area Estimation Methods for Polish Labour Force Survey in One Region of Poland, *Statistics in Transition*, 4, 5, 769—777.

KUBACKI, J. (2000b), (Selected Methods of Small Area Statistics Applied to the Estimation of the Size of the Unemployment (For Counties — Poviats). In proceedings of the conference: "Statystyka regionalna w służbie samorządu lokalnego i biznesu", Kiekrz, 5—7, June, 2000 (In Polish).

RAO J.N.K (2003) Small area estimation, Wiley-Interscience.

SZARKOWSKI A., WITKOWSKI J. (1994), The Polish Labour Force Survey, *Statistics in Transition*, 1, 4, 467—483.

WOLTER K.M. (1985), Introduction to Variance Estimation, Springer-Verlag.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 797—806

OPTIMAL STRATIFICATION USING RANDOM SEARCH METHOD IN AGRICULTURAL SURVEYS

Marcin Kozak[1]

ABSTRACT

Lednicki and Wieczorkowski (2003) presented a method of stratification in subpopulations, based on the paper of Rivest (2002), which leads to obtain a fixed precision of a considered estimator for the particular subpopulations. They solved such a problem using a numerical simplex method of Nelder and Mead (1965). Because it is not the best solution of the question, (in a sense of an optimality of the simplex method in multivariate optimization problems), the algorithm of random search method is proposed in the paper as a more efficient way of considered stratification. Moreover, a simple modification of a data structure is proposed when using the optimization. Five numerical experiments were carried out to compare the proposed algorithm with the simplex method using data from Agricultural Census 2002 regarding the cereals area.

Key words: global optimization, optimal stratification, random search, simplex method, stratified sampling.

1. Introduction

The most often sampling scheme used in agricultural surveys conducted in Poland by Central Statistical Office is a stratified sampling. The stratification aims to divide the population into groups called strata using a stratification variable in a way that a precision of estimation for a variable of interest is minimal. Therefore there is a need for good stratification methods.

There are many methods of stratification using an auxiliary variable. The description of them can be found in several copies, see e.g. Cochran (1977) or Bracha (1996). In the opinion of the latter the best stratification method is the one proposed by Schneeberger (1970). This method has some disadvantages, i.e. it does not take into consideration a problem of a sample allocation and a

[1] Central Statistical Office, Al. Niepodległości 208, 00-925 Warsaw, Poland.
E-mail: m.kozak@stat.gov.pl, Department of Mathematical Statistics and Experimentation, Warsaw Agricultural University, Poland.

distribution of the stratification variable. Therefore some other methods of the stratification were proposed during last years. These methods take into consideration the question of sample allocation, model-assisted stratification, take-all stratum and some others (see e.g. Godfrey *et al.*, 1984, Lavalleé and Hidiroglou, 1988, Sweet and Sigman, 1995, Niemiro, 1999, Dorfman and Valiant, 2000, Hedlin, 2000, Rivest, 2002, Lednicki and Wieczorkowski, 2003).

Lednicki and Wieczorkowski (2003) presented the method of stratification in subpopulations that leads to given precision of the estimators for the particular subpopulations getting. The authors based on the paper of Rivest (2002). They solved the problem using a numerical simplex method of Nelder and Mead (1965). It is not the best solution of the question, because first, the simplex method is not optimal when considering a large number of variables (Findensein *et al.*, 1974, p. 157), and secondly, it is rather slow method (Brandt, 1999). Therefore we see a need for finding the better algorithm of stratification when using the method of Lednicki and Wieczorkowski.

Niemiro (1999) studied a usefulness of a random search method in the stratification problem. The algorithm proposed by the author did not guarantee that it leads to global optimum of a minimizing function. Furthermore, it would go wrong in a case of a large population, as it requires too many iteration steps; some details are pointed in the paper.

An aim of the paper is to present the modified random search algorithm as a method of the optimal stratification presented by Rivest (2002) and Lednicki and Wieczorkowski (2003). The algorithm has two advantages — it leads to very good results and is quite fast. We will use the method for the data from Agricultural Census 2002 regarding cereals area. The results will be compared with the results of the simplex method.

2. Algorithm of stratification using random search method

Consider a population U consisting of N units. An aim of the stratification is to divide the population U into fixed number, say L, of separate groups, (called strata), i.e.

$$\bigcup_{h=1}^{L} U_h = U, \qquad U_h \cap U_g = \varnothing \text{ for } h = 1,\dots,L, g = 1,\dots,L, h \neq g, \tag{1}$$

in such a way that some objective function depending on this division (1) is minimal. The division (1) is defined by the vector of strata boundaries, say $\mathbf{a}=(a_1,\dots,a_{L-1})^T$.

Let us define the strata boundaries as follows: sort a population by the stratification variable; two stratum boundaries a_{h-1} and a_h defines the stratum h

in such a way, that this stratum consists of the units with the index I in an interval $a_{h-1} < I \le a_h, h = 1,...,L, \quad a_0 = 0, \quad a_L = N.$

First, let us describe the problem of optimal stratification given by Rivest (2002) and Lednicki and Wieczorkowski (2003). The approach of the authors assumes that the population U is right skewed; the units with the biggest value of the survey variable have a very big influence on an accuracy of the estimation. Therefore the authors propose to create so called “take-all” stratum from which all units are taken to a sample. Note that such distribution is typical for business and agriculture data. Rivest (2002) proposed a form of an objective function that minimizes a sample size with respect to a given precision of estimation of a mean of the stratification variable; such approach was adapted by Lednicki and Wieczorkowski (2003) to the problem of the stratification in subpopulations with respect to the fixed precision of the estimation in subpopulations.

