AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 3&4, 349–360

# Estimating Unemployment-Rates for Small Areas – A Simulation-Based Approach

Bernhard Meindl
Statistics Austria

**Abstract:** The estimation of Austrian unemployment rates is based on data of the labour force survey (LFS). It is possible to calculate direct, design based estimates with fixed precision for population subgroups for which the sample size is known due to the sampling design.

Sometimes we are interested to estimate unemployment rates for population subgroups in which the sample size is random and often small. In this paper we conducted a simulation study to compare the performance of several model-assisted estimation methods with the direct estimator using auxiliary information from an administrative source when estimating unemployment rates for population subgroups. The results showed that if sample-sizes in a subgroup are small, model-based estimators outperform the design-based estimator.

**Zusammenfassung:** Die Grundlage für die Schätzung von Arbeitslosenquoten bildet in Österreich die in den Mikrozensus eingebettete Arbeitskräfteerhebung. Bedingt durch das Stichprobendesign können für Teilgruppen der Population für die die Stichprobengröße aufgrund des Stichprobendesign bekannt ist, Parameter mit adäquater Genauigkeit geschätzt werden.

Oft ist es jedoch von Interesse, Parameter für Subgruppen zu schätzen, in denen die Stichprobengröße zufällig und meist gering ist. In dieser Arbeit wurde im Rahmen einer Simulationsstudie die Performance von modellbasierenden Schätzverfahren mit jener des direkten Schätzers für die Schätzung von Arbeitslosenquoten für Subgruppen mit kleinen Stichprobenumfängen verglichen. Für die modellbasierenden Schätzverfahren wurde Hilfsinformation aus einer administrativen Quelle herangezogen. Es zeigte sich, dass für kleine Teilgruppen modellbasierende Verfahren hinsichtlich standardisierter Performancekriterien design-basierten, direkten Schätzern überlegen sind.

**Keywords:** Small Area Estimation, Unemployment Rates.

## 1 Introduction

The aim of the EURAREA research project (The EURAREA Consortium, 2004) was to study the practical behaviour of standard small area estimators on 'real-world' data sets. Rao (2003) defines small areas as population subgroups that are geographically partitioned and for which the available sample size is too small to use direct estimation methods for estimating population parameters of interest with adequate precision. Likewise, small domains are population subgroups with small sample size that are partitioned according to some socio-demographic variables. For example, small areas could be municipalities or political districts while the partitioning of a data-set into age/sex groups would lead to small domains.

Small area estimates should only be produced if there is a justified user demand and no other data that serves the same purpose is available (Australian Bureau of Statistics, 2006, p. 9). If these requirements are met, it is necessary to examine different estimation methods concerning their suitability in practice before deciding to publish results on small area/domain level.

The aim of this work was to assess the performance of standard small area estimation methods when estimating Austrian unemployment rates at small area level by definition of the International Labour Organisation (ILO) (Hussmanns et al., 1990). A simulation study has been conducted to compare the performance of model-assisted or model-based estimation methods that make use of auxiliary information and the direct, design-based estimator. The auxiliary information used in the simulation study was available from an administrative source.

The paper is organised in the following way. After introducing the necessary notations, the estimation methods considered and the evaluation criteria that have been used to compare the different estimation methods, we describe the data that was used in this work along with the set-up of the simulation study that has been conducted. Finally, we present the main simulation results for two different applications along with conclusions.

# 2 Estimating Unemployment

## 2.1 Introduction and Notation

Estimating population parameters on small area level with classical direct, design-based estimators is difficult because sample sizes in small areas of interest are often small or even zero. This situation is evident for non-sampled domains for which the actual sample size equals zero. In this special case it is not possible to estimate the parameter of interest or its variance using a design-based approach. Analyzing data from the Austrian labour force survey (LFS) for the 3rd quarter of 2006 we see that the number of unemployed people at NUTS-3 level (Nomenclature des Unités Territoriales Statistiques) is indeed small with the distribution ranging from 0 to 31 with the median located at 5.

*'Borrowing strength'* is the key approach to overcome the limitations of direct estimators when sample sizes in certain domains are small or zero. It means that the effective sample-size available for estimation is increased by using data from either other small areas (area-indirect approach), other time-spans (time-indirect approach) or a combination of both. Furthermore, to improve parameter estimation, model-based methods can make use of auxiliary information which should be highly correlated with the target variable. This additional information has to be available either at unit- or aggregated at area-level. In this work, however, data from previous periods was not used.

