DACSEIS

IST-2000-26057

Workpackage 10

Variance Estimation for Small Area Estimates

Deliverables 10.1 and 10.2

2004

List of contributors:

Kaja Sõstra, Statistics Finland;
Kersten Magg, Ralf Münnich, Katrin Schmidt, Rolf Wiegert, University of Tübingen.

Main responsibility:

Ralf Münnich and Kersten Magg, University of Tübingen

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

Workpackage 10 investigates small area estimation methods applied to Finnish and German data-sets. The aim of the work is to elaborate the recommended small area methodology that was evaluated in the EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs; IST-2000-26290; `http://www.statistics.gov.uk/methods_quality/eurarea/`) project as the so-called standard estimators especially on the German DACSEIS data-sets.

In contrary to the other workpackages, quality of point estimators have been in the center of attention instead of variance estimators. This is due to an additional agreement between the European Commission and the projects EURAREA and DACSEIS in order to optimize co-operation and to avoid overlapping research.

The main emphasis, therefore, was laid on the adaption of the EURAREA programmes of the 8 standard estimators to the DACSEIS simulation study and to investigate these on the German data-sets.

Special thanks go to Patrick Heady as well as to his team, co-ordinator of the EURAREA project, for the smooth and friendly co-operation and the great support in many tasks of the agreement. We would also like to thank Kaja Sõstra, Statistics Finland, who participated in both projects, for her kind co-operation and her valuable contribution to this report, i. e. Chapter 2 and Section 3.2. Further, we would like to express our thanks to Kari Djerf, Statistics Finland, for many valuable comments.

Kersten Magg, Ralf Münnich University of Tübingen

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In general, sample surveys are designed to provide reliable direct estimates for large areas or domains of a population. But more and more these surveys are also being used to provide estimates for small areas or domains. What features exactly identify an area or domain to be considered *small* can be regarded in different ways. No matter if you attach the characteristic *small* to the absolute or the relative size of a subpopulation, the common problem resulting is that of the sample size in the considered small area being random and small or even zero.

Direct estimates use data only from sample units in the area of interest. Due to the small sample size in the small areas or domains, direct estimators yield standard errors which are unacceptably large. Therefore indirect estimators are constructed that *borrow strength* from related areas, increasing the effective sample size and with it the estimation precision. These indirect estimators are based on either explicit or implicit models providing a link between the small area in question and related areas through ancillary information. These auxiliary variables can be miscellaneous, cross-sectional as well as across time, for example information from neighbouring or next higher populations, data from a previous census or administrative records. Due to the growing demand for reliable small area statistics, small area estimation is becoming an important field in survey sampling.

The EURAREA Project (Enhancing Small Area Estimation Techniques to Meet European Needs) was a three year project, funded by the European Community, to analyze the performance of eight „standard“ – applied by NSIs or much discussed in literature – small area estimators. Central part of the investigation was the empirical evaluation of these estimators´ performance using simulation methods. The simulations were based on repeated sampling from real population register and census datasets of six countries – Sweden, Finland, Poland, Italy, Spain and Britain. Target variables were

- the average equivalised household income,
- the proportion of single-person households, and
- the proportion of individuals which are unemployed

at NUTS3, NUTS4 and NUTS5 level.

As the subject of small area estimation was included in EURAREA als well as in DACSEIS, a cluster between the two projects was formed in order to avoid work being done twice. Common part of EURAREA and DACSEIS was the comparison of the performance of the estimators via variance estimation. General recommendations on the use of small area estimators and overall evaluation of their performance is made by EURAREA.

For purposes of co-operation and comparability DACSEIS adopted seven out of the eight standard estimators analyzed by EURAREA. They were used for simulation studies with the German Microcensus (GMC), an annual 1%-sample of the German population. For reasons of feasibility the simulation study was partly restricted to a sensible selection of German federal states. Several central simulations were conducted on the full sample of the GMC. Concerning the classifications of the small areas, four different settings were used. Details will be discussed in Chapter 3. Another set of simulations was conducted on Finnish data.

Following the research done in EURAREA the target variables for the evaluation of small area estimators in DACSEIS are

- the average disposable income,
- the proportion of single-person households, and
- the proportion of unemployment as surveyed in the GMC

As in EURAREA, the simulation studies were based on real data in order to stay as close as possible to the situation in real life.

In accordance with further remits of DACSEIS, the simulations done in the context of small area estimation cover the performance of the considered estimators in the presence of non-response. In order to compensate for non-response, calibration with corrected weighting and multiple imputation will be investigated.

In the following chapter, the standard estimators used in EURAREA and DACSEIS simulations will be presented. As we are interested not only in the estimators alone but mainly in their performance in terms of bias and precision, the estimators for their Mean Squared Errors (MSE) are explained as well.

In Chapter 3 the results of the simulation studies made with German and Finnish data are exposed. The first part deals with the German simulation study based on the GMC. First, a description of the German data is given. This section is followed by a presentation of the simulation study was carried out on the basis of this data. The last section deals with the effects that non-response and imputation have on small area estimates. The second part of Chapter 3 focuses on the small area estimation based on Finnish data. Here, first the general settings, the connection to the simulations done in the context of EURAREA and the data are presented. Afterwards, the Finnish sampling design and the simulation study are explained. The chapter is completed with a comparison of the performance of the six standard estimators applied to the Finnish dataset.

Chapter 2

Estimators for Small Areas

EURAREAs overall aim was to improve small area estimation methods widely used, or rather currently used within European NSIs. Therefore the project concentrated its evaluation on small area estimators currently being used in the European NSIs or discussed largely in literature. In adopting these estimators for its simulation study DACSEIS focussed its research on the same set of estimators. The following eight small area estimators were considered as „standard estimators“ in the EURAREA and DACSEIS projects:

1. National sample mean;
2. Direct estimator;
3. GREG estimator with a standard linear regression model;
4. Three synthetic estimators:
   (a) Synthetic estimator A using a standard linear normal model with unit-level covariates;
   (b) Synthetic estimator B using a linear normal model with area-level covariates;
   (c) Synthetic estimator C (for binary target variable) using a logistic regression model with area-level covariates;
5. Two EBLUP estimators:
   (a) EBLUP estimator A, a weighted combination of the Synthetic A and the GREG estimator;
   (b) EBLUP estimator B, a weighted combination of the Synthetic B and the direct estimator.

2.1 National Sample Mean

The national sample mean is constant for every area. It is calculated using the following formula:

$$\widehat{\mu}_Y = \sum_{i \in s} \omega_i \, y_i / \widehat{N} \qquad , \tag{2.1}$$

where

$$\widehat{N} = \sum_{i \in s} \omega_i$$

and ω_i is the inverse of the sample inclusion probability of individual i. The national sample mean is not a small area estimator, because it contains no area specific information. This estimator was included for aims of comparison with other estimators which take into account the differences between areas. The national sample mean is a very poor estimator for small areas, as it will produce large errors for areas whose true population value differs severely from the national mean.

2.2 Direct Estimator

The direct estimator of the mean in area d is defined as a ratio of the design-weighted Horvitz-Thompson estimators for each area:

$$\widehat{\mu}_{Y_d} = \sum_{i \in s_d} \omega_i \, y_i / \widehat{N}_d \qquad , \tag{2.2}$$

where

$$\widehat{N}_d = \sum_{i \in s_d} \omega_i \qquad .$$

The sums are taken over sample s_d from area d and the design weights are the inverses of the inclusion probabilities, $\omega_i = 1/\pi_i$. The direct estimator of the domain mean is approximately unbiased (SÄRNDAL *et al*., 1992, p. 185). The precision of the estimator is measured by its MSE, estimated by the following formula (SÄRNDAL *et al*., 1992, p. 391):

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}\right) = \sum_{i \in s_d} \sum_{j \in s_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\left(y_i - \widehat{\mu}_{Y_d}\right)}{\pi_i} \frac{\left(y_j - \widehat{\mu}_{Y_d}\right)}{\pi_j} / \widehat{N}_d^2 \qquad . \tag{2.3}$$

Assuming independence, $\pi_{ij} = \pi_i \cdot \pi_j$, whenever $i \neq j$ we get

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}\right) = \sum_{i \in s_d} \omega_i \ \left(\omega_i - 1\right) \ \left(y_i - \widehat{\mu}_{Y_d}\right)^2 / \widehat{N}_d^2 \qquad . \tag{2.4}$$

2.3 GREG Estimator

To allow for the differences between the sample and population area means of the auxiliary variable X, the direct estimator is adjusted, usually using the standard linear model. This yields the generalised regression estimator

$$\widehat{\mu}_{Y_d}^{\text{GREG}} = \widehat{\mu}_{Y_d} + (\mu_{X_d} - \widehat{\mu}_{X_d})^T \widehat{\beta} \quad , \tag{2.5}$$

where

$$\widehat{\mu}_{X_d} = \sum_{i \in s_d} \omega_i \; x_i / \widehat{N}_d \quad .$$

$\mu_{X_d} = \left(\mu_{X_{d,1}}, ..., \mu_{X_{d,p}}\right)^T$ is the vector of true means of p covariates (x_i is p-dimensional) in the area d and $\widehat{\beta}$ is the least squares regression estimate assuming a standard linear model $y_i = x_i\beta + \varepsilon_i$ with independent errors $\varepsilon_i \sim N(0, \sigma^2)$ for each unit i in the sample:

$$\widehat{\beta} = \left(\sum_{i \in s} \omega_i x_i x_i^T \right)^{-1} \sum_{i \in s} \omega_i x_i y_i \quad .$$

Note that β is estimated using the whole sample s. An alternative presentation for the GREG estimator is through g-weights:

$$\widehat{\mu}_{Y_d}^{\text{GREG}} = \sum_{i \in s} \omega_i \; g_{di} \; y_i \quad , \tag{2.6}$$

where g-weights depend on the domain d, element i and whole sample s:

$$g_{di} = \tfrac{N_d}{\widehat{N}_d} z_{di} + N_d \, (\mu_{X_d} - \widehat{\mu}_{X_d})^T \left(\sum_{i \in s} \omega_i x_i x_i^T \right)^{-1} x_i \quad ,$$

with domain indicators $z_{di} = 1$, if $i \in s_d$ and $z_{di} = 0$, otherwise.

The main property of the g-weights is that g-weighted sample sum of the auxiliary values equals the known domain total of these values (Särndal *et al*., 1992, p. 401):

$$\sum_{i \in s} \omega_i \; g_{di} \; x_i = \sum_{i \in U} z_{di} \; x_i = \sum_{i \in U_d} x_i \quad .$$

An estimate for the MSE of the GREG estimator is derived as follows (SÄRNDAL *et al*, 1992, p. 401):

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}^{\mathrm{GREG}}\right) = \sum_{i\in s}\sum_{j\in s} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{g_{di}e_i}{\pi_i} \frac{g_{dj}e_j}{\pi_j} / \widehat{N}_d^2 \quad . \tag{2.7}$$

Assuming that $\pi_{ij} = \pi_i \cdot \pi_j$, whenever $i \neq j$ we get

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}^{\mathrm{GREG}}\right) = \sum_{i\in s} \omega_i \ (\omega_i - 1) \ g_{di}^2 \ e_i^2 / \widehat{N}_d^2 \quad , \tag{2.8}$$

with residuals $e_i = y_i - x_i^T \ \widehat{\beta}$.

2.4 Synthetic Estimators

An estimator is called a synthetic if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for at least one of these small areas. The underlying assumption is that of the small areas having the same sample characteristics as the large area. If this assumption is satisfied, the synthetic estimator is very efficient as its MSE is small. For areas with strong individual effects the synthetic estimator can be heavily biased (RAO, 2003, p. 46). In DACSEIS, two synthetic estimators were considered. Synthetic estimator A uses an individual model with unit-level covariates and synthetic estimator B makes use of an area-level model.

2.4.1 Synthetic Estimator A

Assuming that unit-level auxiliary data $x_{di} = (x_{di1}, ..., x_{dip})^T$ are available for each population element i in small area d, the variable of interest y_{di} is related to x_{di} through a nested error linear regression model:

$$y_{di} = x_{di}^T\beta + u_d + \varepsilon_{di} \quad , \tag{2.9}$$

where β is the regression coefficients vector, u_d is the area-specific effect with $\mathrm{E}(u_d) = 0$, $\mathrm{var}(u_d) = \sigma_u^2$ and ε_{di} is the independent random error with $\mathrm{E}(\varepsilon_{di}) = 0$ and $\mathrm{var}(\varepsilon_{di}) = \sigma_\varepsilon^2$. The synthetic estimator is given by the formula

$$\widehat{\mu}_{Y_d}^{\mathrm{SYNA}} = \mu_{X_d}^T \widehat{\beta} \quad , \tag{2.10}$$

with known area-level covariates vector $\mu_{X_d} = (\mu_{X_{d1}}, ..., \mu_{X_{dp}})^T$.

The MSE of the synthetic estimator A is obtained using estimates of σ_u^2 and the estimated variance-covariance matrix of $\widehat{\beta}$:

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}^{\mathrm{SYNA}}\right) = \widehat{\sigma}_u^2 + \mu_{X_d}^T \widehat{\mathrm{var}}(\widehat{\beta}) \mu_{X_d} \quad . \tag{2.11}$$

2.4.2 Synthetic Estimator B

The synthetic estimator B uses a linear normal model with area-level covariates and a pooled sample estimate of within-area variance. The model is

$$\overline{y}_{d.} = \mu_{X_d}^T \beta + \xi_d \quad , \tag{2.12}$$

where ξ_d is the area effect with $\mathrm{E}(\xi_d) = 0$ and $\mathrm{var}(\xi_d) = \sigma_u^2 + \psi_d$. The sampling variance is $\psi_d = \sigma_\varepsilon^2 / n_d$, where n_d is the sample size of area d.

The variance σ_ε^2 is estimated by

$$\sigma_\varepsilon^2 = \frac{1}{n-m} \sum_{d=1}^{m} \sum_{i=1}^{n_d} (y_{di} - \overline{y}_{d.})^2 \quad . \tag{2.13}$$

In the regression model vector β and σ_u^2 are estimated iteratively by the formula

$$\widehat{\beta} = (\mathrm{x}^T \mathrm{D}^{-1} \mathrm{x})^{-1} \mathrm{x}^T \mathrm{D}^{-1} \mathrm{y} \quad ,$$

where y is the vector of sample elements and x is the matrix composed of rows $\overline{x}_{d.}^T$. Matrix $\mathrm{D} = \mathrm{diag}\{\mathrm{var}(\xi_d)\}$ is updated iteratively.

The synthetic estimator is given by the formula

$$\widehat{\mu}_{Y_d}^{\mathrm{SYNB}} = \mu_{X_d}^T \widehat{\beta} \quad . \tag{2.14}$$

Its MSE is estimated using estimates of σ_u^2 and the estimated variance-covariance matrix $\widehat{\mathrm{var}}(\widehat{\beta}) = (\mathrm{x}^T \mathrm{D}^{-1} \mathrm{x})^{-1}$ as follows:

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}^{\mathrm{SYNB}}\right) = \widehat{\sigma}_u^2 + \mu_{X_d}^T \widehat{\mathrm{var}}(\widehat{\beta}) \mu_{X_d} \quad . \tag{2.15}$$

2.4.3 Synthetic Estimator C

Synthetic estimator C uses a logistic regression model with area-level covariates and is suited for binary target variables. Suppose the target variable y_i is binary and the parameters of interest are the small area proportions

$$\mu_{Y_d} = p_d = \sum_{i \in U_d} y_i / N_d \qquad .$$

The total of each area is assumed to follow the binomial distribution $y_{d.} \sim B(n_d,\, p_d)$ where n_d is the sample size of area d. The probability p_d obeys the following model with random area effect u_d:

$$\text{logit}(p_d) = \log \tfrac{p_d}{1-p_d} = \overline{x}_d^T \beta + u_d,\ u_d \sim iid\ N(0, \sigma_u^2) \qquad . \tag{2.16}$$

The synthetic estimator is given by the formula

$$\widehat{\mu}_{Y_d}^{\text{SYNC}} = \text{logit}^{-1}\left(\mu_{X_d}^T \widehat{\beta}\right) \qquad . \tag{2.17}$$

It is mentioned here because it has been classified as a standard estimator by EURAREA and was used in their study. In the work of DACSEIS this estimator was not made use of.

2.5 EBLUP Estimators

Composite estimators attempt to balance the potential bias of synthetic estimators and the instability of direct estimators. In EURAREA and DACSEIS, two composite estimators were used combining a direct or GREG estimator with a synthetic A or synthetic B estimator. Both composite estimators are BLUPs (best linear unbiased predictor) for small areas.

2.5.1 EBLUP A Estimator

The EBLUP A estimator is a weighted combination of the synthetic A estimator and the GREG estimator. It is given by the formula

$$\widehat{\mu}_{Y_d}^{\text{EBLUPA}} = \gamma_d \widehat{\mu}_{Y_d}^{\text{GREG}} + (1-\gamma_d)\widehat{\mu}_{Y_d}^{\text{SYNA}} = \gamma_d(\widehat{\mu}_{Y_d} - \widehat{\mu}_{X_d}^T \widehat{\beta}) + \mu_{X_d}^T \widehat{\beta} \qquad , \tag{2.18}$$

where

$$\gamma_d = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2/n_d} \quad .$$

$\widehat{\mu}_{Y_d}$ and $\widehat{\mu}_{X_d}$ are the sample means of the target variable y and the vector of auxiliary variables x for area d respectively; $\widehat{\beta}$, $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\sigma}_u^2$ are the parameter estimates of the two-level linear model. The weight γ_d $(0 \leq \gamma_d \leq 1)$ measures the model variance $\widehat{\sigma}_u^2$ relative to the total variance $\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2/n_d$. If the model variance is relatively small, more weight is attached to the synthetic component. On the other hand, more weight is attached to GREG estimator if the domain sample size n_d increases (Rao, 2003, p. 136).