Afterwards, our aim is to find such values

$$a_1 \le a_2 \le ... \le a_{L-1}, \tag{2}$$

(the strata boundaries defined above), that minimize the objective function given by (Rivest, 2002, Lednicki and Wieczorkowski, 2003)

$$n = n(a_1, a_2, ..., a_{L-1}) =$$
$$= N_L + \left(\sum_{h=1}^{L-1} W_h S_h\right)^2 \left(\overline{Y}^2 c^2 + \frac{1}{N}\sum_{h=1}^{L-1} W_h S_h^2\right)^{-1}, \tag{3}$$

where n is the minimizing sample size required to getting the given precision c of the mean value of variable Y when the strata boundaries are $\mathbf{a} = (a_1, a_2, ..., a_{L-1})^T$,

N_h is the size of the stratum h,

$W_h = N_h / N$ is the relative weight of the stratum h in the population U,

$S_h = \sqrt{N_h^{-1}\sum_{i=1}^{N_h}\left(Y_{ih} - \overline{Y}_h\right)^2}, \overline{Y}_h = N_h^{-1}\sum_{i=1}^{N_h} Y_{ih},$ is the known population standard deviation of the stratification variable Y in the stratum h,

$\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ is the known population mean of the stratification variable Y in the population, under constraints

$$N_h \geq 2, h = 1,\ldots,L,$$
$$2 \leq n_h \leq N_h, h = 1,\ldots,L-1, \tag{4}$$

where n_h is the sample size from the stratum *h*.

The objective function (3) assumes a Neyman optimal sample allocation between strata; the mentioned authors gave also a form of the function (3) for a power allocation; we shall not consider that case, for the most important from practical point of view is certainly optimal allocation. Although the sample sizes from the strata n_h are not directly seen in the objective function (4), they are included in the formula; therefore in each step of the algorithm we should evaluate $n_h = nW_hS_h\left(\sum_{h=1}^{L-1} W_hS_h\right)^{-1}, h = 1,\ldots,L-1$ to control the constraints (4).

An aim of such stratification is to minimize the estimation of stratification variable, not the survey variable; it makes the stratification biased, but when a value of a correlation coefficient between the auxiliary, (stratification), and survey variable is big, (especially almost 1), the stratification is efficient enough. When we know from some sources, e.g. previous surveys or preliminary studies, a relationship between these two variables, we can use it in stratification; for details see e.g. Lednicki and Wieczorkowski (2003).

The random search method algorithm solving a problem of the optimal stratification minimizing the function (3) under the constraints (4) is presented hereafter.

1. Sort the population by the values of the stratification variable.
2. Choose an initial point **a**, i.e. the vector of initial strata boundaries. Some random integers satisfying the conditions (4) could be used, but practice shows that the better results give the approximate strata boundaries obtained by applying some classical approximate methods, e.g. Dalenius and Hodges (1959), Eckman (1959) or Mahalanobis (1952). Calculate the function value $n = n(\mathbf{a})$.
3. For $r = 0, 1, \ldots, R$ repeat the following step:
   a. Generate point $\mathbf{a}'$ by drawing one stratum boundary a_i and changing it as follows

$$a_i' = a_i + j,$$
$$a_k' = a_k \text{ for } k = 1, \ldots, L-1, k \neq i, \tag{5}$$

where *j* is the random integer, $j \in \langle -p; -1\rangle \cup \langle 1; p\rangle$; *p* is a given integer according to the size of the population.

   b. Calculate the function value $n' = n(\mathbf{a}')$.
   c. If the conditions (4) are satisfied and $n(\mathbf{a}') \leq n(\mathbf{a})$, accept $\mathbf{a}_{r+1} = \mathbf{a}'$, (where $\mathbf{a}_{r+1}$ is the vector of strata boundaries in a next iteration), else $\mathbf{a}_{r+1} = \mathbf{a}$.

4. Finish the algorithm if the stopping rule is fulfilled, e.g. if $r = R$, where *R* is given number of steps or if in last *m* steps the sample size did not decrease. We take the vector **a** as a vector of final strata boundaries.

The *j* value in the third step has two tasks. First, it makes the algorithm works faster in comparison to a classical case, (see Niemiro, 1999), in which *j* = 1 in all steps. It is significant especially in a case of a large population. Secondly, probable more important in the *j* value, its random propriety protects us against the algorithm stopping in a local minimum what could happen in a case of *j* = 1. Therefore the above algorithm is more efficient than the one proposed by Niemiro (1999).

As it was mentioned, the integer *p*, being the lower and upper bound of the *j* value, should be determined according to the size of the population; it should not be too big; in our investigations it was fixed as *p* = 3. Preliminary studies showed that *p* rather should not be bigger than 5; in smaller populations it might be smaller, e.g. *p* = 2 or 3 for *N* being about a few hundreds. What is important, *p* should not be equal 1, (the reasons are given above).