Methods that make use of sample elements from different small domains and/or different time-points or which are based on explicit statistical models that borrow strength across time or space for parameter estimation are called '*indirect methods*'. Since the effective sample size that can be used for parameter estimation is increased, a reduction of the variance for indirect methods compared to the variance of the design based, direct estimator is expected. One has to keep in mind, however, that - due to the underlying model - indirect methods are usually not unbiased regarding the population parameter.

Considering these facts, it is appropriate to study the statistical properties and the performance of different estimation methods in a (simulation) set-up that is as close as possible to a real life problem. The results may help in deciding if results obtained by indirect and model-based estimation methods on small area level respectively are worth to be officially published. This case is preferable because decision makers often have legitimate interest in small area statistics. Information of small area level may be used for example to allocate funds or to improve regional planning (Rao, 2003, p. 3). Information on cancer rates at small area level is for instance crucial to identify critical regions and let decision makers take appropriate actions.

Before discussing different estimators, some notations need to be introduced. We consider $y$ the variable of interest, $x$ denotes an auxiliary variable and $w$ represents the sampling weight. $\overline{y}$ and $\hat{\overline{y}}$ denote the mean and an estimator for the mean of the target variable, respectively. The subscript $i$ is used in this paper as index for a person, the subscript $d$ is used to index small areas. Thus, $y_{id}$ represents the value of the target variable of the $i$th person within small area $d$. Furthermore, $s$ represents the available sample and $M$ denotes the number of samples in the simulation study.

## 2.2 Estimation Methods

In the following section we give a short overview of the estimation methods that have been used in the simulation study. Detailed information as well as a discussion on the advantages and disadvantages of these estimation methods is given in the EURAREA Project Reference Volume (The EURAREA Consortium). It has been an interesting question how a Bayesian method is performing in comparison to the methods already considered by the EURAREA project team. Thus, a hierarchical Bayes estimator was considered in the simulation approach as well. The hierarchical Bayes estimator was implemented using *R* (R Development Core Team, 2007) and OpenBugs (Thomas et al., 2006).

**National sample mean** The *national sample mean* is defined as

$$\hat{\overline{Y}}_d^{NSM} = \frac{1}{\hat{N}} \sum_{i \in s} w_i y_i \qquad \text{with} \quad \hat{N} = \sum_{i \in s} w_i \,. \tag{1}$$

The weighted expansion estimator $\hat{\overline{Y}}_d^{NSM}$ is used as a benchmark estimator in this simulation study. It should be noted that the national sample mean is an indirect estimation method because all sample elements are used to calculate the estimator for small area $d$.

**Direct estimator** The *direct estimator* which is also known as the classical Horvitz-Thompson estimator only makes use of area-specific sample elements and is defined as

$$\hat{\overline{Y}}_d^{DIR} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} w_{id} y_{id} \qquad \text{with} \quad \hat{N}_d = \sum_{i \in s_d} w_{id} \,. \tag{2}$$

**Generalized regression estimator** The approximately design-unbiased *generalized regression estimator* is defined as

$$\hat{\overline{Y}}_d^{GREG} = \hat{\overline{Y}}_d^{DIR} + \left(\overline{\boldsymbol{X}}_d - \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}\boldsymbol{x}_{id}\right)^T \hat{\boldsymbol{\beta}} \qquad \text{with} \quad \hat{N}_d = \sum_{i\in s_d} w_{id}\,. \tag{3}$$

Thus, $\hat{\overline{Y}}_d^{GREG}$ 'adjusts' the direct estimator $\hat{\overline{Y}}_d^{DIR}$ for differences between the sample mean and the population mean of a vector of auxiliary variables using a regression model. Thus, $\hat{\overline{Y}}_d^{GREG}$ can be considered as a model-assisted estimation method.

**Synthetic estimator A** For the first synthetic estimator auxiliary information needs to be available on unit/person level. It is then possible to link the sample information $y_{id}$ from the domain of interest to the auxiliary information $x_{id}$ on unit level using a regression model. A weighted OLS estimator $\hat{\boldsymbol{\beta}}^{unit}$ is then calculated from this unit-level regression model. The synthetic estimator $\hat{\overline{Y}}_d^{SA}$ is finally calculated as

$$\hat{\overline{Y}}_d^{SA} = \overline{\boldsymbol{X}}_d \hat{\boldsymbol{\beta}}^{unit}\,. \tag{4}$$