The MSE estimator consists of three components (Rao, 2003, p. 139-140):

$$\widehat{\mathrm{MSE}}\left(\widehat{\mu}_{Y_d}^{\mathrm{EBLUPA}}\right) = g_{1d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) + g_{2d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) + 2g_{3d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) \quad . \tag{2.19}$$

The first term $g_{1d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$ takes sampling variances into account:

$$g_{1d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) = \gamma_d\left(\widehat{\sigma}_\varepsilon^2/n_d\right) \quad .$$

The second term $g_{2d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$ accounts for the variability in the estimator $\widehat{\beta}$:

$$g_{2d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) = (1-\gamma_d)^2\left(\mu_{X_d}^T \widehat{\mathrm{var}}(\widehat{\beta})\mu_{X_d}\right) \quad .$$

The third term $g_{3d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$ is due to estimating $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$:

$$\begin{aligned} g_{3d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right) &= (1/n_d)^2\left(\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2/n_d\right)^{-3} \\ &\quad \left\{\widehat{\sigma}_\varepsilon^4 \widehat{\mathrm{var}}(\widehat{\sigma}_u^2) + \widehat{\sigma}_u^4 \widehat{\mathrm{var}}(\widehat{\sigma}_\varepsilon^2) - 2\widehat{\sigma}_\varepsilon^2 \widehat{\sigma}_u^2 \widehat{\mathrm{cov}}(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2)\right\} \quad , \end{aligned}$$

where $\widehat{\mathrm{var}}(\widehat{\sigma}_u^2)$ and $\widehat{\mathrm{var}}(\widehat{\sigma}_\varepsilon^2)$ are the asymptotic variances of the estimators $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\mathrm{cov}}(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2)$ is the asymptotic covariance of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\varepsilon^2$. The leading term of MSE estimator, $g_{1d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$, is of order $O(1)$, whereas the terms $g_{2d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$ and $g_{3d}\left(\widehat{\sigma}_u^2, \widehat{\sigma}_\varepsilon^2\right)$ are of lower order.

The EBLUP estimator provides increase in efficiency if γ_d is small. Models with smaller γ_d should be preferred if they provide adequate fit according to model diagnostics (Rao, 2003, p. 137).

2.5.2 EBLUP B Estimator

The EBLUP B estimator is a weighted combination of the synthetic B and the direct estimator. It is given by the formula:

$$\widehat{\mu}_{Y_d}^{\text{EBLUPB}} = \gamma_d \widehat{\mu}_{Y_d} + (1-\gamma_d)\widehat{\mu}_{Y_d}^{\text{SYNB}} = \gamma_d \widehat{\mu}_{Y_d} + (1-\gamma_d)\mu_{X_d}^T \widehat{\beta} \quad , \tag{2.20}$$

where $\gamma_d = \widehat{\sigma}_u^2/(\widehat{\sigma}_u^2 + \widehat{\psi}_d)$; $\widehat{\mu}_{Y_d}$ and $\widehat{\mu}_{X_d}$ are the sample means of the target variable y and the vector of auxiliary variables x for area d respectively; $\widehat{\beta}$, $\widehat{\psi}_d$ and $\widehat{\sigma}_u^2$ are the parameter estimates of two-level linear model .

The weight γ_d $(0 \leq \gamma_d \leq 1)$ measures uncertainty in modeling the μ_{Y_d}. If the model variance $\widehat{\sigma}_u^2$ is relatively small compared to the total variance $\widehat{\sigma}_u^2 + \widehat{\psi}_d$, more weight is attached to the synthetic estimator. If the design variance $\widehat{\psi}_d$ is relatively small, more weight is attached to the direct estimator. The form for $\widehat{\mu}_{Y_d}^{\text{EBLUPB}}$adjusts the synthetic estimator $\mu_{X_d}^T \widehat{\beta}$ to account for model uncertainty. EBLUP B estimator is valid for general sampling designs because it is modeling $\widehat{\mu}_{Y_d}$'s and not the individual elements of population, and the direct estimator uses the design weights. Estimator $\widehat{\mu}_{Y_d}^{\text{EBLUPB}}$ is design-consistent because $\gamma_d \to 1$ as the sampling variance $\psi_d \to 0$ (RAO, 2003, p. 117).

MSE estimator consists of three terms (RAO, 2003, p. 128):

$$\widehat{\text{MSE}}\left(\widehat{\mu}_{Y_d}^{\text{EBLUPB}}\right) = g_{1d}\left(\widehat{\sigma}_u^2\right) + g_{2d}\left(\widehat{\sigma}_u^2\right) + 2g_{3d}\left(\widehat{\sigma}_u^2\right) \quad . \tag{2.21}$$

The first term $g_{1d}\left(\widehat{\sigma}_u^2\right)$ takes into account design variance:

$$g_{1d}\left(\widehat{\sigma}_u^2\right) = \gamma_d \widehat{\psi}_d \quad .$$

The second term $g_{2d}\left(\widehat{\sigma}_u^2\right)$ accounts for the variability in the estimator $\widehat{\beta}$:

$$g_{2d}\left(\widehat{\sigma}_u^2\right) = (1-\gamma_d)^2 \left(\mu_{X_d}^T \widehat{\text{var}}(\widehat{\beta}) \mu_{X_d}\right) \quad .$$

The third term $g_{3d}\left(\widehat{\sigma}_u^2\right)$ is due to estimating $\widehat{\sigma}_u^2$:

$$g_{3d}\left(\widehat{\sigma}_u^2\right) = \widehat{\psi}_d^2 \left(\widehat{\sigma}_u^2 + \widehat{\psi}_d\right)^{-3} \widehat{\text{var}}(\widehat{\sigma}_u^2) \quad .$$

where $\widehat{\text{var}}(\widehat{\sigma}_u^2)$ is the asymptotic varianc of the estimator $\widehat{\sigma}_u^2$. The leading term of MSE estimator, $g_{1d}\left(\widehat{\sigma}_u^2\right)$, is of order $O(1)$. The terms $g_{2d}\left(\widehat{\sigma}_u^2\right)$ and $g_{3d}\left(\widehat{\sigma}_u^2\right)$ are of lower order. Comparison of leading term $\gamma_d \psi_d$ with MSE of the direct estimator ψ_d shows that $\widehat{\mu}_{Y_d}^{\text{EBLUPB}}$ leads to large gains in efficiency when γ_d is small, that is the variability of the model error u_d is small relative to the total variability (RAO, 2003, p. 117).

Chapter 3

Small Area Problems in DACSEIS

3.1 The German Microcensus

3.1.1 Description of the German Data

The GMC is a sample with 1% sampling fraction based on individual data. It is conducted annually using a rotating panel system. The sampling method used is a stratified random sample, applying the variables region and size of building as stratum variables. Sampled elements are clusters of households, defined as buildings, parts of buildings and, if small buildings are involved, groups of several buildings. The clusters are selected by a semi-systematic procedure leading to the stratified sample mentioned above. The pseudo universe used for the simulation performed in DACSEIS relies on real Microcensus survey data from 1996.

Altogether the data includes 214 regional classes and 5 house size classes. In Table 3.1, the partition of the 16 German federal states into regional classes is illustrated. Federal states used in the simulation study are marked dark, those not included in light grey. The number of households and persons represented in the pseudo universe are displayed in Table 3.2. In total the GMC data includes 37,409,881 households with altogether 82,914,752 persons.

The data includes information about the place of residence as well as information about the individuals living in the households like gender, age, labour force status, income etc. As mentioned above a pseudo universe is used in the simulation study. For description of how this synthetic data is obtained see workpackage 3, Chapter 4. Prior to presenting selected results in Chapter 3.1.2 some more details concerning the universe will be shown. The universe used for the simulation study contains the variables listed in Table 3.3, including their coding and the possible outcomes.

Table 3.1: Partition federal states into regional classes and number of classes

	Federal State	**Regional Classes**	**Number of Classes**
1	Schleswig-Holstein (SWH)	1 - 7	7
2	Hamburg (HAM)	8 - 9	2
3	Niedersachsen (NIE)	10 - 30	21
4	Bremen (BRE)	31 - 32	2
5	Nordrhein-Westfalen (NRW)	33 - 76	44
6	Hessen (HES)	77 - 93	17
7	Rheinland-Pfalz (RLP)	94 - 106	13
8	Baden-Württemberg (BAW)	107 - 132	26
9	Bayern (BAY)	133 - 166	34
10	Saarland (SAL)	167 - 169	3
11	Berlin (BER)	170 - 174	5
12	Brandenburg (BRA)	175 - 179	5
13	Mecklenburg-Vorpommern (MVP)	180 - 185	6
14	Sachsen (SAC)	186 - 199	14
15	Sachsen-Anhalt (SAA)	200 - 205	6
16	Thüringen (THN)	206 - 214	9

Table 3.2: Number of households and number of persons in the German pseudo universe

Federal State	**Number of Household**	**Number of Persons**
SWH	1,343,312	2,924,508
HAM	952,755	1,794,844
NIE	3,282,729	7,363,264
BRE	347,640	687,042
NRW	7,858,812	17,357,578
HES	2,748,344	6,128,783
RLP	1,824,484	4,155,488
BAW	4,774,624	10,590,105
BAY	5,612,198	12,719,037
SAL	514,464	1,087,930
BER	1,870,256	3,580,162
BRA	1,116,202	2,647,942
MVP	747,752	1,793,669
SAC	2,078,257	4,655,381
SAA	1,209,769	2,799,846
THN	1,128,283	2,629,173
total	37,409,881	82,914,752

Table 3.3: Variables included in the GMC pseudo universe

Variables	**Abbreviation**	**Possible Outcomes**
regional stratum	RS	number of regional stratum
housesize class	GGK	1 small buildings 2 medium-sized buildings 3 large buildings 4 institutions 5 new buildings
sampling unit	AWB	number of sampling unit
household number	HH	number of household
age	AGE	0 - 94 age in years 95 95 or more
gender	SEX	0 male 1 female
ethnicity	NAT	0 German 1 EU foreigner 2 non EU foreigner
duration of job-seeking	DJS	0 missing or non-seeking 1 up to 6 months 2 6 to 12 months 3 more than 12 months
labour force status	ELO	0 employed labour force 1 unemployed labour force 2 non labour force
registered at the employment center	ALO	0 employed 1 registered unemployed
level of education	EDU	0 children under 15 1 practical training 2 training 3 foreman, technician 4 technical school in former DDR 5 university of applied sciences 6 university 7 missing
class of income	INC	0 $<$ 300 DM 1 [300;600) DM 2 [600;1,000) DM 3 [1,000;1,400) DM 4 [1,400;1,800) DM 5 [1,800;2,200) DM 6 [2,200;2,500) DM 7 [2,500;3,000) DM $\vdots$ $\vdots$ 16 [7,000;7,500) DM 17 $\geq$ 7,500 DM 18 self-employed farmer in main activity 19 non income 20 missing

For purposes of comparability of the outcomes of research, we adopted the variables EURAREA used in their simulation study. Thus, the target variables considered in DACSEIS are the rate of unemployment, the household composition and the net income. The rate of unemployment will be computed as number of unemployed divided by the true population size. The target variable household composition represents the ratio of quantity of single-person households by total of all households. The third and last variable which will be considered is the net income. Thereby we estimated the average equivalised household net income. With respect to the first target variable mentioned above there was indeed individual data in use. Concerning the last two (household composition and income) it was used an aggregated data set on household level. It was computed the proportion of single-households and accordingly the average household net income based on auxiliary variables concerning the head of the household or the first named person in the household.

In the simulation study different kinds of auxiliary information were tested. In the context of the estimation of unemployment there are two sets of covariates. The first one only uses the information of the status of registration at the employment center. The second set of covariates uses the age class [25;65], gender and highest educational level. Considering the estimation the proportion of single-person households the covariates age class [25;65], gender and equivalised net income were used. With respect to the estimation of net income the auxiliary information age class [25;65], gender, higher education, unemployment and number of persons per household were made use of. The variable higher education is a binary variable which is 1 if the person has the university degree and 0 else. The supplement *equivalised* at the auxiliary variable income corresponds to the modified OECD definition used by EUROSTAT (cf. THE EUAREA CONSORTIUM, 2003). The average equivalised household net income will be computed by

$$\frac{\text{Total household net income}}{1 + 0.5 * (\text{No of people aged 14 and over} - 1) + 0.3 * (\text{No of children aged under 14})}.$$

As far as the categorisations of small areas are concerned there are four different cases examined in the study. The small areas will be represented by (i) regional strata, (ii) house size classes, (iii) regional strata and house size classes and (iv) by a classification related to NUTS5. The classification mentioned last is formed by dividing each regional stratum in order to define various comparable small areas. In this case a divisor of eight was chosen. The areas formed that way have about the same size.

The four small area categorisations mentioned above were computed on the basis of eight different universes. The universes are the states marked in Table 3.1, that is the federal states Baden-Württemberg, Berlin, Hamburg, Mecklenburg-Vorpommern, Sachsen, Saarland and Schleswig-Holstein as well as a universe containing all 16 federal states. For all these cases $R = 1,000$ replications were done.

Below the relevant variables with respect to distributions and contingencies will be described. Thereby it will be distinguished between the four different kinds of small area categorisations. For reasons of lucidity however, only a choice of federal states will be considered in the following graphical analysis. For that purpose Baden-Württemberg and Saarland will be analysed.

At first the following graphs (Figure 3.1 to 3.8) will illustrate the distribution of the variables without any small area categorisations.

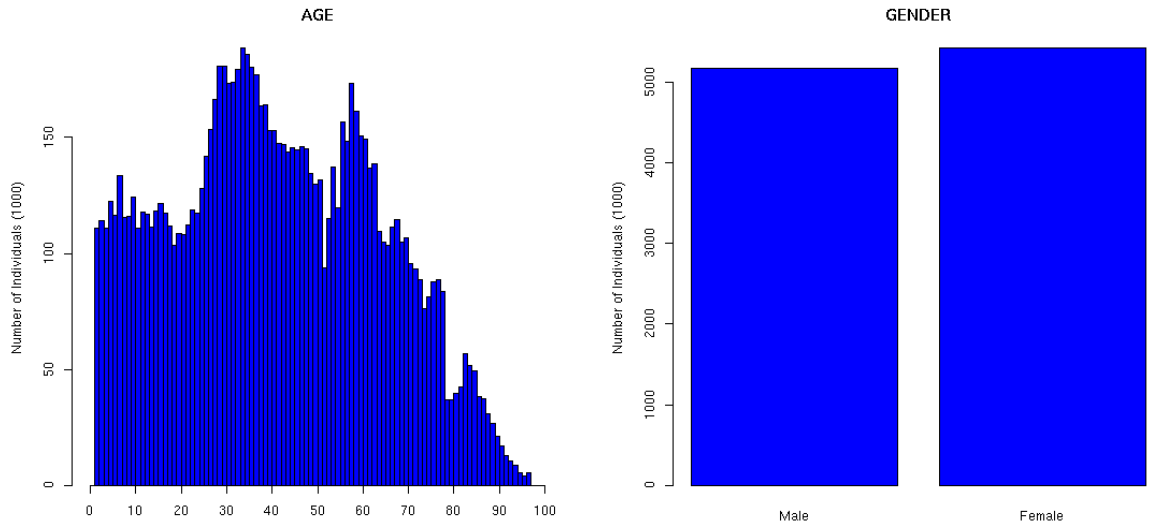

Figure 3.1: Distribution of age and gender in Baden-Württemberg

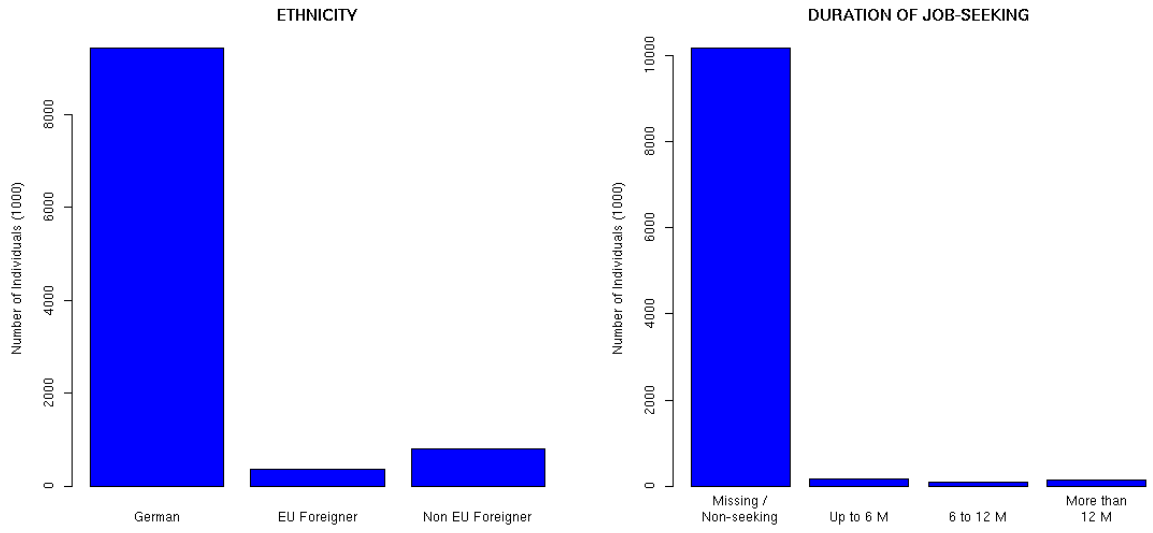

Figure 3.2: Distribution of ethnicity and duration of job-seeking in Baden-Württemberg