In agricultural surveys conducted by Central Statistical Office of Poland we can often meet a case in which stratification variable values are equal in some units, (farms). For instance, let us consider the farms population. Many of farms have no cattle and some others have only little number of it. The similar situation can be considered in a case of other agricultural variables. Table 1 contains the number of farms in particular provinces and number of farm groups having different cereals and potatoes area. The originating from the Agricultural Census 2002 data regards the farms of larger than 2 ha area. As one can see, many farms have the same cereals and, mostly, potatoes area.

The above propriety can be helpful in making the random search algorithm faster. Let us create a new variable, say Y^u, which has the same values as *Y*, but consists only of unique units, (so we remove the duplicated ones). Calculate weights $g_i, i = 1,...,N^u$, where N^u is the Y^u size. These weights inform how many units in the population have the Y_i^u value.

Table 1. Number of farms with the agricultural land area larger than 2 ha, and number of farms with different cereals and potatoes area by province in Poland

Province Code	Number of farms with:		
	Area larger than 2 ha	different cereals area	different potatoes area
02	54401	3408	575
04	64013	3795	458
06	170496	3172	429
08	18869	2322	298
10	128429	2803	755
12	124659	1589	367
14	223677	3382	604
16	28368	2984	287
18	119769	1553	399
20	84052	2897	482
22	38899	3265	428
24	57437	2056	341
26	90688	1896	357
28	39786	3238	379
30	107931	4211	639
32	27837	3069	497
Total	1379311	45640	7295

Source: own calculations based on CSO data from the Agricultural Census 2002

The above algorithm can be now applied for the Y^u variable. The objective function (3) has to be modified as follows:

$$n = n(a_1^u, a_2^u, \ldots, a_{L-1}^u) =$$

$$= N_L + \left(\sum_{h=1}^{L-1} W_h^u S_h^u\right)^2 \left(\overline{Y}^2 c^2 + \frac{1}{N}\sum_{h=1}^{L-1} W_h^u \left(S_h^u\right)^2\right)^{-1}, \qquad (6)$$

where $a_1^u, a_2^u, \ldots, a_{L-1}^u$ are the strata boundaries for Y^u,

N is the population size, ($N = \sum_{i=1}^{N^u} g_i$),

$N_h = \sum_{i=a_{h-1}^u+1}^{a_h^u} g_i, h = 1, \ldots, L-1,$ is the weighted size of the stratum *h*, (*h*=1,...,*L*–1),

$N_L = \sum_{i=a_{L-1}^u+1}^{N^u} g_i$ is the weighted size of the last stratum L,

$W_h^u = W_h = N_h / N$,

$S_h^u = \sqrt{N_h^{-1} \sum_{i=a_{h-1}^u+1}^{a_h^u} g_i \left(Y_i^u - \overline{Y}_h^u\right)^2}, \overline{Y}_h^u = N_h^{-1} \sum_{i=a_{h-1}^u+1}^{a_h^u} g_i Y_i^u$, is the weighted population standard deviation of Y^u in the stratum h.

Such modification contributes to faster work of the algorithm because of working on the smaller sets, (see tab. 1). The same modification can be used when applying the simplex method.

3. Numerical example

In this section an application of the proposed method and its comparison with the simplex method is presented. The latter will be applied in two versions – first will be the original, with the objective function (3), and second will be the modified one, i.e. the simplex method applied for the weighted data with the objective function (6). Data from Agricultural Census 2002 regarding cereals area are used in the study. The frame consists of the farms with the agricultural land larger than 2 ha. The subpopulations, (provinces), were stratified using three investigated methods. The cereals area was a stratification variable. The approximate strata boundaries from the Dalenius and Hodges method (1959) were used as the initial strata boundaries. A sample size required to getting a precision of the stratification variable total in subpopulations equal c = 0.005 have been used as a comparative criterion.

Five experiments have been carried out; in each one a different number of strata have been created, i.e. $L = \{6, 8, 10, 12, 14\}$. Overall sample sizes, (i.e. the sample sizes from the whole Poland), required to get the fixed precision of estimation in subpopulations c obtained by using three methods are presented in table 2, (for each experiment).

A comparison of the results are based on the values of the objective function (3) or (6), (according to the method), obtained by using the three studied methods. First of all, the results of the simplex method for the modified data were much better than the results of the classical simplex used by Lednicki and Wieczorkowski (2003). The more strata were used the difference was bigger. It confirms the results of Findensein *et al.* (1974); in their opinion the simplex method is efficient in a case of 3 or 4 dimensions and markedly less efficient in a case of more dimensions.

Table 2. Sample size from whole Poland obtained be using three optimization algorithms, i.e. classical simplex (S1), simplex for weighted data (S2), random search (RS), for cereals area and different number of strata (precision of estimation in provinces $c = 0.005$)

Number of strata (L)	Sample size for optimization algorithm:		
	S1	S2	RS
6	36425	36078	36030
8	20706	20702	20457
10	13585	13264	13053
12	13032	9422	9054
14	10776	6916	6699

Source: own calculations based on CSO data from the Agricultural Census 2002.