**Synthetic estimator B** In contrast to the synthetic estimator $\hat{\overline{Y}}_d^{SA}$, the second synthetic method $\hat{\overline{Y}}_d^{SB}$ is based on an area-level regression model which links sample information to auxiliary information for the corresponding area on small area level. A weighted OLS estimator $\hat{\boldsymbol{\beta}}^{area}$ has to be computed. The synthetic estimator B is then given as

$$\hat{\overline{Y}}_d^{SB} = \overline{\boldsymbol{X}}_d \hat{\boldsymbol{\beta}}^{area}\,. \tag{5}$$

**Empirical best linear unbiased predictor A** The empirical best linear unbiased predictors are also referred to as EBLUP's (Rao, 2003, p. 95ff). EBLUP A is given as

$$\hat{\overline{Y}}_d^{EBA} = \gamma_d \hat{\overline{Y}}_d^{GREG} + (1-\gamma_d)\hat{\overline{Y}}_d^{SA}\,. \tag{6}$$

It should be noted that the concept of unbiasedness applies with respect to the underlying model. Using the area-specific variance ratio $\gamma_d$ (The EURAREA Consortium, PRV Part I, p. 20ff) as weights, $\hat{\overline{Y}}_d^{EBA}$ is given as a linear combination of $\hat{\overline{Y}}_d^{GREG}$ and the synthetic estimator $\hat{\overline{Y}}_d^{SA}$.

**Empirical best linear unbiased predictor B** $\hat{\overline{Y}}_d^{EBB}$ is given by expression

$$\hat{\overline{Y}}_d^{EBB} = \gamma_d \hat{\overline{Y}}_d^{DIR} + (1-\gamma_d)\hat{\overline{Y}}_d^{SB}\,. \tag{7}$$

$\hat{\overline{Y}}_d^{EBB}$ is therefore a linear combination of the direct estimator $\hat{\overline{Y}}_d^{DIR}$ and the synthetic estimator $\hat{\overline{Y}}_d^{SB}$ using the area-specific variance ratio $\gamma_d$.

**Hierarchical Bayes estimator** The hierarchical Bayes estimator $\hat{\overline{Y}}_d^{HB}$ is a model-based estimator. Since it is built in stages this estimator is considered hierarchical (Suciu et al., 2001, p. 18). In a first step the underlying model (8) has to be defined. We may assume that the number of unemployed people in small area $d$ is normally distributed with an area-specific mean $\mu_d$ and variance $\tau$. $\mu_d$ is linked to the auxiliary information available using a regression model.

$$y_d \sim N(\mu_d, \tau)\,, \qquad \mu_d = \alpha_d + \overline{\boldsymbol{X}}_d^T \boldsymbol{\beta}\,. \tag{8}$$

In a second step the prior information on the model parameter has to be specified. In case of no prior information, non informative prior distributions are typically assumed for all model parameters. We may define the non informative priors for model (8) as

$$\alpha_d \sim N(0, \tau_\alpha)\,, \qquad \beta_i \sim N(0, 0.0001)\,, \qquad \tau, \tau_\alpha \sim \Gamma(0.001, 0.001)\,. \tag{9}$$

Using Markov-Chain Monte-Carlo (MCMC) methods (Rao, 2003, p. 224ff) it is possible to sample from the posterior distribution of the parameter we are interested in given the data. An estimate for the mean of the parameter of interest is finally derived by averaging over draws from the posterior distribution.

## 2.3 Performance Criteria

In order to assess the performance of the estimation methods discussed above concerning their statistical properties such as variance or bias, it is necessary to use well-known performance measures. In this work we have concentrated on the following criteria:

**RRMSE (relative root mean square error) in %**

$$RRMSE_d(\%) = 100\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\overline{Y}}_d^m - \overline{Y}_d}{\overline{Y}_d}\right)^2} \tag{10}$$

**RB (relative bias) in %**

$$RB_d(\%) = 100\left(\frac{1}{M}\sum_{i=1}^{M}\frac{\hat{\overline{Y}}_d^m - \overline{Y}_d}{\overline{Y}_d}\right) \tag{11}$$

**ARRMSE (average relative root mean square error) in %**

$$ARRMSE(\%) = \frac{1}{D}\sum_{d=1}^{D} RRMSE_d(\%) \tag{12}$$

Detailed information on the performance criteria (10)-(12) is available in the documentation of the EURAREA project (The EURAREA Consortium, PRV Part I, p. 65f). It is important to note that the $RRMSE_d$ and the $RB_d$ are area specific performance criteria while the $ARRMSE$ measures the performance of an estimator globally.