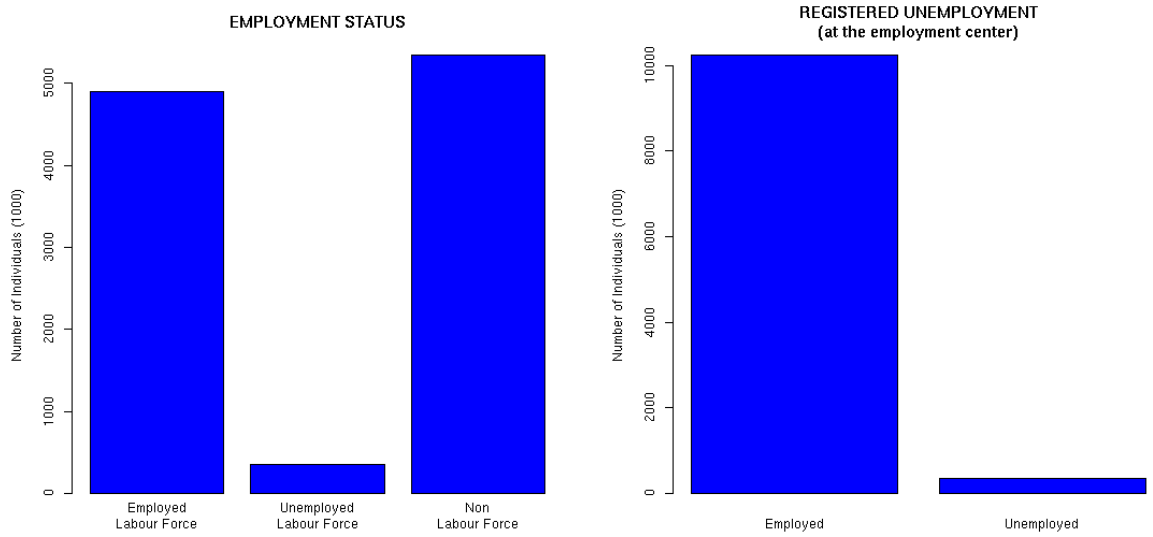

Figure 3.3: Distribution of unemployment and registration at the employment center in Baden-Württemberg

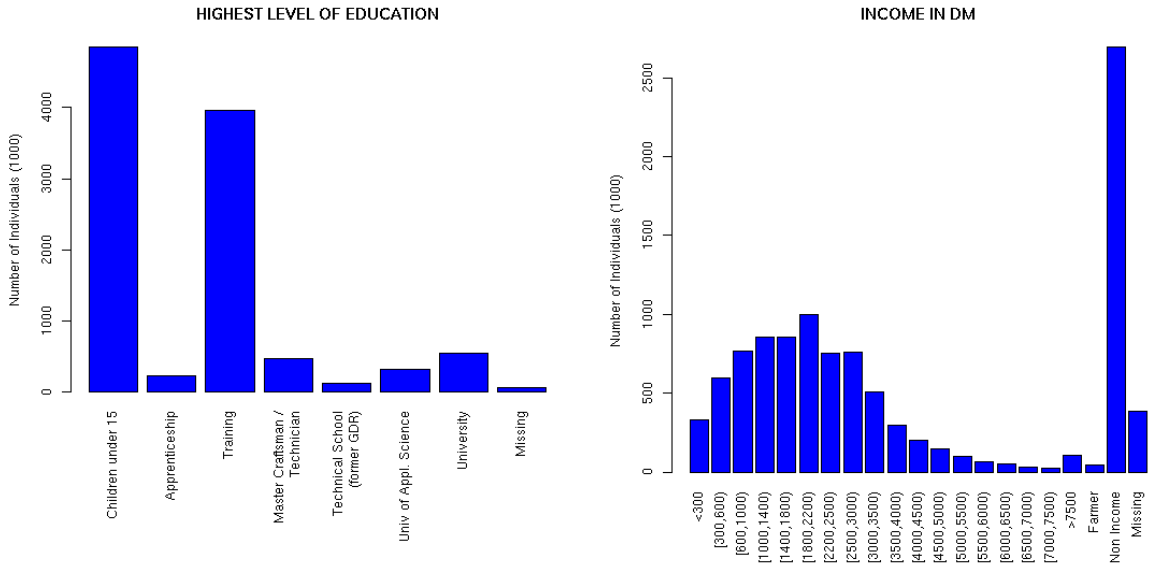

Figure 3.4: Distribution of education and income in Baden-Württemberg

The graphs above show the specifics concerning the variables age, gender, ethnicity, duration of job-seeking, unemployment, registration at the employment center, education and income as well as concerning the subunit federal state Baden-Württemberg. These graphs are generated to get a better overview of the variables used. Below one can see the same illustration but with regards to the subpopulation Saarland as a smaller federal state.

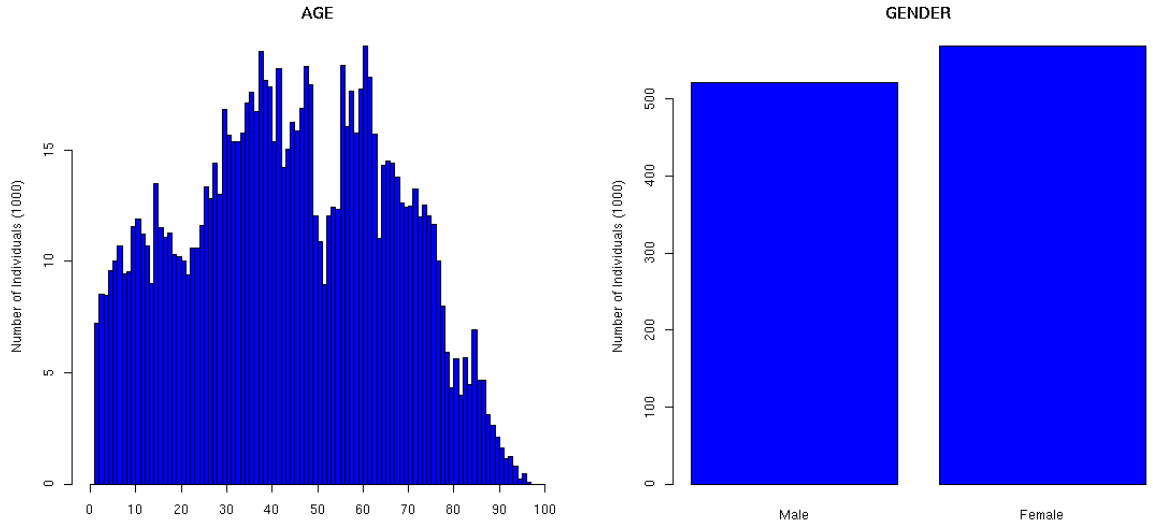

Figure 3.5: Distribution of age and gender in Saarland

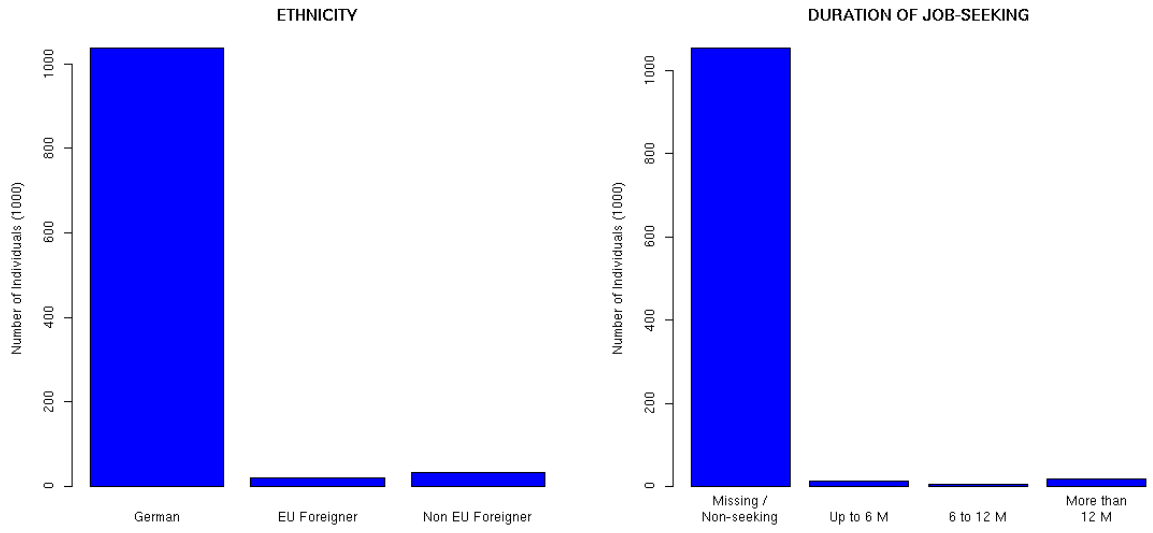

Figure 3.6: Distribution of ethnicity and duration of job-seeking in Saarland

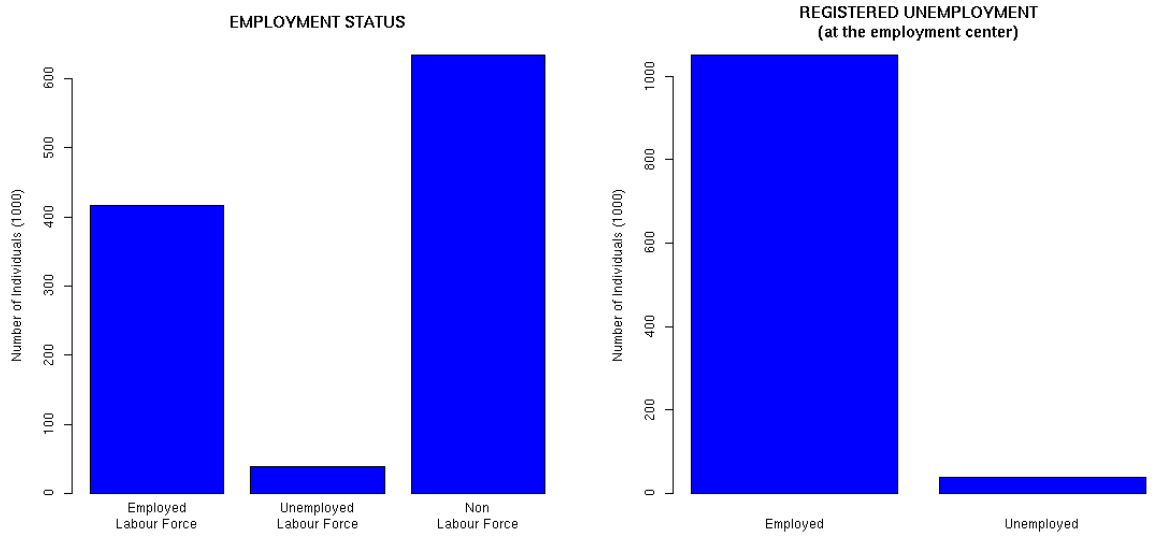

Figure 3.7: Distribution of unemployment and registration at the employment center in Saarland

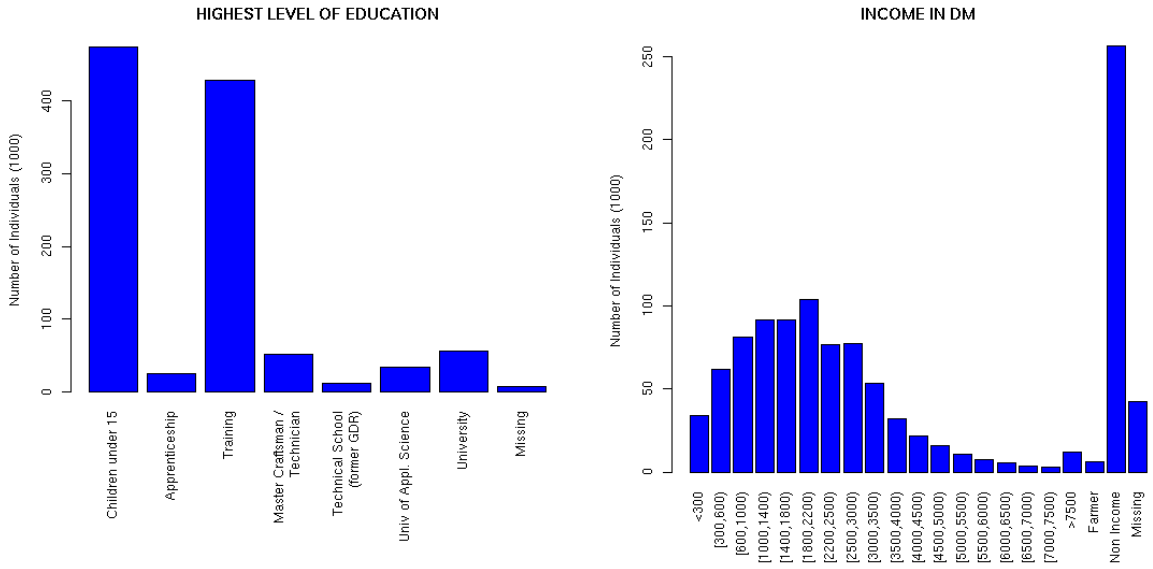

Figure 3.8: Distribution of education and income in Saarland

Another approach, adapted to the context of small area estimation is the character of the following figures. These graphs will present the conditional distribution of the known variables but now concerning the classification of the different small areas. Thereto the distinction will be made between the categorisations (i) to (iv) mentioned above. The graphs with respect to the classification (iv) will not be considered in the following. Because of to many plotted units in the graphs the characteristic could not be identified. For this reason only an extract will be shown below.

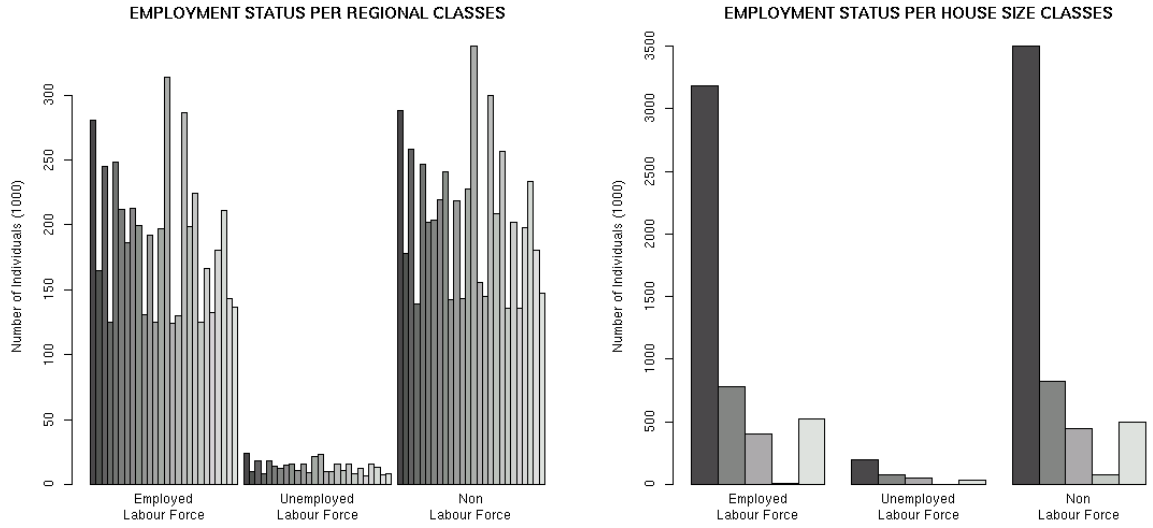

Figure 3.9: Distribution of unemployment in Baden-Württemberg concerning the classification according 26 regional classes and 5 house size classes

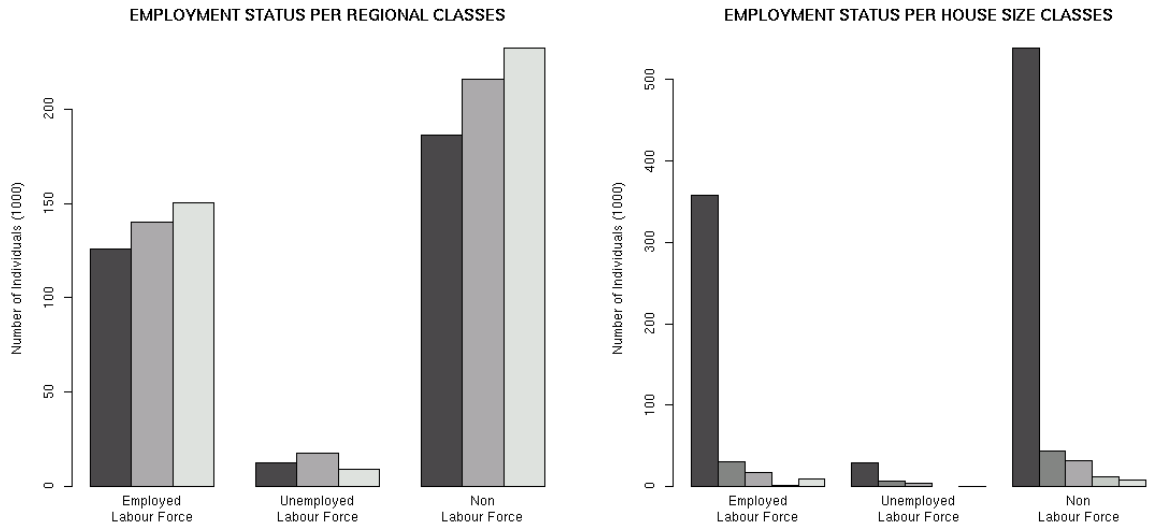

Figure 3.10: Distribution of unemployment in Saarland concerning the classification according 3 regional classes and 5 house size classes