The results of the stratification using the random search algorithm were also much better than results of simplex method for the original data. There were also differences between the results of the random search method and simplex method for the modified data, although they were rather small, (but still satisfying). Note that the results of both applications of the simplex method could not be improved on more, in contradiction to the results of the random search method; its results could be improved in some cases by longer working of the algorithm.

Table 3. Sample sizes from provinces obtained by using three optimization algorithms, i.e. classical simplex (S1), simplex for weighted data (S2), and random search (RS), for cereals area and $L = 12$ strata, (precision of estimation in provinces $c = 0.005$)

Province Code	Sample size for optimization algorithm:			Province Code	Sample size for optimization algorithm:		
	S1	S2	RS		S1	S2	RS
02	746	663	625	18	823	675	667
04	710	467	471	20	603	474	443
06	873	530	525	22	1085	588	523
08	633	647	571	24	1015	747	746
10	656	506	500	26	686	579	572
12	927	778	767	28	1172	597	539
14	788	560	532	30	591	487	488
16	603	570	547	32	1121	554	538

Source: own calculations based on CSO data from the Agricultural Census 2002.

Table 3 contains an example of the results of stratification, ($L = 12$), for particular provinces. Note that in some cases the simplex methods led to better results than the random search method. Certainly the results of the latter are random; therefore multiple repetitions of the algorithm would be a more efficient way of computing.

The R language was used to do all computations in the paper (see R Development Core Team, 2003).

4. Conclusions

The algorithm of the random search method for the optimal stratification in the agriculture surveys has been introduced in the paper. Its application for the real agricultural data was presented. The algorithm is easy in implementation, (for instance in R language, as in the paper), quite fast and, first of all, more efficient than other presented stratification approaches. The sample sizes required to obtaining the precisions of the mean estimators were smaller in comparison to the results of other two methods. Note that the simplex method used for the modified data were also quite efficient, so such data modification presented in the paper can be proposed in practical agricultural surveys.

Unfortunately, we cannot guarantee that the algorithm does lead to the global minimum of the studied objective function; but the results of the presented comparison with the simplex method showed, that it is more efficient in a sense of leading to the smaller value of the objective function. It makes us recommend the data modification and random search method as the better way of stratification using the approach proposed by Rivest (2002) and Lednicki and Wieczorkowski (2003) than the simplex method used by Lednicki and Wieczorkowski (2003).

REFERENCES

BRACHA Cz. 1996. *Teoretyczne podstawy metody reprezentacyjnej* (*Theoretical Basis of Survey Sampling*). PWN, Warsaw, Poland.

BRANDT S. 1999. *Data Analysis. Statistical and Computational Methods for Scientists and Engineers*. Ed.3. Springer Verlag, New York.

COCHRAN W. G. 1977. *Sampling Techniques*. John Wiley & Sons, New York.

DALENIUS T., HODGES J.L. 1959. Minimum Variance Stratification. *Journal of the American Statistical Association*, 54, 88—101.

DORFMAN A. H., VALIANT R. 2000. Stratification by Size Revisited. Journal *of Official Statistics*, 16, 139—154.

ECKMAN G. 1959. An Approximation Useful in Univariate Stratification. *Annals of Mathematical Statistics*, 30, 219—229.

FINDENSEIN W., SZYMANOWSKI J., WIERZBICKI A. 1974. *Metody obliczeniowe optymalizacji* (*Computing Methods of Optimization*). Wydawnictwa Politechniki Warszawskiej, Warsaw, Poland.

GODFREY J., ROSHWALB A., WRIGHT R. L. 1984. Model-Based Stratification in Inventory Cost Estimation. *Journal of Business and Economic Statistics*, 2, 1—9.

HEDLIN D. 2000. A Procedure for Stratification by an Extended Eckman Rule. *Journal of Official Statistics*, 16 (1), 15—29.

LAVALLEÉ P., HIDIROGLOU M. 1988. On the Stratification of Skewed Population. *Survey Methodology*, 14, 3—43.

LEDNICKI B., WIECZORKOWSKI R. 2003. Optimal Stratification and Sample Allocation between Subpopulations and Strata. *Statistics in Transition*, 6, 287—306.

MAHALANOBIS P. C. 1952. Some Aspects of the Design of Sample Survey. *Sankhya*, 1—7.

NELDER J.A., MEAD R. 1965. A Simplex Method for Function Minimization, *Computer Journal*, 7, 308—313.

NIEMIRO W. 1999. Konstrukcja optymalnej stratyfikacja metodą poszukiwań losowych. (Optimal Stratification using Random Search Method). *Wiadomości Statystyczne*, 10, 1—9.

R DEVELOPMENT CORE TEAM 2003. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org.

RIVEST L.P. 2002. A Generalization of Lavallee and Hidiroglou Algorithm for Stratification in Business Survey. *Techniques d'enquete*, 28, 207—214 (www.mat.ulaval.ca/pages/lpr/).

SCHNEEBERGER H. 1970. Optimierung in der Stichprobentheorie durch Schichtung. *Statistische Hefte*, 11.4, 242—253.