# 3 Simulation Study

## 3.1 Data

This work is based on data of the Austrian LFS from which information on a person's labour force status is available. The LFS in Austria is conducted as a quarterly survey. The sampling plan of the LFS is described in detail by Haslinger and Kytir (2006). Information on a person's labour force status is available from the Austrian Association of Social Insurance Providers (ASIP) as well. An advantage of the administrative data from ASIP is that they are complete.

Unfortunately, no common identifier exists that would allow to merge the two data-sources easily. However, for LFS data from the 3rd quarter of 2006 which will be used in this work, matching techniques have been used to link information from both data sources. For more than 92% of the respondents of the LFS a definitive link to the ASIP data could be established. Thus, for approximately 45.000 persons the information on their labour force status was available from both the LFS as well as from the ASIP data. Additionally, further demographic variables such as age, sex or information on the educational status of these persons was available from the LFS survey.

One should note that the employment status is measured differently in the LFS survey and the ASIP data. While the concept of measuring the employment status in the LFS survey is based on the definitions of the ILO, the employment status available from data from the ASIP is based on national concepts. Even though the underlying concepts of measuring the employment status are different, the variables containing information on the labour force status from the two data sources are positively correlated. Thus, the additional information from the ASIP may be exploited to improve parameter estimation of ILO unemployment rates at small area/domain level. The general idea for all further work was to sample from the merged data from LFS and ASIP which is considered as a pseudo-population. Then the estimation procedures described in Subsection 2.2 are applied to each of the samples and the results are analysed.

We will now discuss two applications where the performance of the estimation methods considered is compared for two different setups. In the first setup geographically divided subgroups are considered while in the second example we discuss the performance of methods when estimating ILO unemployment rates for population subgroups that are divided into age/sex groups. It has to be noted that two steps are necessary to estimate ILO-unemployment rates. In a first step the number of unemployed persons in small area $d$ is estimated. Afterwards, the number of employed persons in small area $d$ is estimated too. The ILO-unemployment rate is finally calculated as the number of unemployed divided by the number of persons that build the workforce.

## 3.2 Estimation of ILO-Unemployment Rates for 9 Small Areas

**Simulation Environment** The data base for the simulation study was the 'pseudo-population' already described in Subsection 3.1. A total of 1000 simple random samples have been selected from the pseudo-universe. The sampling fraction of 5% was chosen so that the sample sizes at NUTS2-level in the samples used for the simulation study

Table 1: Distribution of the number of ILO unemployed people in the simulation samples.

| | SA1 | SA2 | SA3 | SA4 | SA5 | SA6 | SA7 | SA8 | SA9 |
|---|---|---|---|---|---|---|---|---|---|
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Median | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 10 |
| Max | 12 | 11 | 13 | 10 | 12 | 15 | 10 | 12 | 23 |

are comparable to those of the Austrian LFS at NUTS3-level (political districts). Due to this fact, for the simulation samples NUTS2-regions are considered as small areas. The numbers of persons unemployed for the resulting nine small areas are listed in Table 1.

**Results and diagnostic criteria** We will now present the main simulation results and discuss the performance of the estimation methods when estimating ILO unemployment rates for the nine small areas defined above.

**RB (relative bias) in %:** In Table 2 the relative bias in % is listed for each of the estimation methods for the small areas of interest. We learn from Table 2 that the relative bias of the direct estimator $\hat{\overline{Y}}^{DIR}$ is less than 2.1% in each of the small areas. The maximal relative bias for $\hat{\overline{Y}}^{GREG}$ is $-2.56\%$. We find that the relative bias for indirect estimation methods is small for subgroups in which the underlying model works well while it can be rather large in subgroups where the model assumptions do not hold. The relative bias for $\hat{\overline{Y}}^{SA}$ in SA2 is only 0.93% while it is more than 25% in SA4. The relative bias of the hierarchical Bayes estimator is generally lower than those of the other indirect estimation methods. For example the RB of the hierarchical Bayes estimator is almost $-5\%$ in SA9 while it is nearly 28% for $\hat{\overline{Y}}^{SA}$ in the same small area. Since $\hat{\overline{Y}}^{EBA}$ and $\hat{\overline{Y}}^{EBB}$ are calculated as linear combinations from a design-based and a model-based estimator, the relative bias in any small area is between the RB of the direct and the RB of the synthetic estimator from which they are calculated.

**RRMSE (relative root mean square error) in %:** The RRMSE is a criteria that may be used to assess the performance of various estimation methods for a selected small area. In Table 3 the values of the RRMSE in % are listed for all estimation methods in the small areas considered.