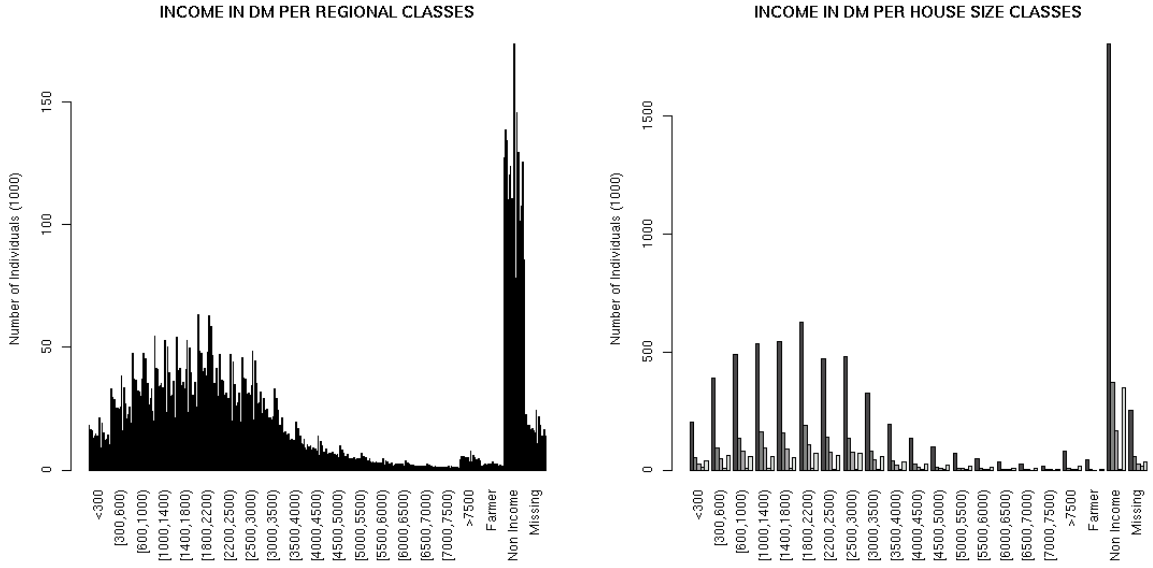

Figure 3.11: Distribution of income in Baden-Württemberg concerning the classification according 26 regional classes and 5 house size classes

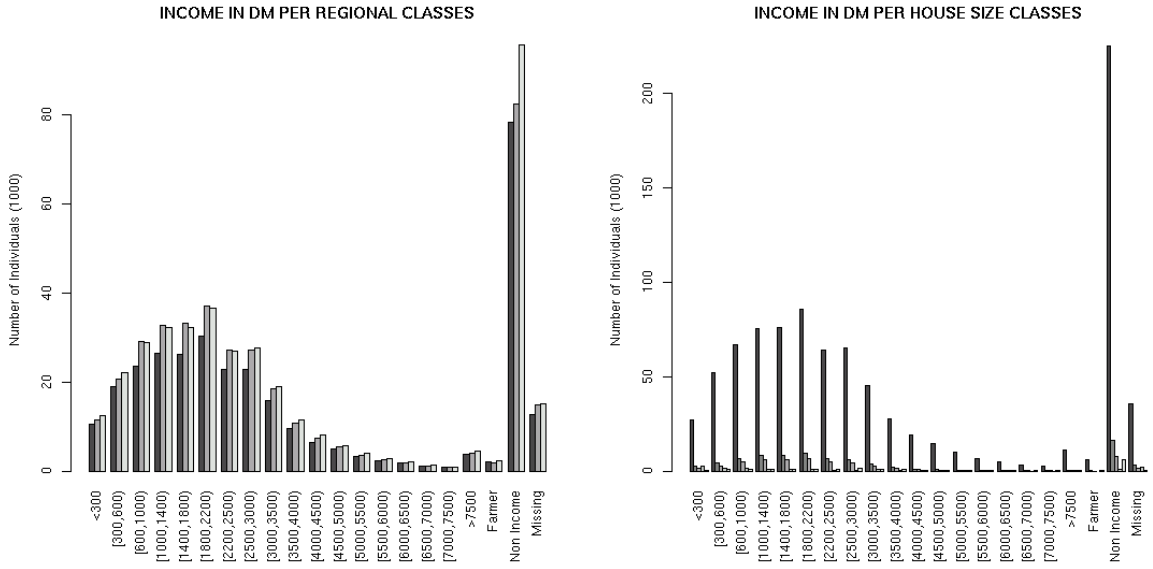

Figure 3.12: Distribution of income in Saarland concerning the classification according 3 regional classes and 5 house size classes

The Figures 3.9 to 3.12 show the distribution of unemployment and income in Saarland and Baden-Württemberg in respect to regional classes and house size classes. The differences between the federal states are not very remarkable but between the small area classes. Particularly in the cases of classification by house size classes one can observe that the distribution processes are very heterogeneous. House size class number four, the class of institutions, differs extremely from the another four. Both the target variable unemployment and the target variable income present that specific characteristic (cf. Figure 3.9 and 3.11 or Figure 3.10 and 3.12).

Another interesting view is to have a look on the distribution of the auxiliary variables. Below one can see an overview to the variables education, gender, ethnicity, registered unemployment as well as age. The graphs belong to the classification by house size classes. The figures in Baden-Württemberg are similar to Saarland and therefore omitted here.

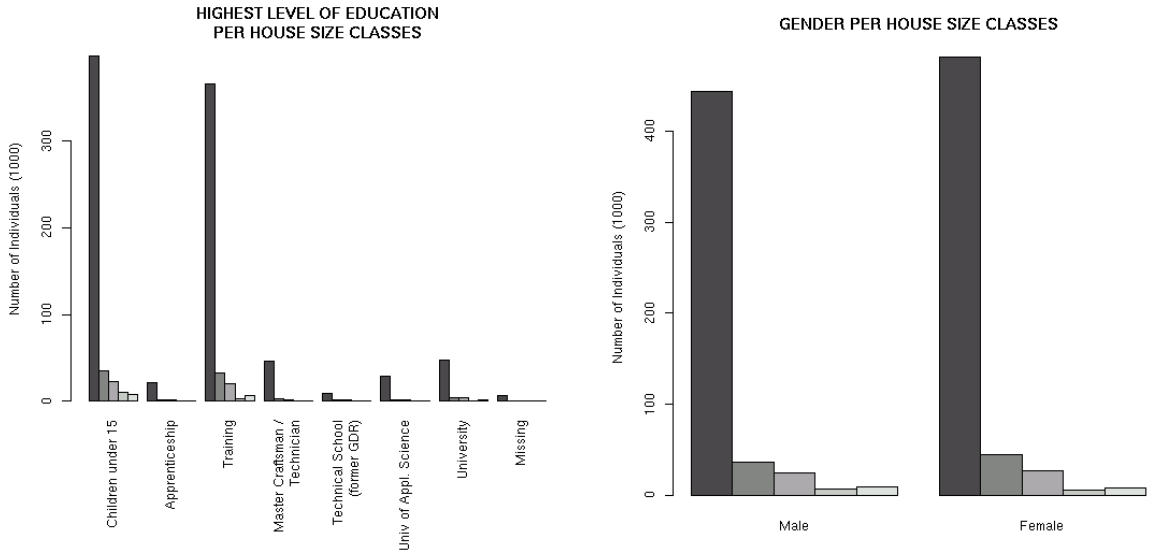

Figure 3.13: Distribution of education and gender in Saarland concerning the classification according 5 house size classes

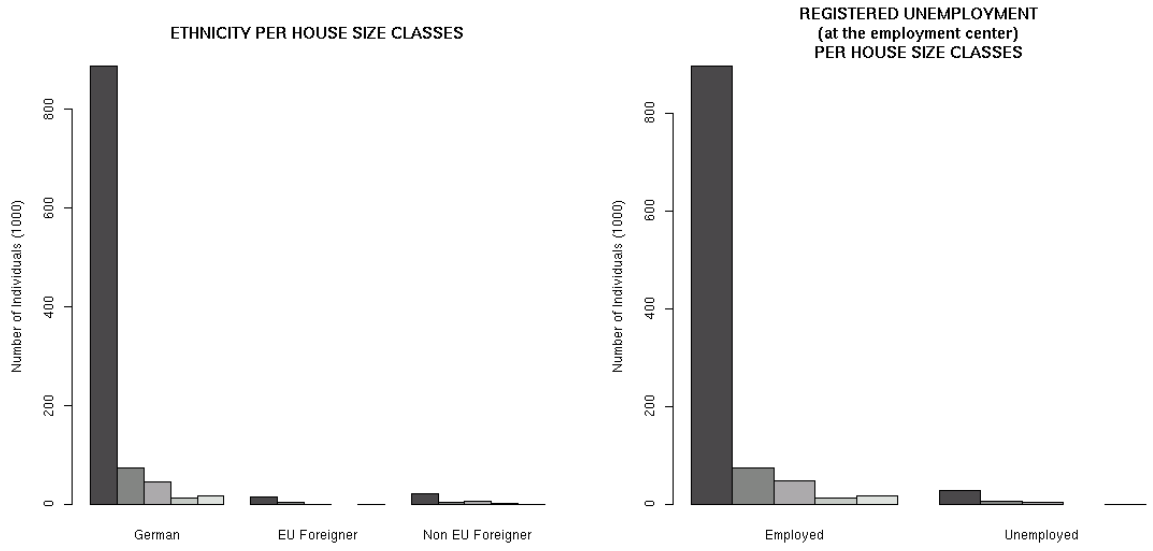

Figure 3.14: Distribution of ethnicity and registered unemployed in Saarland concerning the classification according 5 house size classes

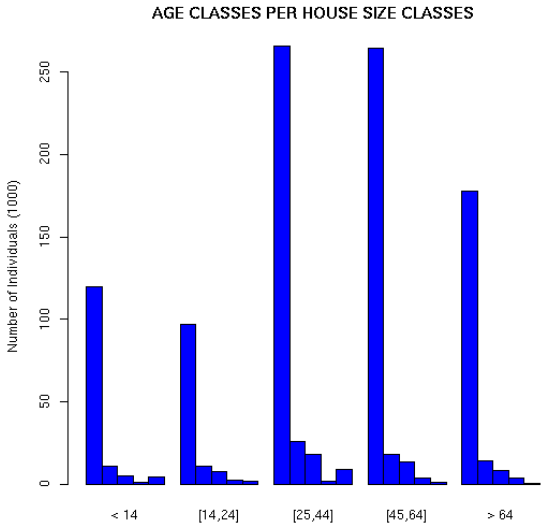

Figure 3.15: Distribution of age classes in Saarland concerning the classification according 5 house size classes

Table 3.4: Contingency coefficients of all considered variables in Saarland and Baden-Württemberg

Saarland	age	gender	ethn.	dojs	unempl.	reg. unem.	edu.	inc.
age	1.0000	0.1055	0.1305	0.1284	0.5335	0.1289	0.4933	0.5834
gender	0.1055	1.0000	0.0378	0.0494	0.2027	0.0599	0.1879	0.3609
ethn.	0.1305	0.0378	1.0000	0.0654	0.0759	0.0519	0.1253	0.0998
dojs	0.1284	0.0494	0.0654	1.0000	0.6224	0.6355	0.0607	0.0646
unempl.	0.5335	0.2027	0.0759	0.6224	1.0000	0.6802	0.3453	0.3323
reg. unem.	0.1289	0.0599	0.0519	0.6355	0.6802	1.0000	0.0656	0.0714
edu.	0.4933	0.1879	0.1253	0.0607	0.3453	0.0656	1.0000	0.5553
inc.	0.5834	0.3609	0.0998	0.0646	0.3323	0.0714	0.5553	1.0000
Baden-Württ.								
age	1.0000	0.0901	0.1523	0.1504	0.5628	0.1263	0.5005	0.5952
gender	0.0901	1.0000	0.0276	0.0299	0.1621	0.0390	0.1695	0.3391
ethn.	0.1523	0.0276	1.0000	0.0716	0.0864	0.0728	0.1765	0.1264
dojs	0.1504	0.0299	0.0716	1.0000	0.6192	0.6210	0.0616	0.0706
unempl.	0.5628	0.1621	0.0864	0.6192	1.0000	0.6688	0.3682	0.3686
reg. unem.	0.1263	0.0390	0.0728	0.6210	0.6688	1.0000	0.0539	0.0666
edu.	0.5005	0.1695	0.1765	0.0616	0.3682	0.0539	1.0000	0.5604
inc.	0.5952	0.3391	0.1264	0.0706	0.3686	0.0666	0.5604	1.0000

One can see that the characteristic concerning house size class four is sustainable. Besides it is obvious that the weight of house size class one is dominant. That result could be helpful in interpreting some simulation results following in Chapter 3.1.2. For a better feeling how the variables are correlated Table 3.4 gives a matrix of contingency coefficients between all considered variables in Saarland and Baden-Württemberg. There one can see that e.g. the coherence between the variable unemployed (unempl.) and registered

unemployed (reg. unem.) is extremely high. On the other side the correlation between duration of job-seeking (dojs) and gender or ethnicity (ethn.) is negligibly small. In addition the coefficients concerning income (inc.) and registered unemployed as well as unemployed look interesting. As the value with respect to unemployed tends to 0.3 the coefficient between income and registered unemployed is also negligibly small.

3.1.2 Selected Tasks of the Simulation Study

The simulation study on small area estimators in DACSEIS was performed in a varying general framework. The purpose of this approach was to gain information about the influence of the general conditions on the performance of the estimators considered. Therefore we will compare the outcomes of the simulation study with respect to the impact of the following factors of possible influence:

- the type of target variable,
- the set of ancillary variables,
- the regional differences between the federal states, and
- the classification of the small areas.

Of course this all leads to a comparative examination of the different estimators, their strenghts and failings.

The following comparative studies often refer to the estimations done with the data of Baden-Württemberg, mostly with the small area classification (iv) and the target variable unemployment. This restriction does not imply that all the conclusions made in the following only work with this combination of variables and settings. In general they hold for most or all federal states, variables or small area classifications. But the purpose of this section is to analyse the effects, a change in one part of the general settings has. For this, a kind of ceteris paribus analysis seems adequate which means referring to the same set of estimators again and again.

As already mentioned above, the target variables considered in this workpackage were unemployment, the household composition and the net income. For the estimation of unemployment, two different sets of auxiliary variables were used. For the simulation on the proportion of single households and for the net income only one set was used.

As the set of auxiliary information differs among the three target variables, it is difficult to separate their particular impact, so the analysis of these two influence factors will be done together. For a comparison of the findings concerning different target and auxiliary variables, the results of the estimations done with data of Baden-Württemberg will be used as showcase outcomes, but they also hold for other federal states.

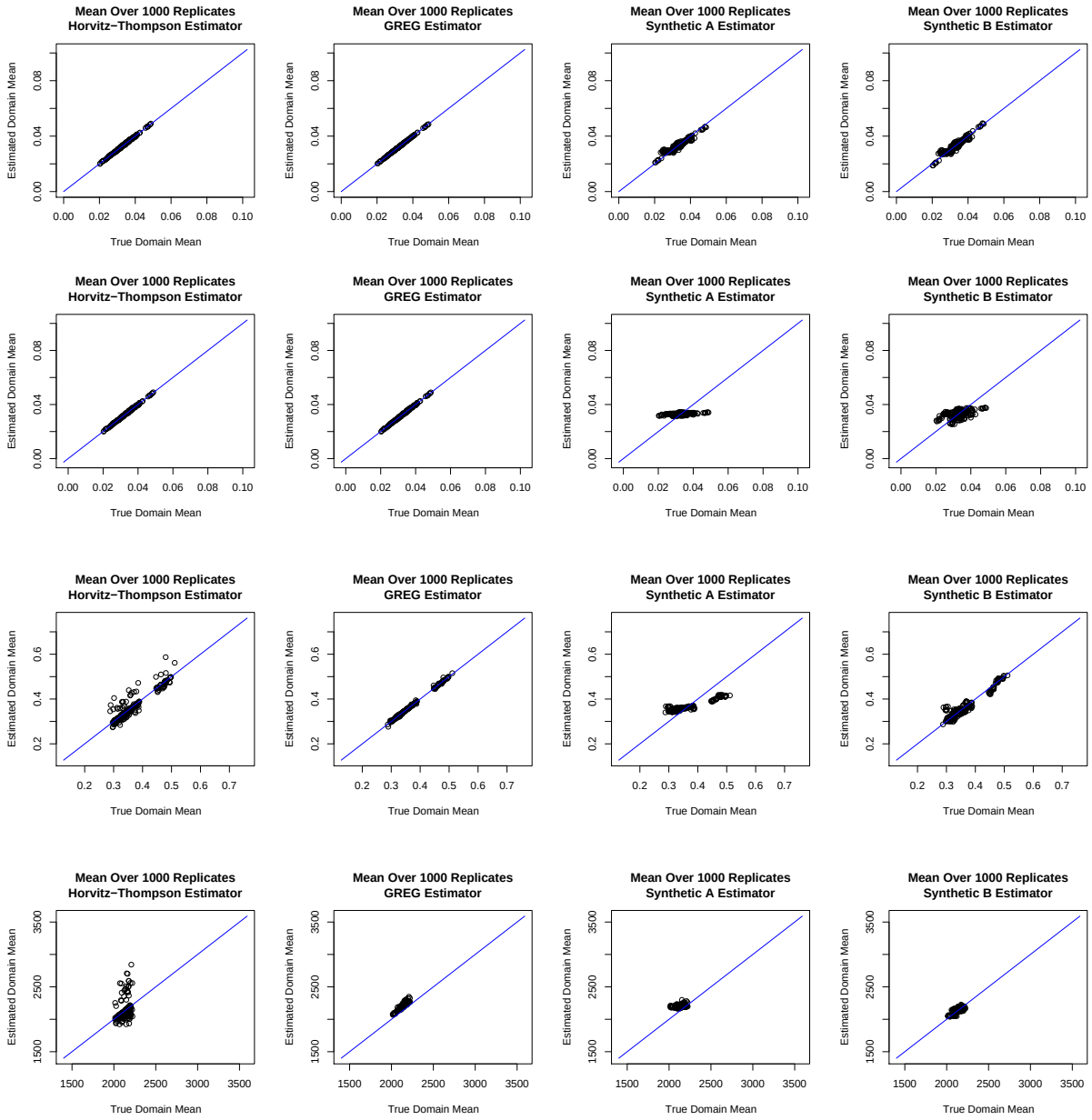

Figure 3.16: Direct and indirect estimators for all target variables in Baden-Württemberg and small area classification (iv).