SWEET E.M., SIGMAN R. 1995. Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations using Auxiliary Data, U. Bureau of the Census (www.censugov/srd/papers/pdf/sm95—22.pdf).

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 807—816

UTILIZATION OF ADMINISTRATIVE REGISTERS IN THE POLISH OFFICIAL STATISTICS

Ewa Walburg[1], **Agnieszka Prochot**[2]

ABSTRACT

Development of Polish Official Statistics Information System, including the field of changes in data sources for statistical surveys — wider utilisation of administrative registers, is a continuous/permanent process.

The purpose of undertaken activities is a modernisation of sources providing statistical surveys with data, consisting in an increase of administrative data share in total number of data provided for statistics, i.e. reducing use of a traditional way of data collecting based on statistical forms and extended utilisation of administrative registers instead.

A subject of work being conducted at CSO is wider utilisation of administrative registers priority for official statistics i.e.: tax system, system of social security, systems of employment agencies and social welfare, system of health insurance, geodesy records, real estate tax register, Integrated Administration and Control System IACS

The paper presents previous stages of Polish Official Statistics Information System development in the field of administrative registers, executed and planned work. The paper also describes the outline of perspectives for wider provide for statistics with administrative data and related background.

1. Present status

Since 2000 intensive work on wider utilisation of administrative data has been carried out in Central Statistical Office. The purpose of these activities is a modernisation of sources providing statistical surveys with data, consisting in an increase of administrative data share in total number of data provided for statistics, i.e. reducing use of a traditional way of data collecting based on statistical forms and extended utilisation of administrative registers instead (Dmochowska, 2001; Kordos, 1995).

[1] Central Statistical Office, Warsaw, Poland, e:mail: e.walburg@stat.gov.pl
[2] Central Statistical Office, Warsaw, Poland, e:mail: a.prochot@stat.gov.pl

In respect of the registers and information systems of public administration one can distinguish two stages of Information System of Statistics development. In the first one, a legal basis for providing statistics with administrative data was created — law on official statistics[1]. In the second stage, activities were focused on gathering of information on administrative registers, evaluation of their usefulness for statistics as well as building of metainformation system — a tool for coherence analysis of Official Statistics System and administrative systems. Administrative registers were assigned to one of three groups according to criteria for verifying their usefulness for statistics. Systems, which are presently relevant and foreseen in perspective for use in many official statistics surveys, were included in the first group. It comprises among others: tax system, system of social security, system of health insurance, systems of employment agencies and social welfare, geodesy records, real estate tax register, Integrated Administration and Control System (IACS) (Kordos, Paradysz, 2000).

The second group covers "specific" systems, being a data source for a single survey, e.g. vehicles and drivers' records, systems concerning environment protection. The third group embraces systems which are not potential data sources for statistics in view of the fact that their information content is not useful for the official statistics at present or they are under construction, but the date of their implementation is unknown.

142 systems were identified in the course of work on recognition of administrative data sources. They are mostly nation-wide, computerised systems. Centrally maintained administrative sources make up over a half of such data sources. The following identification standards are used in systems: the National Official Register of the National Economy Units number — REGON, General Electronic System of Population Registration number — PESEL, Tax Identification Number — NIP and National Official Register of Territorial Division of the Country number — TERYT.

In 2003, 80 of 197 investigations, covered by Programme of Statistical Surveys of Official Statistics, were supplied with information from the administrative data sources. Administrative registers of the second group were utilised most often, including systems concerning environment protection, whereas administrative registers included in the first group — in a lesser degree.

[1] Law on official statistics issued on 29 June 1995 (Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej.) (Polish — English version), Dz.U. Nr 88 poz.439.

2. Development of work

2.1. Legal basis for administrative data transfer

Work, carried out at CSO, is aimed at extended use of administrative registers of the first group in official statistics. With the purpose of forming a legal basis for statistics reinforcement with new administrative data sources, i.e. systems of employment agencies and social welfare, system of health insurance, tax system, system of social security, appropriate records were put in the following surveys included in Programmes of Statistical Surveys of Official Statistics in 2002-2004: Households Budgets Survey, Health of Population - Monitoring of Health Survey, Medical Stuff Survey, Pharmacies Survey, Enterprises Current Financial Results and Enterprises Fixed Assets Expenditures Survey, Current Business Survey, Gross Domestic Product and its Elements by Geographical Regions Survey.

Besides, relevant information was put in the project of Programme of Statistical Surveys of Official Statistics in 2005 to enable utilisation of real estate tax register, system of population registration and IACS in the following surveys: Forest Resources Survey, Management of Housing Stock Survey, Land Use Survey.

2.2. Tax system

Since 2002, Ministry of Finance as a tax system administrator has provided CSO with datasets with predetermined scope of data concerning:

- direct taxes from the personal income tax database (PIT) and the corporate income tax database (CIT), which describe results of activity of economic units, legal persons, organisational units without legal status and the taxpayers obtaining incomes from work, pension and other similar payments.
- taxpayers from the National Register of Taxpayers.

In 2004, Ministry of Finance will also provide CSO with data concerning Value Added Tax (VAT).