It can be observed from Table 3 that the RRMSE of indirect methods is almost always

Table 2: Relative bias in % for small areas considered.

| | NSM | DIR | G | SA | SB | EBA | EBB | HB |
|---|---|---|---|---|---|---|---|---|
| SA1 | $-9.49$ | $-0.10$ | $-1.74$ | $-6.59$ | $-3.78$ | $-5.42$ | $-3.51$ | $-0.87$ |
| SA2 | 5.70 | $-0.13$ | $-0.76$ | 0.93 | $-5.85$ | 0.20 | $-5.56$ | $-0.97$ |
| SA3 | 12.43 | 0.04 | $-0.31$ | 7.17 | 2.63 | 4.30 | 2.38 | 2.32 |
| SA4 | 37.76 | 1.55 | 0.78 | 25.67 | 14.87 | 16.67 | 13.94 | 10.00 |
| SA5 | 29.54 | $-2.07$ | $-2.56$ | 7.81 | $-11.66$ | 4.25 | $-11.11$ | $-4.26$ |
| SA6 | 15.03 | $-0.32$ | 0.01 | 15.22 | 16.57 | 9.99 | 15.82 | 7.31 |
| SA7 | 38.50 | $-1.48$ | $-1.00$ | 12.22 | $-12.99$ | 7.45 | $-12.40$ | $-4.66$ |
| SA8 | 11.46 | 1.52 | 1.45 | 9.77 | 4.54 | 6.36 | 4.14 | 3.99 |
| SA9 | $-51.49$ | 0.78 | 0.25 | $-27.96$ | $-4.07$ | $-17.40$ | $-3.61$ | $-4.86$ |

Table 3: Relative root mean square error in % for small areas considered.

| | NSM | DIR | G | SA | SB | EBA | EBB | HB |
|---|---|---|---|---|---|---|---|---|
| SA1 | 16.86 | 51.08 | 40.11 | 13.64 | 29.01 | 17.00 | 29.56 | 43.28 |
| SA2 | 17.16 | 45.03 | 39.66 | 13.29 | 21.49 | 17.67 | 21.73 | 37.08 |
| SA3 | 21.21 | 47.25 | 40.65 | 15.74 | 32.49 | 19.07 | 32.58 | 39.46 |
| SA4 | 43.17 | 49.54 | 43.16 | 30.73 | 32.01 | 28.71 | 31.70 | 44.33 |
| SA5 | 35.51 | 53.34 | 47.14 | 16.72 | 25.89 | 20.01 | 26.56 | 42.62 |
| SA6 | 23.12 | 49.15 | 40.97 | 21.34 | 24.48 | 22.48 | 24.78 | 40.99 |
| SA7 | 43.87 | 51.23 | 45.65 | 19.88 | 31.31 | 21.88 | 31.18 | 43.02 |
| SA8 | 20.53 | 44.83 | 38.77 | 17.46 | 25.21 | 19.09 | 25.45 | 37.81 |
| SA9 | 52.07 | 31.89 | 25.39 | 29.40 | 29.26 | 26.02 | 29.42 | 30.40 |

considerably smaller than those of the direct estimator. The basic estimator $\hat{\overline{Y}}^{NSM}$ has a very small RRMSE in subgroups in which the unemployment rate is approximately equal to the national unemployment rate. We see that indirect methods that are using auxiliary information on unit-level ($\hat{\overline{Y}}^{SA}$ and $\hat{\overline{Y}}^{EBA}$) are performing better in terms of the RRMSE than those procedures that make use of area-specific auxiliary information only ($\hat{\overline{Y}}^{SB}$ and $\hat{\overline{Y}}^{EBB}$). Looking at the performance of the hierarchical Bayesian method, we can state that there is a reduction of the RRMSE compared to the RRMSE of the direct estimator $\hat{\overline{Y}}^{DIR}$. However, the reduction in terms of the RRMSE is smaller than the reduction of the other indirect methods.

**ARRMSE (average relative root mean square error) in %:** The main advantage of using the ARRMSE is that it is an overall performance measure. It allows to evaluate the performance of an estimation method globally and not only for specific small domains.

In Table 4 the values of the ARRMSE for each of the eight estimation methods are listed. The best global performance was achieved by the synthetic estimator $\hat{\overline{Y}}^{SA}$. Estimating ILO-unemployment rates on small area level by this method, it is possible to reduce the average relative root mean square error compared to the Horvitz-Thompson estimator by about 49%. If auxiliary information is only available aggregated at small area level, it is possible to reduce the ARRMSE, compared again to the value of the direct estimator, by more than 37%.