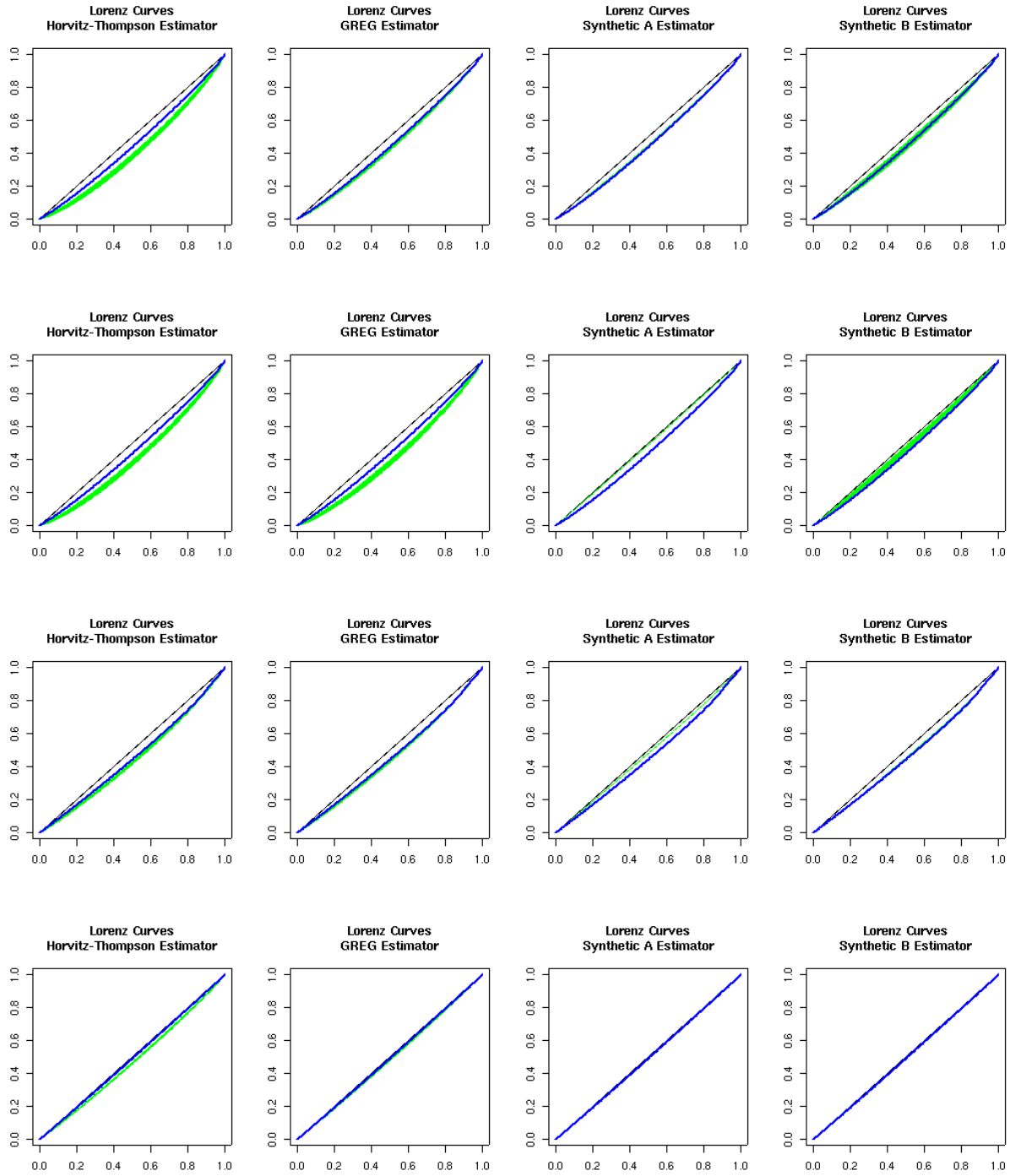

Figure 3.17: Direct and indirect estimators for all target variables in Baden-Württemberg and small area classification (iv).

Figures 3.16 and 3.17 illustrate the performance of a design-based, a model-assisted and two model-based estimators for all three target variables in small area classification (iv). In each graph the four rows stand for estimation of unemployment (row one), estimation of unemployment but with another set of covariates (row two), estimation of the household composition (row three) and estimation of income in the last row. In columns one can see in the first column on the left hand side the Horvitz-Thompson estimator, the second col-

umn presents the GREG estimator and the last two one illustrate the performance of the synthetic estimator A and B. One major difference in the target variables stems from the heterogeneity. Whereas the income is fairly homogenous, the household composition and the percentage of unemployed differ significantly between the small areas. This disparity can be seen in the Lorenz curves displaying the disparity of the true values as a blue line and the disparity of all estimators considered as a band of 1,000 green lines due to the 1,000 replications. As the income shows nearly perfect homogeneity, the Lorenz curve of the true values almost meets the bisector.

Having a closer look at Figures 3.16 and 3.17 one can see that the direct estimators, especially the Horvitz-Thompson estimator, overestimate the disparity in the variables whereas the synthetic estimators yield better results. This effect is increased by a rise in the dispersion of the true variable as can be seen comparing the plots for the different target variables. The synthetic estimators on the other side manage to reproduce the disparity given in the true values more exactly.

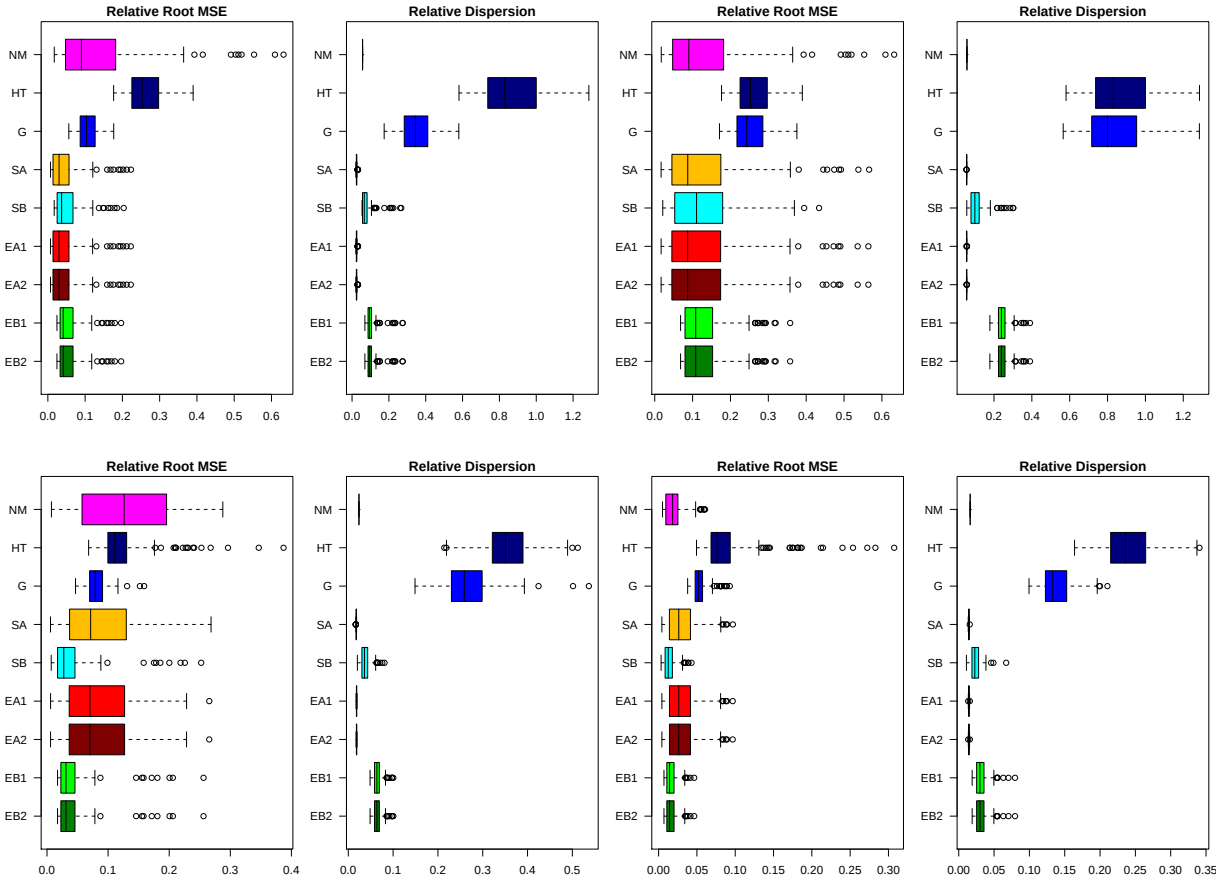

Figure 3.18: RRMSE and RD for the estimators for all target variables in Baden-Württemberg and small area classification (iv).

Another striking fact with respect to the different variables is the performance of the estimators measured with the relative root MSE (RRMSE) and the relative dispersion (RD). Especially the target variable net income, but also the variable household composition yield relatively low values of RRMSE and RD. This effect can be seen from Figure 3.18 containing the RRMSE and RD for all target variables in small area classification (iv) in

Baden-Württemberg. This result can be observed in all federal states. In fact, the effect increases the smaller the federal state is, at the same time showing more differences in level between the RRMSE and RD of the various estimators. For detailed figures on this topic see the electronic RPM (`www.rpm.dacseis.de`). The order in which the estimators are plotted is as follows: National Sample Mean, Horvitz-Thompson, GREG, Synthetic A, Synthetic B, EBLUP A1, EBLUP A2, EBLUP B1 and EBLUP B2. The difference between A1 and A2 as well as between B1 and B2 is due to a various variance estimation. In order to view the differences within the graphs unequal scalings between the graphs had to be used.

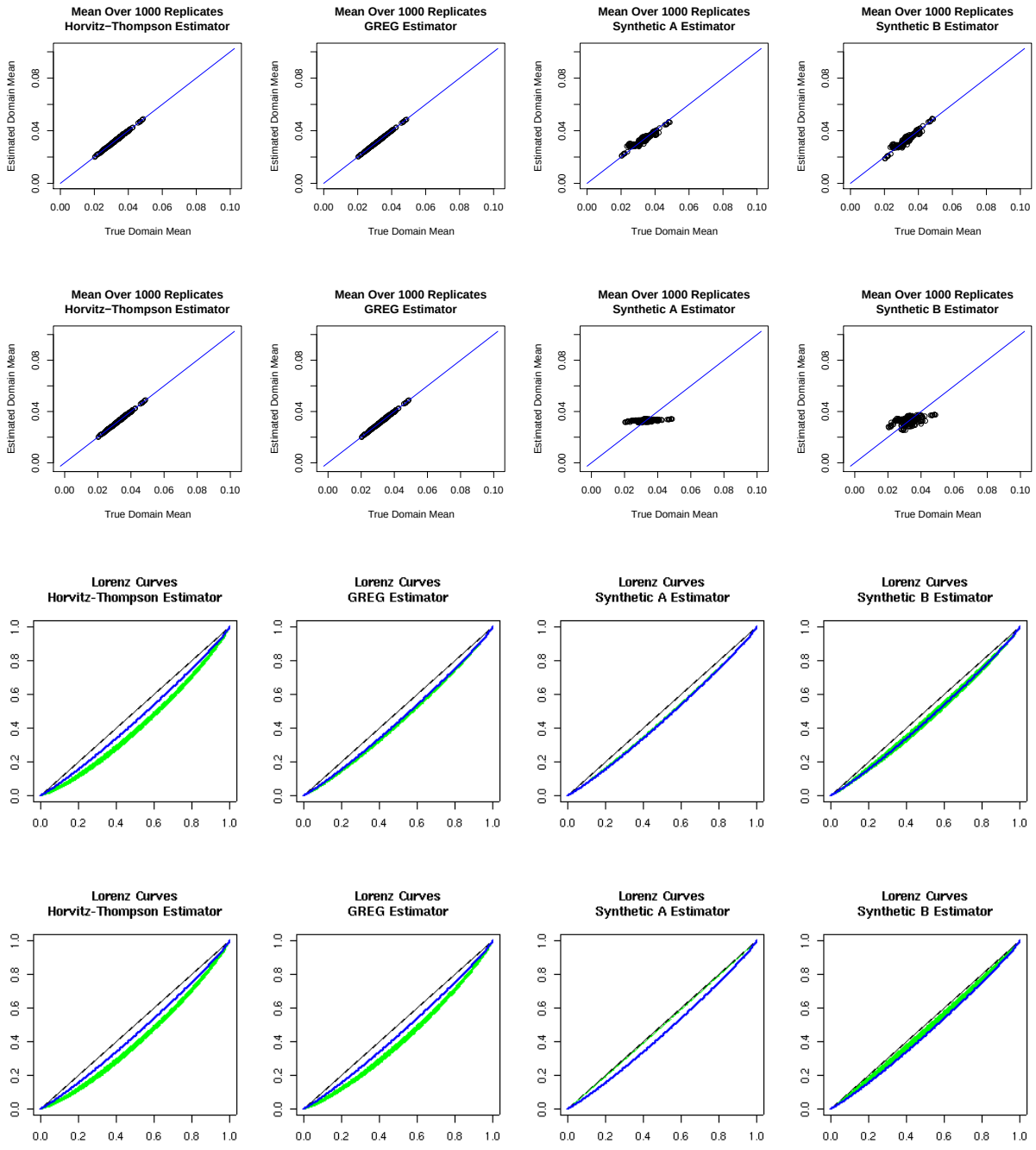

Figure 3.19: Direct and indirect estimators for the unemployed in Baden-Württemberg and small area classification (iv).

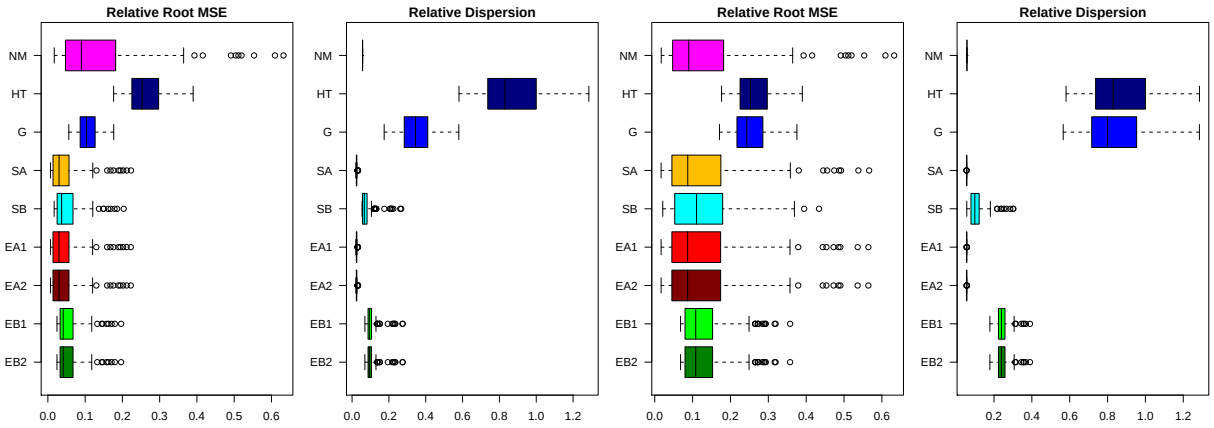

Figure 3.20: RRMSE and RD for small area estimators for the unemployed in Baden-Württemberg and small area classification (iv).

Figures 3.19 and 3.20 show the effect the choice of auxiliary information has on the performance of the small area estimators. Of course it has no effect at all on the Horvitz-Thompson estimator since this estimator is based only on information concerning the target variable. Looking at the performance of the GREG estimator the impact of the ancillary variables used shows best in the Lorenz curve. While the mean over 1,000 replicates suggests similar values of estimation no matter which variables are used, the Lorenz curves differ significantly. When using the auxiliary variable *status of registration at the employment office (ALO)*, the disparity of the true values is overestimated only to a very small extent. Using the second set consisting of a combination of age, sex and educational level leads to serious overestimation. These results were to be expected, as the first auxiliary variable is highly correlated with the target variable (see Table 3.4). Nevertheless the extent of the first variables supremacy is intriguing.

Concerning the synthetic estimators the predominance of the auxiliary variable *ALO* shows even more significantly. Both synthetic estimators, but especially the synthetic estimator A, produce adequate estimation for the unemployed when using this ancillary variable, but fail to account for the differences between the small areas when using the second set. As a result they underestimate the disparity as illustrated in the Lorenz curves. Similar to the discrepancies between the different target variables, this effect is increasing when smaller federal states are looked at. For illustration compare Figures 3.21 and 3.19. Figure 3.21 shows in line one the estimation of unemployment with registered unemployment as covariate and in line two the same target variable but with worse covariates. One possible explanation for this inability of the synthetic A estimator might lie in the fact, that its parameters are calculated with the overall sample therefore blurring the differences between the small areas.

A comparison of performance of all estimators for the two sets of additional information can be taken from Figure 3.20. Contrasting the RRMSE and the RD shows an evident improvement of precision when using the auxiliary variable *ALO*. Especially the effect on the GREG estimator is remarkable. As a conclusion it can be said that the quality of the small area estimators depend largely on the quality of the auxiliary information used.