The datasets from the National Register of Taxpayers have been used for comparative analysis of units from the Statistical Business Register (BJS). This register has been a sampling frame for statistical surveys and a basis for creating lists for surveys since 2001. From the point of view of business surveys, providing the Register with up-to-date information is very important. For this reason, use of administrative data for updating sampling frame in the scope of units' activity in order to exclude non-active units from surveys, is of great importance to statistics.

In the framework of the performed work, data from the National Register of Taxpayers on natural persons carrying out economic activity, legal persons and organisational units without legal status were compared with Statistical Business Register data with the help of identification numbers of General Electronic

System of Population Registration — PESEL and National Official Register of the National Economy Units — REGON. BJS was provided with the Tax Identification Numbers from the National Register of Taxpayers, in which missing Statistical Identification Numbers of the National Official Register of the National Economy Units were supplied. The Ministry of Finance was provided with a verified dataset of the National Register of Taxpayers in order to make a comprehensive analysis of system quality. Providing information scope of Statistical Business Register (BJS) with Tax Identification Numbers (NIP) has made possible BJS updating with data on direct taxes from the personal income tax database and the corporate income tax database. BJS was also supplied with data concerning information on revenues. In order to make comparison of data on revenues between statistical and tax systems, three datasets comprising data from 2001 were used: the results of survey "Report on economic activity concerning enterprises employing up to 9 persons", data from the personal income tax database and the corporate income tax database. A correlation of revenues from this survey and the revenues from the Ministry of Finance databases was investigated and the analysis of their coherence was carried out. This investigation enabled relative error estimation. For 79% of analysed units, this error was less than 3%. At present, information on revenues from the tax system is used as an auxiliary variable in sample design for survey "Report on economic activity concerning enterprises employing up to 9 persons".

Within work on the use of data from the tax system for statistical purposes, methodologies accepted in tax system and in statistical system were also compared. Information scope of administrative systems (subjects and characteristics scope) was described; a range of coherence and a possibility of systems integration were defined. In this way, feasibility of reduction of present structural survey information scope including surveys describing micro-enterprises, were outlined. The next area of work on the use of data from tax system is connected with an elaboration and presentation of enterprise activity survey results. It comprises imputation of missing data and generalization of survey results, with the use indirect information as auxiliary variables.

2.3. System of Social Security

For statistical surveys purposes information resources of social security system contained in Complex Computing System of Social Security Service (KSI ZUS) are also very important. For the sake of long duration of KSI ZUS implementation and due to the problems connected with quality of datasets, including low degree of completeness, this system is not a source of data for official statistics at present. In 2005, CSO will be provided with dataset describing contribution payers of social security with the number of person covered by the social security service. These data will be used for BJS updating - for determination of economic activity, verification of units address characteristics and for determination of employed persons number. This last variable is used in official statistics as a criterion of units grouping. However, it requires previous determination of information scope (subject and characteristics scope) and methodology applied at Social Security Service (ZUS). A comparison of concept definitions between social security system and official statistics should be also performed. Only in case of obtaining coherence of these two systems, the comparison between number of persons employed by subjects — contribution payers with appropriate information from statistical register will be possible. At present, KSI ZUS is the only source of information on number of employed which can be used for verification of these data in BJS.

Work on utilisation of Agricultural Social Insurance Fund (KRUS) system in statistical surveys has been carried out since 2000. Work on building of KRUS data warehouse is still on and information resources of that warehouse are not used in official statistics at present. Providing CSO with such data will be possible after full implementation of computerised system, covering information on insured persons and beneficiaries.

2.4. System of Health Insurance

Work concerns also System of Health Insurance - Central Register of Insured Persons (CWU), which is administrated by National Health Fund. CWU is being built on the grounds of local databases. After elimination of duplicated or multiplicated records a uniform database on insured persons has been set and currently provided with data from Social Security Service, Agricultural Social Insurance Fund and Ministry of Interior Administration. Health care services (medical benefits) rendered to insured persons are not registered in CWU. That function is still fulfilled by local computerised systems of NFZ branches integrated with CWU through sending of updating information. Appropriate record was put in Programme of Statistical Surveys of Official Statistics in 2005 concerning Health of Population - Monitoring of Health survey in order to obtain datasets from CWU by CSO.

2.5. System of Social Assistance

Since 2002 Ministry of Economy, Labour and Social Policy (MGPiPS) has provided CSO with data from System of Social Assistance SI POMOST in a limited scope for the sake of incomplete implementation of the system. At present only approximately 60% of municipalities — gminas (social assistance centres) transfer data to MGPiPS. Amended provisions of the act on social welfare will accelerate full implementation of the system in 2005. That system is not a source of data for official statistics at present. Due to lack of possibility of obtaining full data for all voivodships, statisticians have undertaken activities for utilisation in statistics of data from SI POMOST for voivodships, where social assistance centres implemented the system. Datasets from these centres are transferred to MGPiPS quarterly and from there they are next transferred to CSO with the same frequency. A sample survey based on datasets from SI POMOST has been designed at CSO.