The overall performance of the generalized regression estimator $\hat{\overline{Y}}^{GREG}$ and the hierarchical Bayes method $\hat{\overline{Y}}^{HB}$ is practically equal. It is possible to reduce the ARRMSE compared to that of the direct estimator by approximately 14% using either of these estimators. However, even though the ARRMSE of $\hat{\overline{Y}}^{EBA}$ is slightly higher than that of $\hat{\overline{Y}}^{SA}$ it should be preferred over $\hat{\overline{Y}}^{SA}$. The reason is that the composite procedure performs much better in terms of the relative bias.

Table 4: Average relative root mean square error for small areas considered in %.

| Method | NSM | DIR | G | SA | SB | EBA | EBB | HB |
|---|---|---|---|---|---|---|---|---|
| ARRMSE (%) | 27.02 | 32.08 | 27.48 | 16.40 | 20.01 | 17.03 | 20.03 | 27.77 |

Table 5: Distribution of the number of ILO unemployed people in the simulation samples.

| | SD1 | SD2 | SD3 | SD4 | SD5 | SD6 | SD7 | SD8 |
|---|---|---|---|---|---|---|---|---|
| Min | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Median | 7 | 8 | 7 | 7 | 3 | 3 | 3 | 5 |
| Max | 17 | 18 | 14 | 16 | 10 | 12 | 8 | 13 |
| | SD9 | SD10 | SD11 | SD12 | SD13 | SD14 | SD15 | SD16 |
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Median | 3 | 7 | 3 | 7 | 2 | 6 | 4 | 4 |
| Max | 10 | 16 | 12 | 16 | 8 | 14 | 11 | 11 |

## 3.3 Estimation of ILO-Unemployment Rates for 16 Small Domains

**Simulation Environment** We will now discuss another application and estimate ILO-unemployment rates for several small domains. The domains of interest in this case were age/sex groups. Therefore, we partitioned the pseudo-population into a total of eight 5-year age groups crossed by sex. We obtained a total of 16 small domains (SD1 to SD16) for which the parameter of interest is the domain specific ILO-unemployment rate.

As in the previous example a total of 1000 samples have been drawn from the pseudo-population. The sampling fraction was set at 0.1. The minimal, median and maximal number of unemployed people according to the LFS in the simulation samples for the small domains of interest are listed in Table 5.

For this example we concentrated on an application of the estimators (1)–(7). We did not apply the hierarchical Bayes estimator in this example for several reasons. One reason was its weak performance in the previous example which needs to be examined further. Another reason is that we wanted to focus on feasible methods that have a chance of being considered to be applied in practice.

### Results and Diagnostic Criteria

**RB (relative bias) in %:** In Table 6 the relative bias in % of the estimation methods applied is listed for each of the 16 age/sex groups considered. We see that the maximum relative bias for $\hat{\overline{Y}}_d^{DIR}$ is 2.48%. $\hat{\overline{Y}}_d^{GREG}$ performs slightly better with a maximum relative bias of $-2.09\%$. For the other estimation methods we see the relative bias is exceeding 100% for $\hat{\overline{Y}}_d^{SA}$ in SD13 and for $\hat{\overline{Y}}_d^{NSM}$ in SD11 and SD13. The relative bias of the composite estimators $\hat{\overline{Y}}_d^{EBA}$ is below 25% in each of the small domains. The relative bias of $\hat{\overline{Y}}_d^{EBB}$ is below 36% for all small domains except for SD13 where it is more than 62%.

**RRMSE (relative root mean square error) in %:** The relative root mean squared errors for the estimation methods used in this example are listed in Table 7. We see that the RRMSE for the synthetic estimators $\hat{\overline{Y}}_d^{SA}$ and $\hat{\overline{Y}}_d^{SA}$ are lower than those of the direct estimator in eleven out of the sixteen small domains considered but are considerably larger for example in SD13. The RRMSE of the composite estimator $\hat{\overline{Y}}_d^{EBA}$ is lower than the RRMSE of $\hat{\overline{Y}}_d^{DIR}$ in each of the sixteen small domains. $\hat{\overline{Y}}_d^{EBB}$ is also performing

Table 6: Relative bias in % for small domains considered.