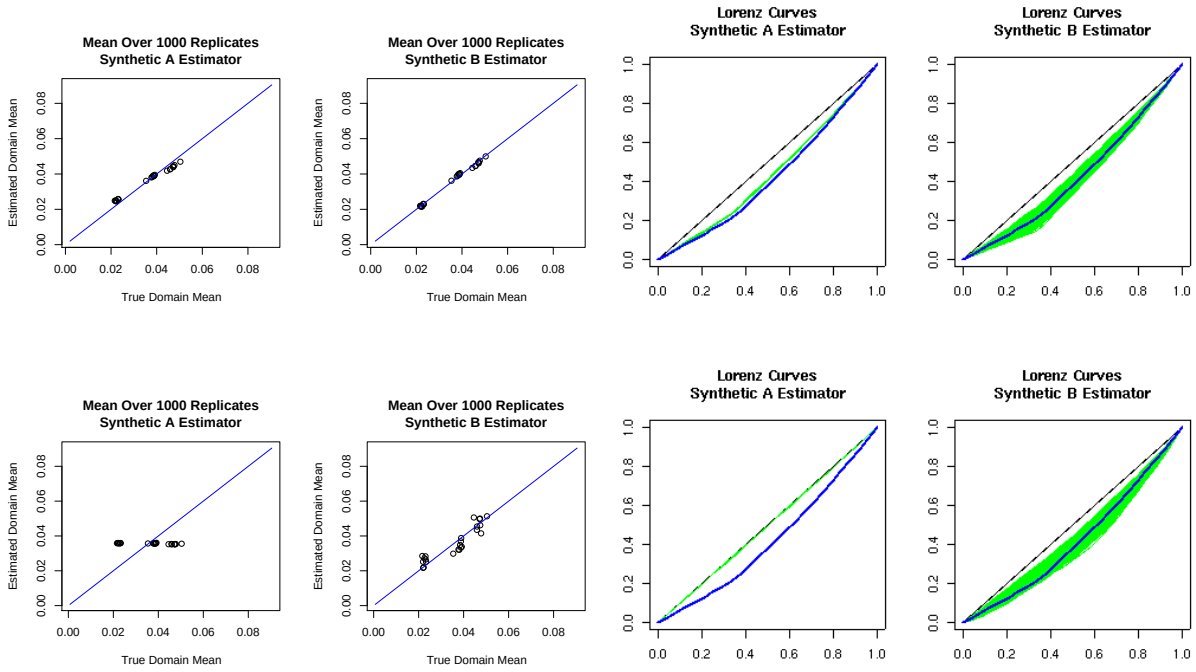

Figure 3.21: Synthetic estimators A and B for the unemployed in Saarland and small area classification (iv).

As already mentioned before the performance of the estimators considered here and the effects variations in the general framework have on it differ significantly with the federal states considered that is with the regional surroundings. The German federal states vary a lot in size and population composition and this differences also affect the estimators. For the simulation study a sensible mixture of bigger and smaller federal states as well as two city states was selected.

The performance of the standard estimators for these states can be seen in Figure 3.22. The order of graphs is as follows: the first seven graphs are the RRMSE and the second one shows the RD. In each block one can see at first the federal states Baden-Württemberg, Berlin, Hamburg and Mecklenburg-Vorpommern and below the states Sachsen, Saarland and Schleswig-Holstein. One can state that the RRMSE and the RD of all estimators assimilate when larger federal states are considered. For the two city states Hamburg and Berlin, evident differences between the estimators are visible. This is apparent when comparing the figures for those two city states with the figures for bigger states, say Baden-Württemberg or the Saarland. One striking fact is the different level of the RRMSE for the two synthetic estimators and the EBLUP estimators derived from those, a difference which diminishes visibly when turning to a bigger federal state. On the other hand the general level of the RRMSE rises when bigger states are regarded. Reacting a little bit strange in this context is the federal state Mecklenburg-Vorpommern. Although a fairly big state, it shows the features of the city states Berlin and Hamburg. A possible explanation for this phenomenon might be the uncommon composition of population in this state which is suffering strongly from movement of labour.

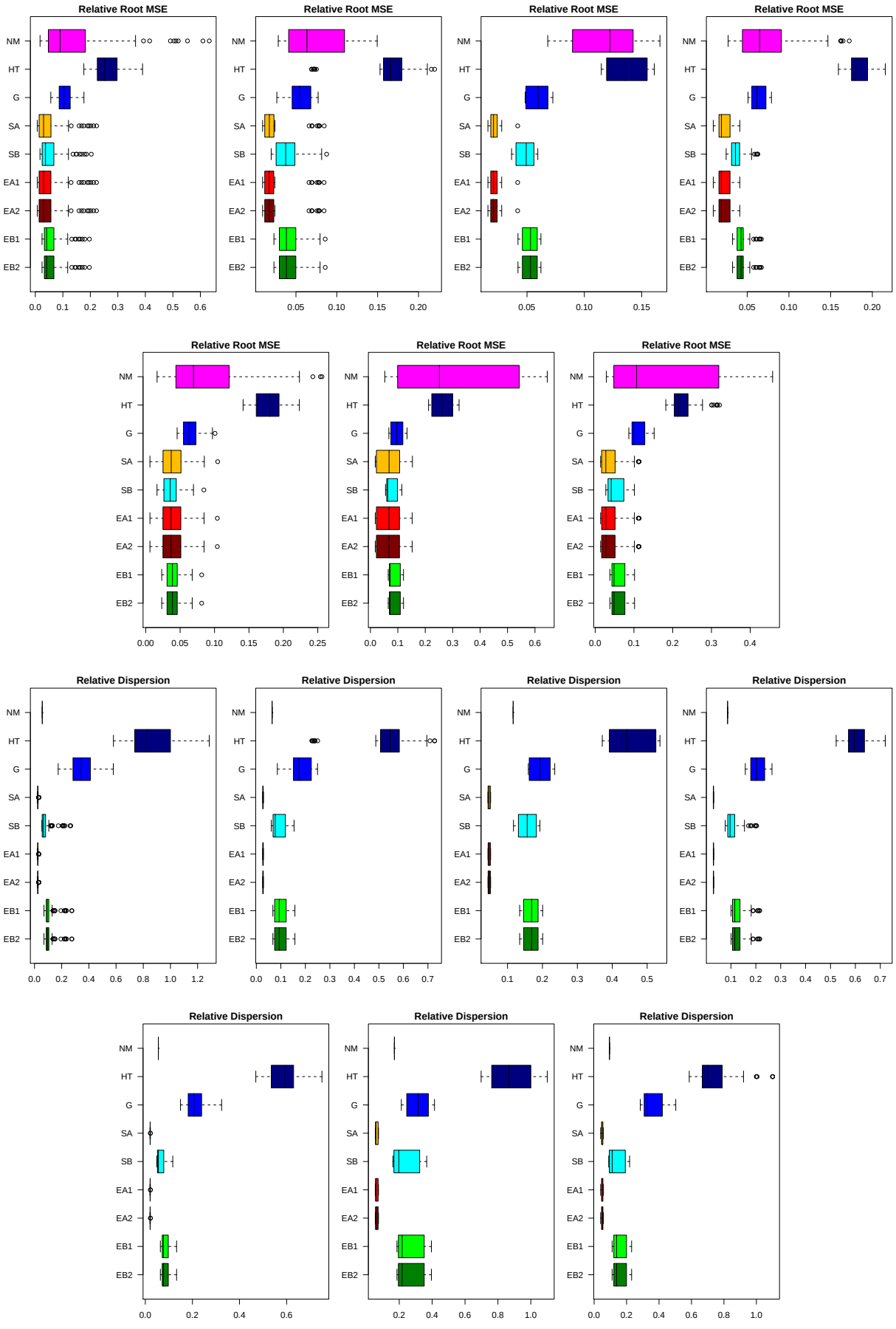

Figure 3.22: RRMSE and RD for small area estimators for the unemployed in small area classification (iv).

A very strong effect on the small area estimators emanates from the structure of the underlying small areas or domains. The performance of the estimators differs a lot depending on the homogeneity or heterogeneity of the small areas. In the simulation study four different classifications of small areas were used displaying different levels of heterogeneity. Classifications (ii) and (iii) show more disparity as they depend on the house size class and of course the fact whether a person lives in a one-family house or in a large apartment building interacts with the target variables examined here, most significantly with the employment status. The effect this has on the disparity of the true values can be seen looking at the figures illustrating the universes in the previous section and at Figure 3.23. The disparity is relatively small for small area classifications (i) and (iv) and high for those classifications including the house size class. In columns one can see the small area categorisations from regional classes in the left column to Nuts5 regions in the column on the right hand side. In rows the Figure 3.23 shows at first an overview to all estimators and in row two to five from the top to the bottom the estimators Horvitz-Thompson, GREG, Synthetic A and B.

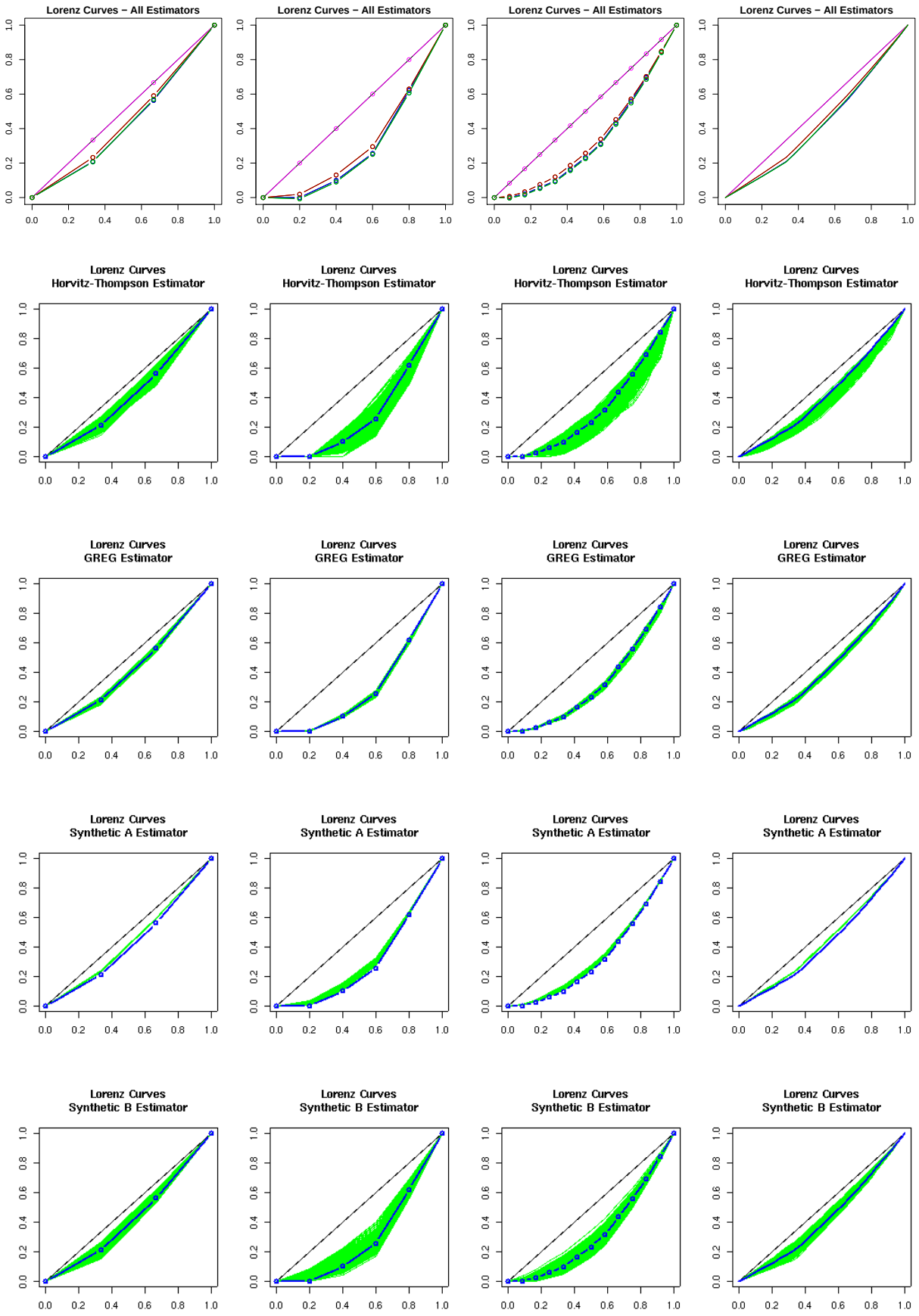

Figure 3.23: Lorenz curves for the unemployed in Saarland.

The influence the different level of heterogeneity have on the ability of the estimators to reproduce this heterogeneity is not too powerful. Viewing Figure 3.24 one can see that e.g. the general tendency in overestimating or underestimating the true disparity is not reversed, only the dimension varies among the various small areas. This result holds for the direct estimators as well for the synthetic estimators. The order of graphs is equivalent to Figure 3.23.

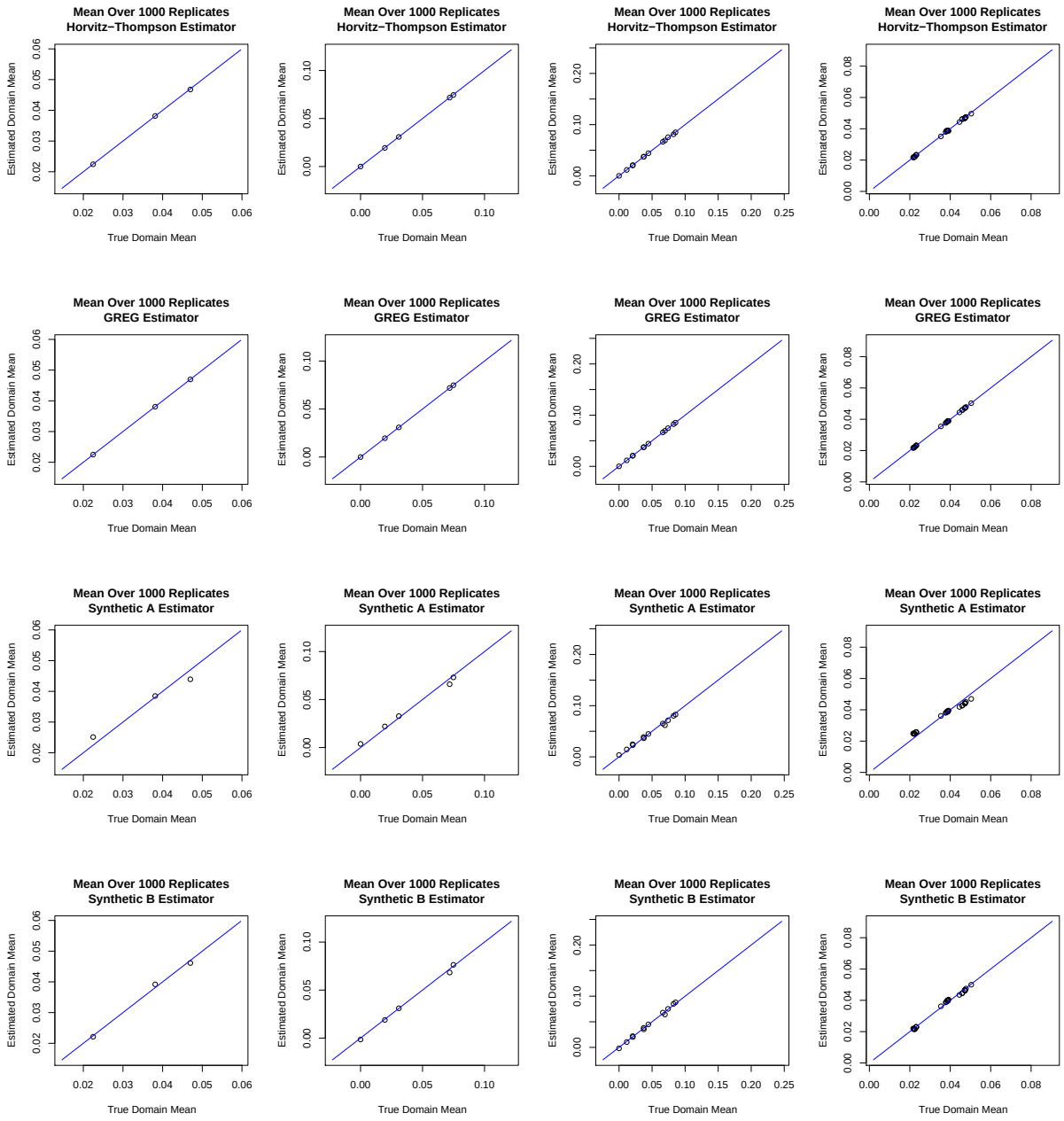

Figure 3.24: Direct and synthetic estimators for the unemployed for all small area classifications in Saarland.

On the other hand, viewing Figure 3.25 it can be seen that the performance of the direct estimators visibly depend on the homogeneity of the small areas, whereas the indirect estimators deal fairly well with heterogenous small areas. Concerning the order of graphs:

in columns one can see the small area classifications (cf. Figure 3.24) and in rows there are first boxplots in respect to the RRMSE und second graphs to the RD.