Datasets from SI POMOST, transferred to CSO, have been analysed from viewpoint of their coherence with methodology accepted in statistical surveys. The analysis result concerning three voivodships was described in a report [9]. Conclusions resulting from the work on datasets from SI POMOST, regarding among others incoherence with the system of official statistics, are reported to the administrator of the system — MGPiPS currently and used in developing next versions of the system.

Statisticians are also interested in system concerning registered unemployment SI PULS. At present reports prepared for CSO are based on information resources of that system. In 2003 MGPiPS has begun work on building a new system, which will accumulate data on social assistance (SI POMOST) and unemployment (SI PULS).

2.6. Other administrative registers

Work on Real Estate Tax Register is also conducted at CSO within activities regarding utilisation of administrative registers in official statistics. That system is run in municipality offices for the purpose of taxation and levies of estate duty, agricultural tax and forest tax. That register is to be a basis for building a full information system on real estates, a system ensuring collecting, updating and exchange of information on real estates, including legal information (legal cadaster), registration information (resulting from lands and buildings register) as well as information for tax purposes (fiscal cadaster). It contains data on taxpayers and subjects of taxation.

A subject of current work is the Integrated Administration and Control System (IACS), which embraces records of producers, agricultural holding and applications for subsidies (area aid applications).

In order to obtain datasets and start an analytical work on the use of the real estate tax register and the IACS for the statistical surveys and for Statistical Register of Agricultural and Forest Holdings purposes, relevant records to be put in annual Programme of Statistical Surveys of Official Statistics 2005, were defined at CSO for the following surveys: Forest Resources Survey, Management of Housing Stock Survey, Land Use Survey.

3. System of Statistical Metainformation on Administrative Registers

For the use of administrative registers in official statistics, information on outer data sources is necessary. A necessity of building of knowledge compendium on administrative data sources for official statistics purposes comes from this fact. In 2002, work on building of Metainformation System on Administrative Data Sources (SMA) has been started. This System describes, in a complex way, administrative registers and comprises: Database of descriptions of administrative data sources and Dictionary of metainformation system concepts and Dictionary of classifications, nomenclatures and groupings that are used in administrative data sources (Olenski, 2001).

Within work on building of Metainformation System on Administrative Data Sources, building of exchange information system with Official Statistics System is planned. It will enable comparative analyses of the Official Statistics System and systems of public administration, generating specifications, making assessments and drawing conclusions concerning the use of the administrative registers for statistical purposes. The analysis of information gathered in Metainformation System on Administrative Data Sources will be a basis for an assessment of: administrative data usefulness for Official Statistics System and for current utilisation of administrative sources for statistical survey purposes.

4. Planned work

4.1. New perspectives

Further work on extension of the use of administrative registers in the official statistics will particularly concern the systems from the first group and will be proceeded towards:

- gradual change from supplementary use to extensive use of administrative data,
- building of coherent statistical system based on the traditional data sources and on the registers (Szarek, 2002),
- increasing influence of statistical services both on planned and implemented administrative systems.

Expected results of planned activities are: increase in a range of administrative data provided for official statistics, especially from Tax System and Social Security System, obtaining of integrated data from Employment Agencies and Social Welfare System and implementation of administrative sources into the statistical practice in a wider scope than at present.

An advisable direction of changes is the use of administrative systems as:

- a direct data source for the present surveys based on statistical forms (Zagozdzinska, 2001),
- a direct data source for new surveys,
- a source of auxiliary variables used in the small area indirect estimation technique (Kordos, Kubacki, 2000),
- an information source for imputation of missing data from surveys,
- a tool for quality control of data from surveys,
- a data source for Statistical Business Register and Statistical Register of Agricultural and Forest Holdings updating (Witkowski, 2003).

4.2. Phare 2003 Project

Within Phare 2003 Programme, Twinning Covenant on Upgrading of the quality of Polish statistics was concluded in March 2004 between the President of CSO and the General Director of Statistics Sweden. The Covenant covers among other a question of administrative data use extension for statistical purposes. Planned work within that Project is aimed at development of methods enabling extended use of administrative information systems for statistical purposes. The Project will cover especially: tax system, system of population registration, system of social security. A description of the conditions and procedures indispensable for use of the systems covered by the Project will be accomplished and general conclusions on utilisation of other administrative systems in official statistics will be drawn.

Within the Project a feasibility study on the use of administrative data systems essential for statistical purposes will be accomplished, too. An analysis of coherence of the official statistics system and selected administrative systems will be carried out, problems that have to be dealt with will be identified as well as countermeasures against barriers to providing official statistics with administrative data will be defined. A strategy for the improvement of the use of administrative registers in statistical surveys will be elaborated.

5. Background of a wider utilisation of administrative registers in official statistics

Along with increasing possibilities of providing statistics with new administrative sources, problems concerning their utilisation have been growing. A solution of methodological and organisational problems calls for many conditions to be met and much work to be executed. Accomplishing the following tasks is indispensable:

1. Evaluation of administrative data sources having view of: deadlines and frequency of data transfer to statistical units, stability of data source, comparison of administrative source utilisation cost to classic survey cost.
2. Evaluation of administrative data usefulness for official statistics, including assessment of coherence between key variables (definitions, classifications). Definition of a way and a scope of administrative data use. Establishing of data combining level (microdata, aggregate data at regional level, national level). Establishing of criteria for evaluation of data quality (Dmochowska, 2001; Platek, Särndal, 2001). Making a proposal for changes of statistical surveys methodology.
3. Starting of experimental work on administrative datasets. Elaboration of innovative methods of small area indirect estimation using data from statistical surveys and administrative registers. Description of manners for combining data from surveys with administrative data. Evaluation of new

methods usefulness in practice. Use of administrative registers in a new form depends on a wider co-operation of statisticians and different specialisation's researchers. Such a work organization will be also helpful in creating of knowledge of practical problem solutions in statistics (Witkowski, 2003).