| | NSM | DIR | G | SA | SB | EBA | EBB |
|---|---|---|---|---|---|---|---|
| SD1 | −61.43 | −1.06 | −1.25 | −48.93 | −38.81 | −15.67 | −28.33 |
| SD2 | −74.90 | 0.88 | −0.24 | −61.09 | −48.33 | −17.98 | −34.47 |
| SD3 | −40.95 | 0.47 | −0.25 | −33.21 | −29.48 | −13.06 | −22.00 |
| SD4 | −46.15 | −0.72 | −1.00 | −38.87 | −35.93 | −15.00 | −27.34 |
| SD5 | 12.38 | −0.39 | −1.47 | −9.98 | −14.02 | −8.10 | −11.41 |
| SD6 | −8.17 | 0.56 | 0.70 | −10.99 | −11.18 | −3.17 | −7.69 |
| SD7 | 74.34 | 1.07 | 1.68 | 18.10 | 7.34 | 2.85 | 4.58 |
| SD8 | −11.97 | 0.20 | 0.56 | −13.78 | −13.43 | −4.34 | −9.36 |
| SD9 | 90.34 | −1.90 | −2.09 | 34.26 | 23.48 | 4.45 | 14.65 |
| SD10 | −18.50 | 1.41 | 0.49 | −14.55 | −12.79 | −4.98 | −8.73 |
| SD11 | 138.75 | 2.48 | 1.17 | 66.33 | 52.47 | 15.06 | 35.27 |
| SD12 | −7.01 | −0.21 | −0.98 | −9.78 | −9.16 | −5.58 | −7.27 |
| SD13 | 149.89 | −1.60 | −0.15 | 100.61 | 91.18 | 24.59 | 62.44 |
| SD14 | 0.82 | 0.53 | 0.47 | 2.22 | 4.47 | −1.21 | 1.93 |
| SD15 | 33.97 | −1.14 | −0.52 | 31.09 | 31.53 | 6.27 | 21.53 |
| SD16 | 10.48 | 0.15 | −0.41 | 22.55 | 29.17 | 5.71 | 20.36 |

Table 7: Relative root mean square error in % for small domains considered.

| | NSM | DIR | G | SA | SB | EBA | EBB |
|---|---|---|---|---|---|---|---|
| SD1 | 61.55 | 33.79 | 29.08 | 49.14 | 39.33 | 26.95 | 33.20 |
| SD2 | 74.94 | 30.93 | 26.33 | 61.18 | 48.67 | 28.13 | 39.29 |
| SD3 | 41.38 | 35.86 | 30.14 | 33.70 | 30.60 | 25.50 | 27.79 |
| SD4 | 46.47 | 34.87 | 30.49 | 39.21 | 36.83 | 26.64 | 32.14 |
| SD5 | 16.76 | 52.90 | 43.15 | 12.85 | 18.27 | 28.17 | 22.87 |
| SD6 | 12.33 | 50.00 | 44.06 | 13.37 | 14.71 | 27.62 | 18.66 |
| SD7 | 76.38 | 53.59 | 46.26 | 21.30 | 25.61 | 30.30 | 29.21 |
| SD8 | 14.88 | 41.97 | 35.74 | 15.64 | 16.25 | 24.49 | 18.60 |
| SD9 | 92.34 | 50.35 | 41.87 | 36.47 | 33.76 | 30.26 | 32.52 |
| SD10 | 20.23 | 34.47 | 29.04 | 16.27 | 16.39 | 22.06 | 19.13 |
| SD11 | 140.81 | 54.08 | 46.81 | 68.15 | 61.31 | 37.79 | 49.96 |
| SD12 | 11.68 | 36.80 | 29.32 | 12.51 | 13.03 | 22.29 | 18.30 |
| SD13 | 151.98 | 55.83 | 52.13 | 102.19 | 94.26 | 46.20 | 70.89 |
| SD14 | 10.16 | 39.20 | 33.08 | 9.08 | 11.68 | 23.92 | 17.72 |
| SD15 | 36.54 | 47.38 | 37.58 | 33.07 | 34.47 | 26.12 | 29.15 |
| SD16 | 15.26 | 46.65 | 39.62 | 24.87 | 32.42 | 27.86 | 28.98 |

well, however for SD2 and SD13 its RRMSE is higher than the RRMSE of $\hat{\overline{Y}}_d^{DT}$ in the corresponding small domains.

**ARRMSE (average relative root mean square error) in %:** To be able to evaluate the global performance of the estimation methods for the given task, we are looking at the ARRMSE for the estimation methods considered. The results are given in Table 8.