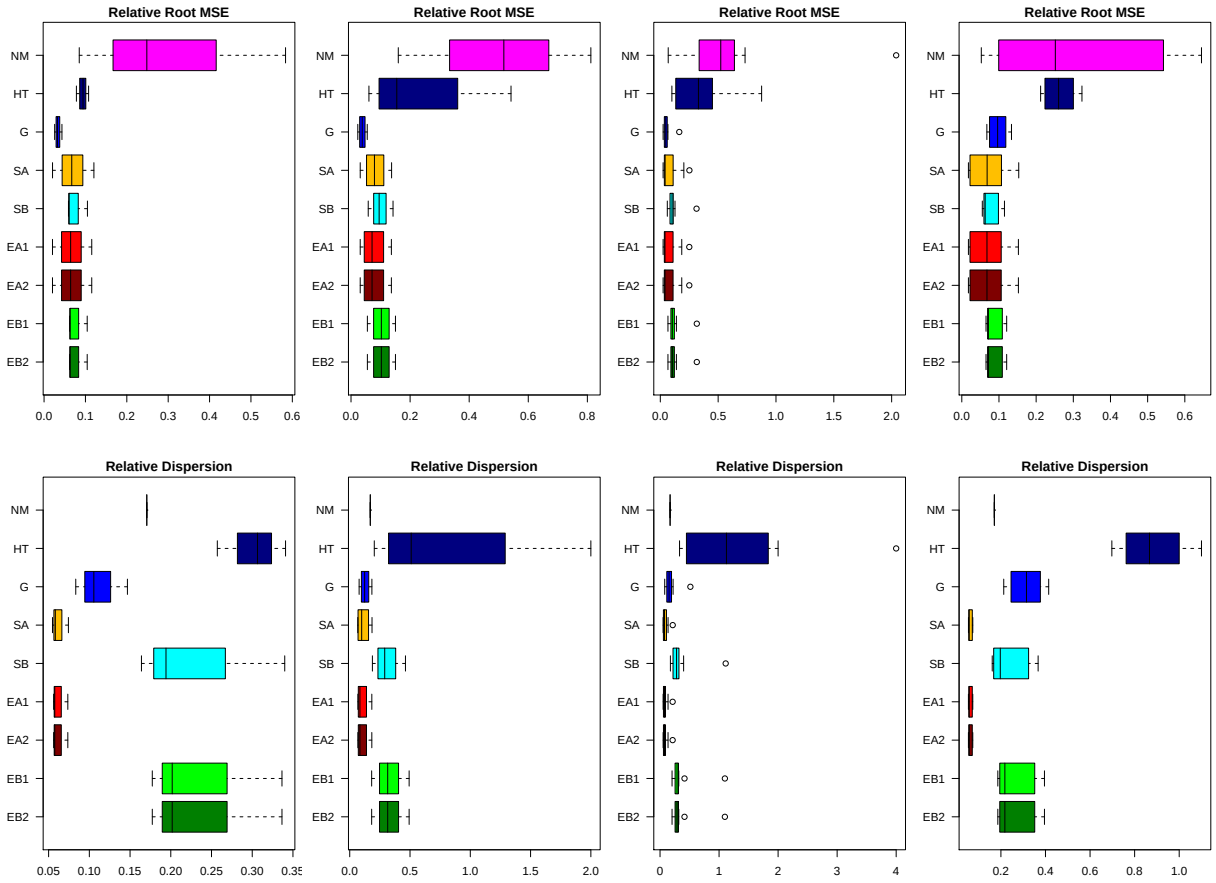

Figure 3.25: RRMSE and RD of estimators for the unemployed in Saarland for all small area classifications.

As already mentioned in the first section, the simulation study on small area estimators in the context of DACSEIS was done on a basis of 1,000 replications for each set-up. Of course, for measuring the performance of the estimators it is also interesting to analyse how they work on one replication only. The differences that can occur here become evident when looking at Figure 3.26 that contains information about the estimation of the unemployed in Baden-Württemberg using the second set of auxiliary variables. The order of graphs is as follows: the first nine graphs concern to a single sample and the second nine refer to the mean over all 1,000 samples. The estimators have the same order as in the RRMSE graphs above. Regarding the outcomes for one randomly chosen sample, one can see that when using this suboptimal set of auxiliaries, all estimators, the direct and the indirect, fail to reproduce the true value structure correctly. But whereas the Horvitz-Thompson estimator and the GREG estimator manage to overcome this problem and produce unbiased estimates when the mean of all 1,000 replications is considered, the indirect estimators, especially the synthetic estimators, still yield estimates that are heavily biased. The synthetic estimator A transmits this problem to the EBLUP A estimator of which it is a part, despite of the influence of the approximately unbiased GREG estimator. The EBLUP B however, consisting of the synthetic estimator B and the Horvitz-Thompson estimator, performs better, due to a greater impact of the unbiased direct estimator.

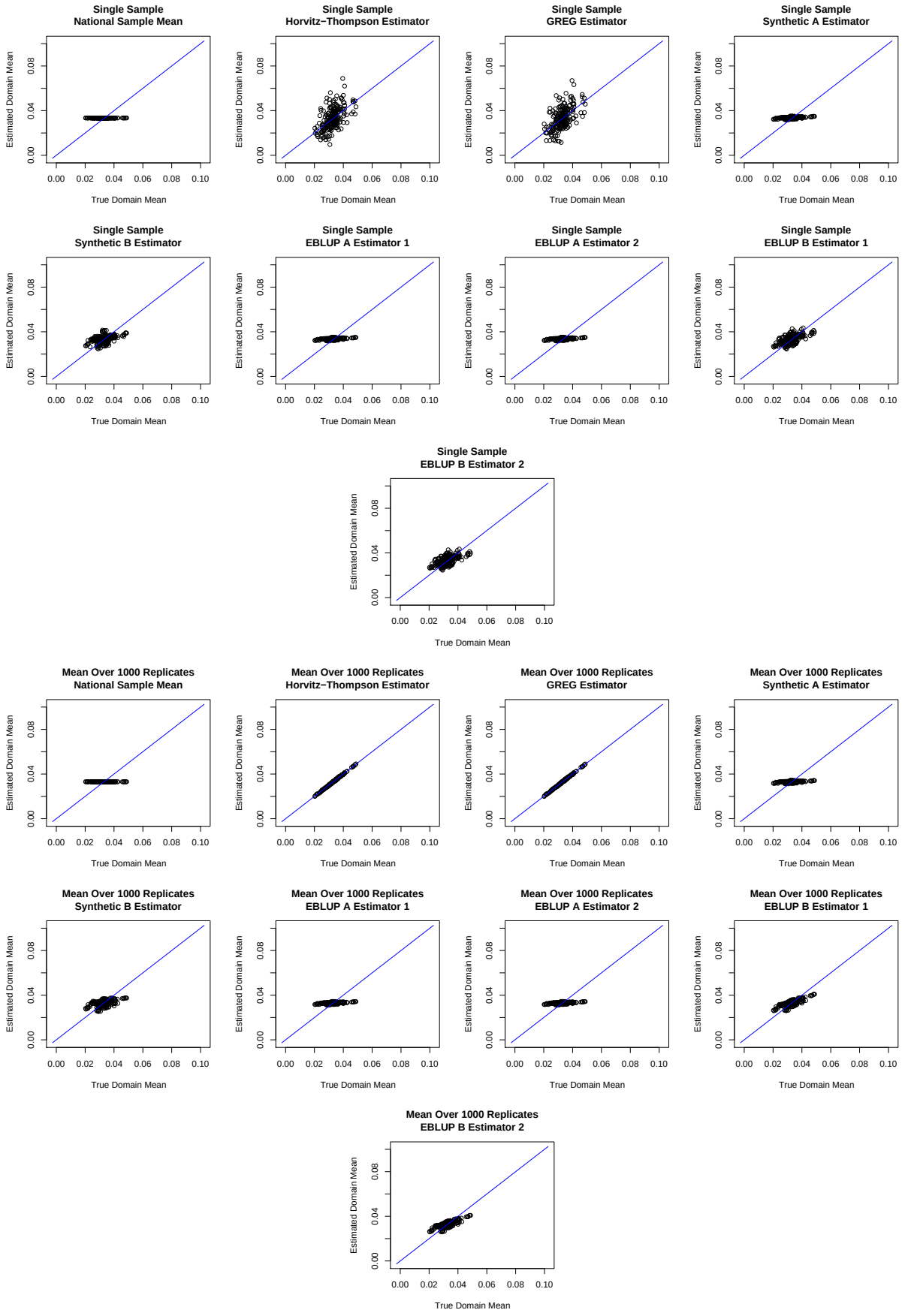

Figure 3.26: Performance of estimators in one randomly chosen sample and in the mean over all samples for unemployed in Baden-Württemberg and small area classification (iv).

Both EBLUP estimators are a combination of an unbiased direct estimator and a biased synthetic estimator having the advantage of showing less variability. Therefore they balance the instability of the direct and the bias of the indirect estimator. The observation made above that in the EBLUP B estimator the impact of the unbiased Horvitz-Thompson estimator is relatively strong in contrast to the EBLUP A holds for most combinations of variables and framework. As a result, the EBLUP B estimator turns out better in terms of RRMSE, as can be seen from Figures 3.27 and 3.28. Figure 3.27 shows the estimator in the following order: HT, GREG, Synthetic A, Synthetic B, EBLUP A1, EBLUP B1.

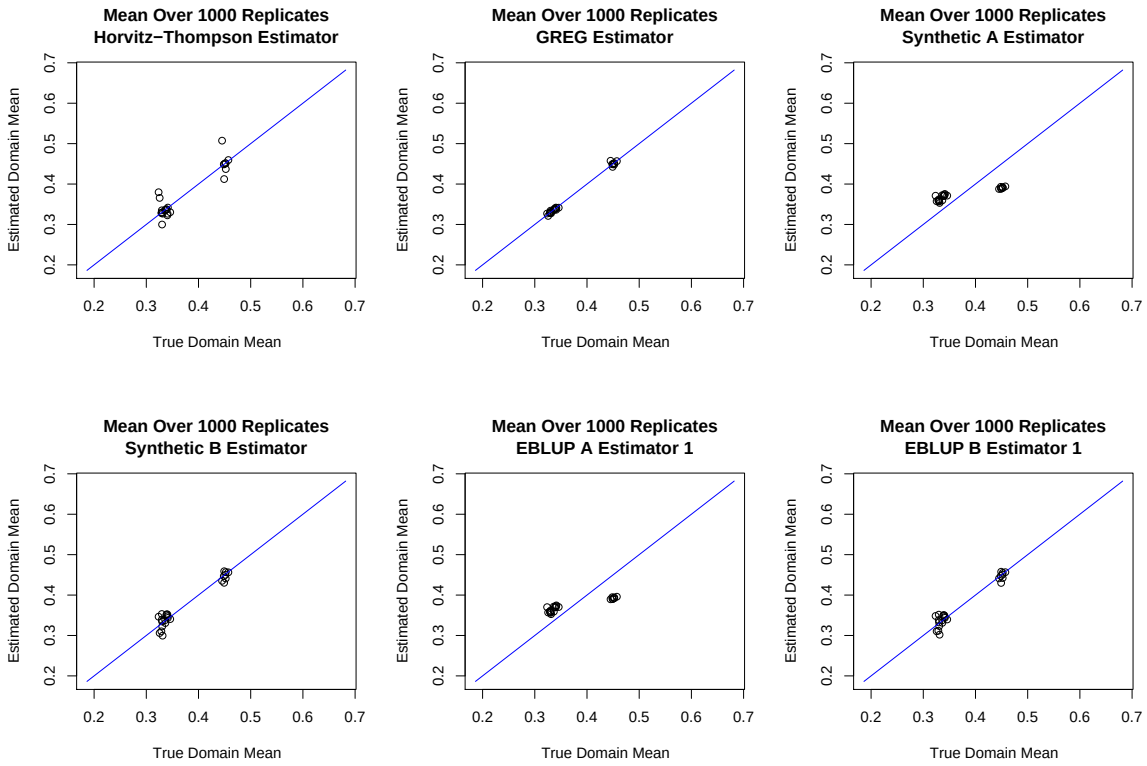

Figure 3.27: Estimators for household composition in Saarland for small area classification (iv).

In general, it can be said that the performance of the EBLUP estimators used in the simulations of DACSEIS largely follows the performance of the respective synthetic estimators. In situations when the synthetic estimator A performs well, this also holds for the EBLUP A estimator and vice versa. The performance of the synthetic estimates however are largely dependent on the auxiliary variables. When a *good* set of ancillary information is used, the synthetic estimator A outperforms the synthetic estimator B and vice versa.

To conclude this section we would like to have a look at the performance of the confidence intervals for the different estimators. The figures below show the empirical coverage rates of the estimators considered in the DACSEIS simulations. By empirical coverage rate we understand the proportion of the 1,000 replications that the intervals included the true value. The simulations were done for face values of 90% and 95%. Looking at Figure 3.29 (row one: regional class and Nuts5 regions concerning unemployment with the first covariate, row two, three, and four, the other target variables) one can see that

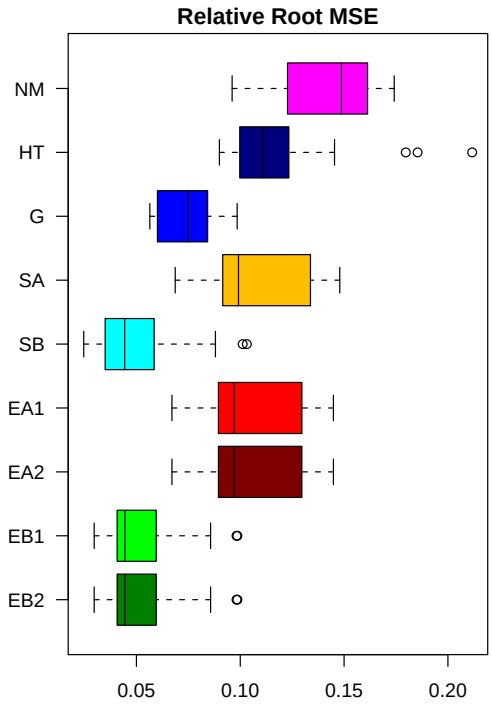

Figure 3.28: RRMSE for estimators for household composition in Saarland for small area classification (iv).

the proportion often does not meet the face value, especially when the small areas have small sample size like given with the classification (iv). This problem might be due to the fact that the variance of the variance estimator was not considered in constructing the confidence intervals.

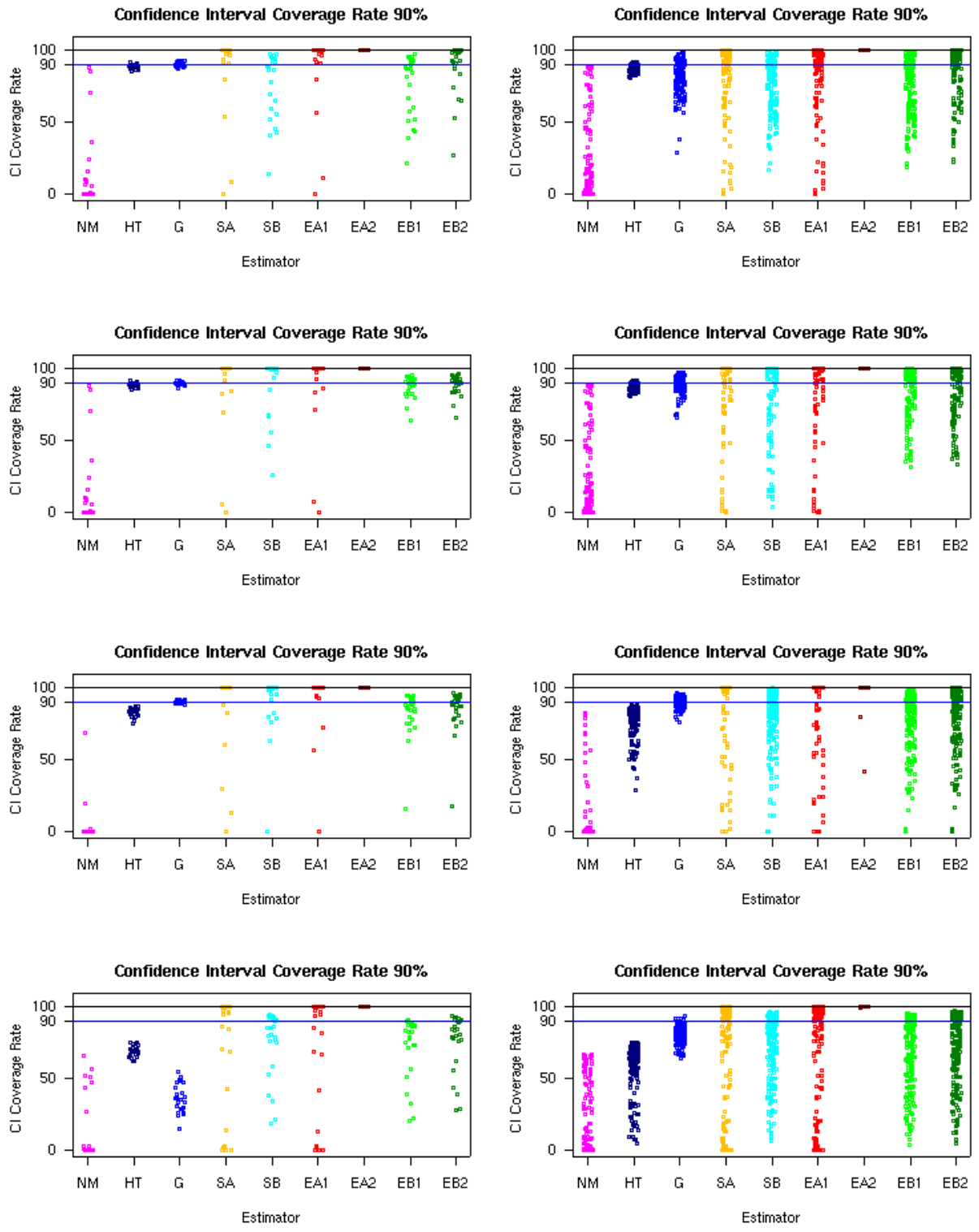

Figure 3.29: 90%-confidence intervals for small area classifications (i) and (iv) for Baden-Württemberg.

From comparing Figure 3.29 with Figure 3.30 the conclusion can be drawn that the empirical confidence rate also depends on the small area classification used. While the classifications (i) and (iv), shown in Figure 3.29, yield relatively high coverage rates, these

rates decrease when calculating the intervals for the more heterogeneous small areas given in classification (iii) as a result of the change in the estimators´ performance.