4. Elaboration of methods for automatic coherence checks of data from different sources (administrative and statistical) and making data coherent in the possible scope.
5. Elaboration of rules of administrative datasets correctness verification and administrative data editing.

A problem of work on administrative data incorporation into Official Statistics System becomes a matter of growing importance, because of constant increase in their use.

REFERENCES

DMOCHOWSKA, H. (2001), Quality in statistics (in Polish), *Wiadomości Statystyczne,* nr 12, pp. 8—14.

KORDOS, J. (1995), Data quality issues during transition period of the Polish Official Statistics (in Polish), *Wiadomości Statystyczne,* nr 8, pp. 6—14.

KORDOS, J., KUBACKI, J. (2000), Regional statistics – methods and sources, Part II: Small area statistics — Poznań — Kiekrz 1999 (in Polish), Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

KORDOS, J., PARADYSZ, J. (2000), Some experiments on utilisation of small area estimation methods in Poland (in Polish), *Wiadomości Statystyczne* nr 11, pp. 1—22.

OLEŃSKI, J. (2001), Economics of Information (Ekonomika informacji), (in Polish), Polskie Wydawnictwo Ekonomiczne, Warszawa.

PLATEK, R., SÄRNDAL, C.E. (2001), Can a statistician deliver? Journal of Official Statistics, Vol.17, No.1, 2001 (in Polish), *Wiadomości Statystyczne*, nr 4, pp. 1—21.

SZAREK, J. (2002), Administrative sources for census data in Finland (in Polish), *Wiadomości Statystyczne* nr 8, pp. 72—83.

SZAREK, J., Statistical data for Podkarpackie voivodship in 2001. Households of beneficiaries in Opolskie, Podkarpackie i Świętokrzyskie voivodships in 2001, based on dataset of Social Assistance System - SI POMOST. Social Assistance System SI POMOST — data analysis for Podkarpackie voivodship in 2001. (in Polish), *http://www.politykaspoleczna.gov.pl/index.php?strona=statystyka2*

WITKOWSKI, J. (2003), Development of social statistics in Poland — important challenge (in Polish), *Wiadomości Statystyczne* nr 7/8, pp. 25—35.

ZAGOŹDZIŃSKA, I. (2001), Results Quality in Polish Business Statistics — reflections on paper by PLATEK, R., SÄRNDAL, C.E., “Can a statistician deliver?” (in Polish), *Wiadomości Statystyczne*, nr 7, pp. 1—7.

STATISTICS IN TRANSITION, April 2004
Vol. 6, No. 5, pp. 817—818

Book Review

Some Contributions to Multivariate Methods in Survey Sampling,

by Janusz Wywiał, University of Economics,
Katowice, 2003, ISBN 83-7246-273-9

This book provides valuable results and interpretations of multivariate statistical analysis to the problem of estimation of an unknown vector of population parameters. In practical applications of survey methods, a statistical inference on a vector of unknown parameters is more common than estimation or hypothesis testing confined to a single population parameter. Therefore, a large number of well justified formulae in this book will be regarded as interesting and useful by both researchers and statisticians responsible for survey designs. Although, the majority of problems discussed in this book consider the problem of estimation of a vector of mean values of variables in finite populations, applying various sampling strategies, most results can be generalized to vectors of other population parameters, such as total value or population fraction.

The book consists of an introduction, six chapters, index of expressions, list of references and summary in Polish.

Chapter 1 presents foundations of sampling strategies, including basic notions and characteristics of particular strategies, as well as interpretations of selected measures of accuracy of vector estimates.

Chapter 2 covers discussion of a number of issues related to applications of Horvitz-Thompson estimator. In particular, approximate expressions of the variance of this estimator for the mean value in different sampling schemes are derived and the role of parameters of auxiliary variables is considered.

Chapters 3 through 6 elaborate on specific topics related to properties and characteristics of vector estimates and sampling strategies, including problems of optimization of sample sizes, and ways of increasing precision owing to proper stratification or clustering in the population. Original techniques of stratification of population on the basis of auxiliary variables, two-phase sampling for stratification, and stratification of population after sample selection are discussed and estimation problems, including a new class of estimators are analyzed. Estimation accuracy, clustering algorithms and a number of connected topics are developed in relation to other two sampling techniques: cluster sampling and two-stage sampling.

Valuable and new results on properties of vectors of regression estimators under different sampling strategies are described by the author in Chapter 6.

In summary, this is a fair book, varied in content, in which Professor Wywiał presents his own results and interpretations of essential survey sampling problems. The language of this book is English.

Mirosław Szreder, University of Gdańsk, Poland