We see that the composite estimators $\hat{\overline{Y}}_d^{EBA}$ and $\hat{\overline{Y}}_d^{EBB}$ performed best globally with the ARRMSE being 28.39 and 30.53 respectively. Thus, using the composite estimator

Table 8: Average relative root mean square error for small domains considered in %.

| | NSM | DIR | G | SA | SB | EBA | EBB |
|---|---|---|---|---|---|---|---|
| ARRMSE | 51.48 | 43.67 | 37.17 | 34.31 | 32.97 | 28.39 | 30.53 |

$\hat{\overline{Y}}_d^{EBA}$ for estimation ILO-unemployment rates in small domains leads to an overall reduction of the ARRMSE compared to that of the direct estimator $\hat{\overline{Y}}_d^{DT}$ of almost 35%. Using $\hat{\overline{Y}}_d^{SA}$ or $\hat{\overline{Y}}_d^{SB}$ would lead to a reduction of the ARRMSE of more than 21% or 24.5% respectively compared to the direct estimator. However as we can see from Table 6, the relative bias of these procedures is unacceptably high in some small domains.

# 4 Conclusions

After conducting the simulation study we state that in both examples indirect estimation methods were able to outperform the direct, design-based estimator in terms of overall performance criteria like the ARRMSE for domains with small sample sizes. In the first example we could see that the ARRMSE of the synthetic estimator $\hat{\overline{Y}}_d^{SA}$ was approximately 49% smaller than the ARRMSE of the Horvitz-Thompson estimator. In the second example we observed that the ARRMSE of the composite estimator $\hat{\overline{Y}}_d^{EBA}$ was almost 35% smaller than the ARRMSE of $\hat{\overline{Y}}_d^{DIR}$. The results of the simulation study showed as well that one should pay special attention to the bias component of the MSE when comparing different estimation methods. It turned out that both composite estimators performed very well in both examples since they offer a trade-off between bias and variance.

An interesting finding was the performance of the hierarchical Bayes estimator using only auxiliary information on area level in the first application. Even though its variance was smaller than the variance of the direct estimator, we had expected the Bayesian estimator to perform better. The reasons should be investigated further especially since the hierarchical Bayes method is very attractive in a way that valid inference for confidence-intervals can be achieved by simply analyzing simulated data from the resulting posterior distribution. However, the Bayesian procedure is computational intensive and the question remains if the additional efforts are worth the possible gains in performance compared to other indirect estimators. We have then concentrated the remaining model-based estimation procedures that may be implemented in a straightforward way.

We have showed that model-assisted or rather model-based estimation methods provided acceptable results estimating unemployment rates for the small domains under investigation in the simulation. The estimation of unemployment rates at small area level can indeed be improved by using auxiliary information on unit-/area level. However, it is important that suitable auxiliary information on either unit/person or aggregated at area-level needs to be available from administrative sources such as the ASIP.

However, publishing model-based or model-assisted estimation results in official statistics remains often difficult or even impossible. Some of the main difficulties that need to be addressed are the traditional against model-based thinking or the problems in explaining the trade-off from bias and variance to the user. Thus, additional work has still to be done to overcome these problems. An important point is that a detailed manual covering methods, underlying assumptions and possible help on how results can be interpreted should be produced and released together with the estimates (demanded for example by: Heady et al. (2003, p. 61ff)). Providing the user with this kind of information will surely help to overcome the difficulties of accepting model-based estimates in official statistics.

## References

Australian Bureau of Statistics. (2006). *A guide to small area estimation* (Tech. Rep.). Canberra, Australia: Australian Bureau of Statistics.

Haslinger, A., and Kytir, J. (2006). Stichprobendesign, Stichprobenziehung und Hochrechnung des Mikrozensus ab 2004. *Statistische Nachrichten*, *6*, 510-519.

Heady, P., Clarke, P., and al et. (2003). *Small Area Estimation Project Report. Model-Based Small Area Estimation Series No.2* (Tech. Rep.). London, England: Office for National Statistics.

Hussmanns, R., Mehran, F., and Verma, V. (1990). *Surveys of Economically Active Population, Employment, Unemployment, and Underemployment: An ILO Manual on Concepts and Methods*. Geneva, Switzerland: International Labour Office.

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rao, J. (2003). *Small area estimation* (1st ed.). New York: Wiley.

Suciu, G., Hoshaw-Woodard, S., Elliott, M., and Doss, H. (2001). *Uninsured Estimates by County: A Review of Options and Issues* (Tech. Rep.). Ohio: Ohio Department of Health, Center for Public Health Data and Statistics.

The EURAREA Consortium. (2004). *Reference Volume, Part I-III* (Tech. Rep.). Luxemburg: The EURAREA Consortium.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS Open. *R News*, *6*, 12-17.

Author's Address:

Bernhard Meindl
Statistik Austria
Guglgasse 13
A-1110 Wien

E-Mail: Bernhard.Meindl@statistik.gv.at