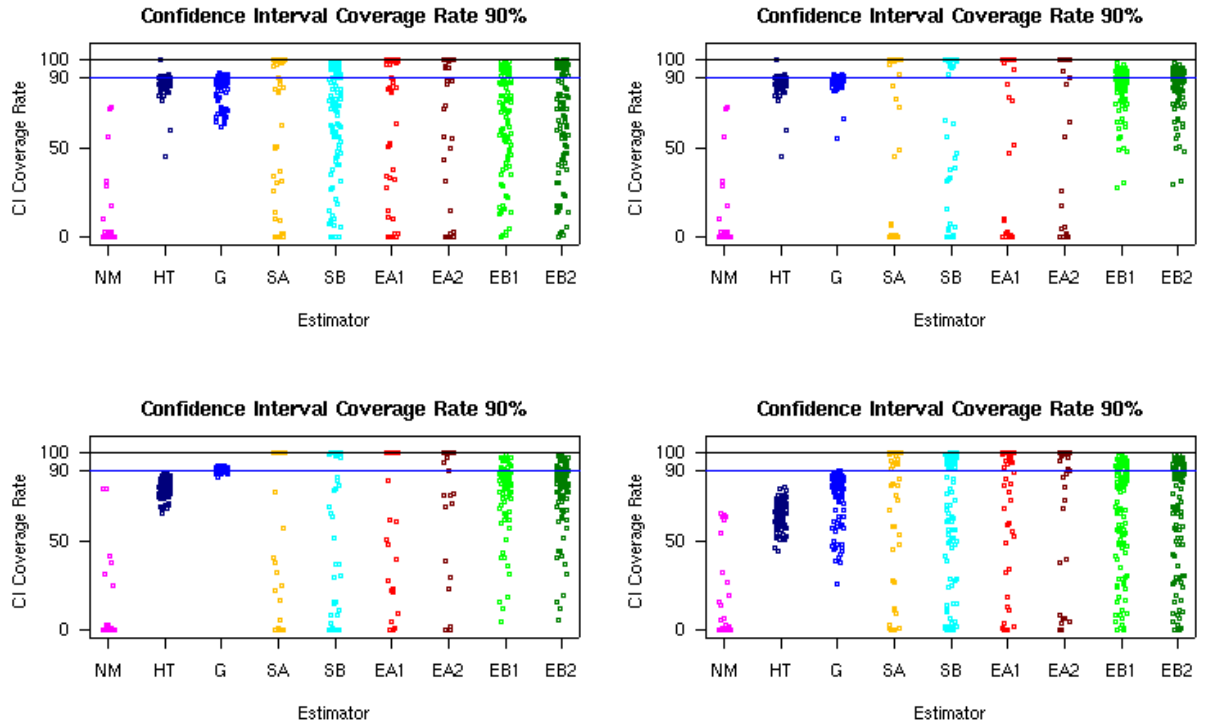

Figure 3.30: 90%-confidence intervals for small area classification (iii) for Baden-Württemberg.

3.1.3 The Influence of Non-response and Imputation on the Small Area Estimates

According to the remits of DACSEIS the simulations presented in the previous section were supplemented with another simulation study considering non-response that will be presented now. Purpose of this study was to examine the accuracy of the methods applied to small area estimation problems under non-response, given different levels of non-response. Two methods for correcting for non-response were applied, multiple imputation on the one hand and a calibration estimator with non-response correction on the other hand.

The first method of correcting for non-response in this workpackage was using a GREG estimator with design-weights correcting for the actual non-response. For more details on this estimator see WP 8.1. The second method used here was multiple imputation with logit imputation. Details on that can be taken from deliverable D11.2. As a modified multiple imputation routine, a specialized linear regression imputation for binary data was taken into account but will not be looked into in detail here. For further information on that way of correcting for non-response, also see Münnich and Rässler (2004). The full simulation results can be drawn from deliverable D12.2.

The analysis of the influence of non-response is not based on all set-ups used for small area estimation without non-response. As the effects of the non-response were to be

examined and not to be blurred by other effects like regional and structural differences between German federal states, the simulation was only performed on the dataset for the Saarland. The target variable in this part of the study was the proportion of unemployed like defined in subsection 3.1.1. As auxiliary data and for the non-response correction the variable *status of registration at the employment office (ALO)* was used. The classification of small areas were made as in the simulation study without non-response.

The Saarland only consists of three regional strata, so using the regional classes as variable yields few but relatively big small areas. Therefore, the small area estimators do not always work optimally. For a comparison of the point estimators and their variance estimators see Figures 3.31 to 3.40. It shows that while the GREG estimator with non-response correction performs well on all regional classes, even with respect to its variance, this does not hold for the estimators with multiple imputation correction. Comparing all regional strata and estimators, the Horvitz-Thompson at least manages to produce suitable point estimates for all regional classes, even if its variance estimator often fails to produce a value corresponding to the true one marked by a blue line. The GREG estimator also yields acceptable results, but the model-based estimators only perform well in regional class 1, but not in classes 2 and 3. As those two classes show common features, only the second is illustrated here.

The following graphs use the abbreviations RS1 and RS2 for the regional class number, GGK1 for the house size class, and RSxGGK1 for the classification regional stratum cross house size class number 1. Further NUTS51 and NUTS514 denote the NUTS5 classification with small area 1 or 14 respectively.

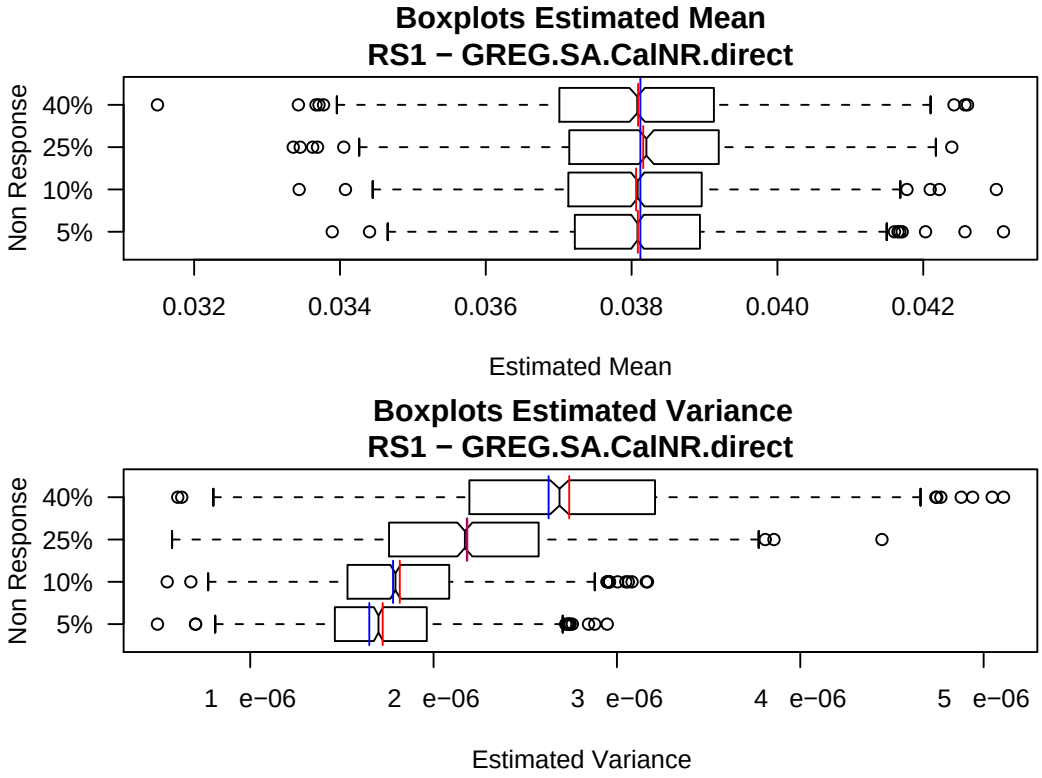

Figure 3.31: Point estimator GREG with non-response correction and its variance estimator for regional class 1.

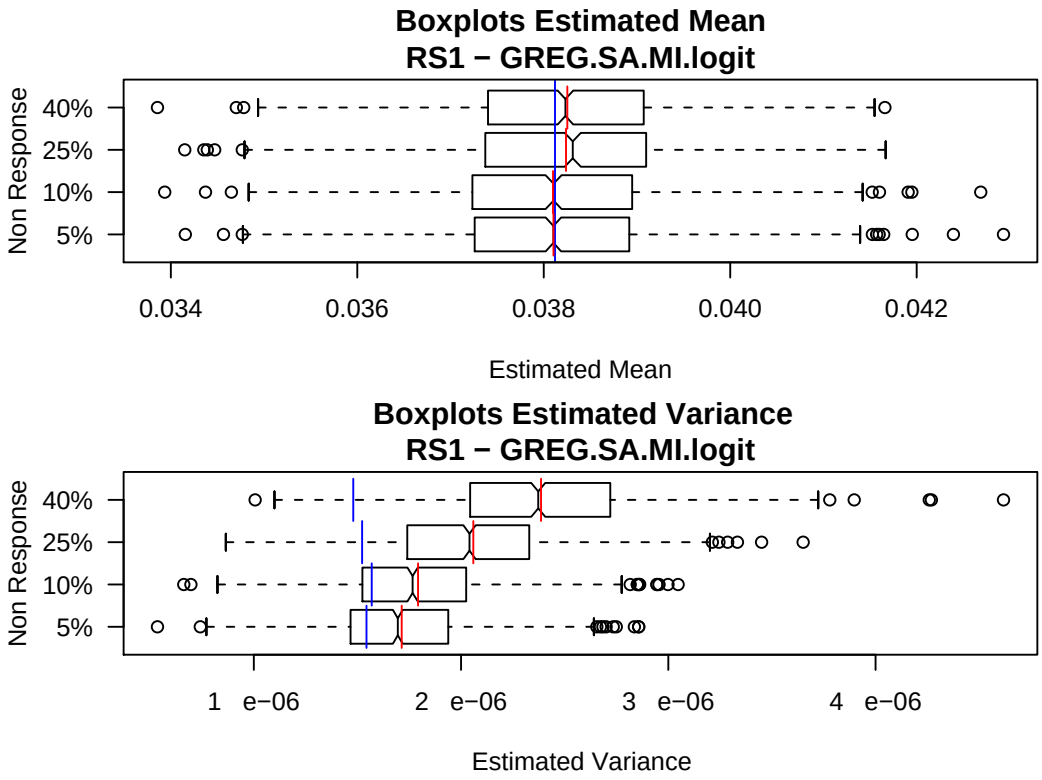

Figure 3.32: Point estimator GREG with MI and its variance estimator for regional class 1.

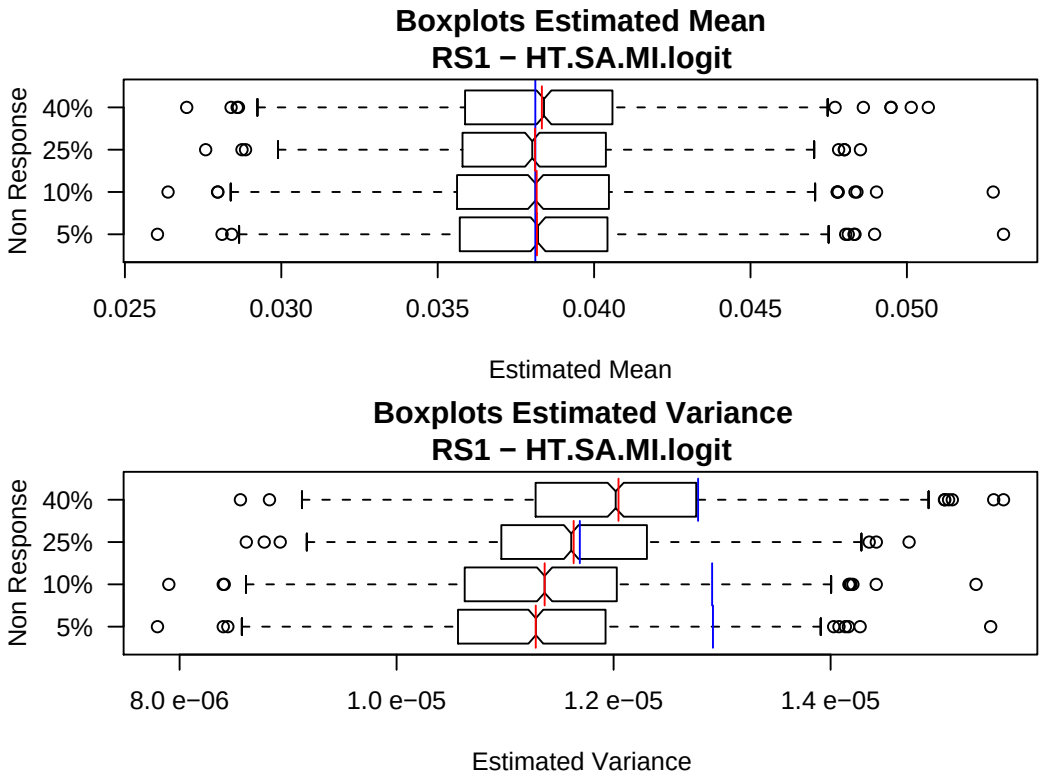

Figure 3.33: Point estimator HT with MI and its variance estimator for regional class 1.

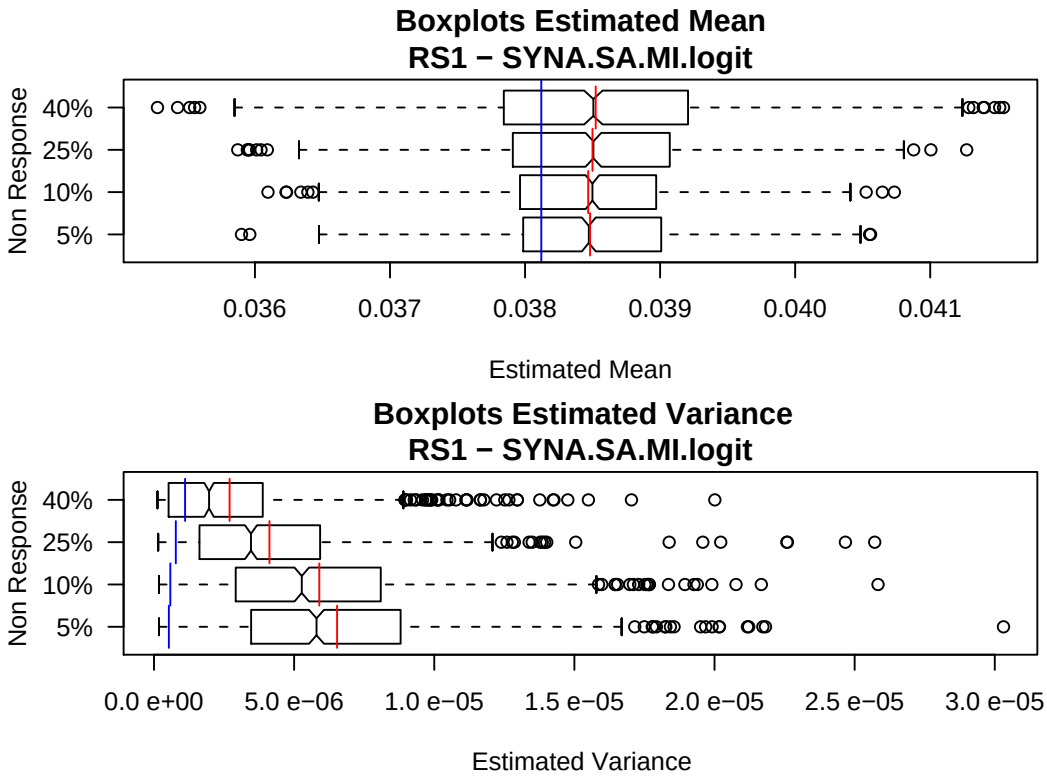

Figure 3.34: Point estimator Synthetic A with MI and its variance estimator for regional class 1.

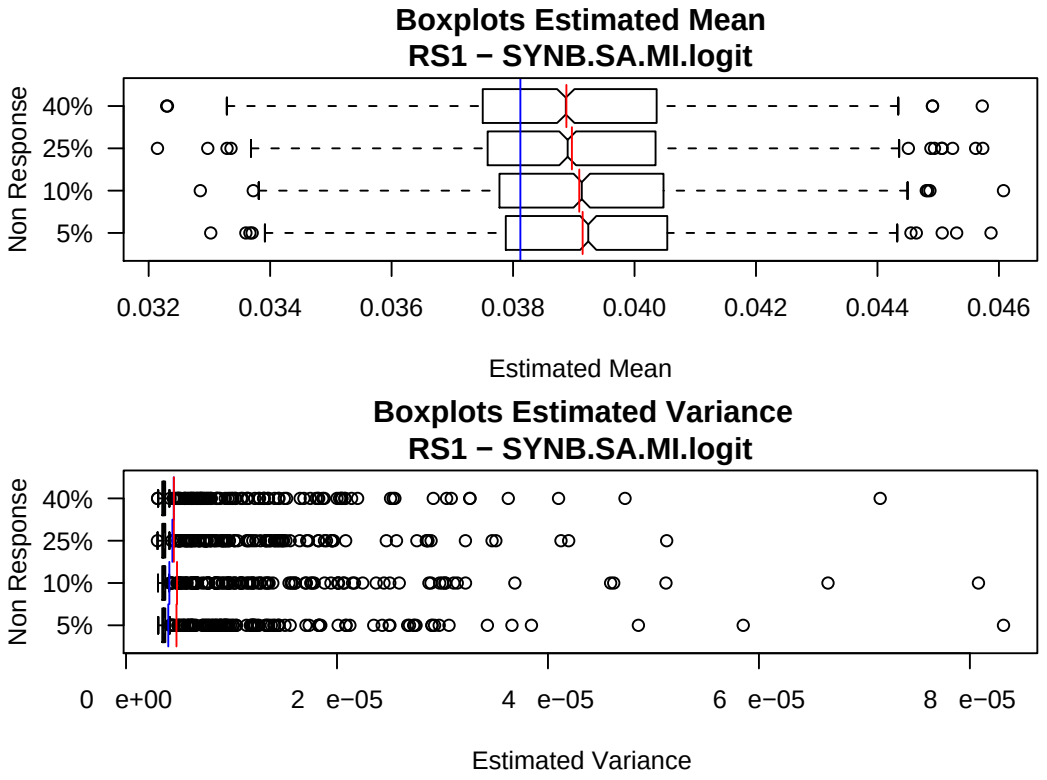

Figure 3.35: Point estimator Synthetic B with MI and its variance estimator for regional class 1.

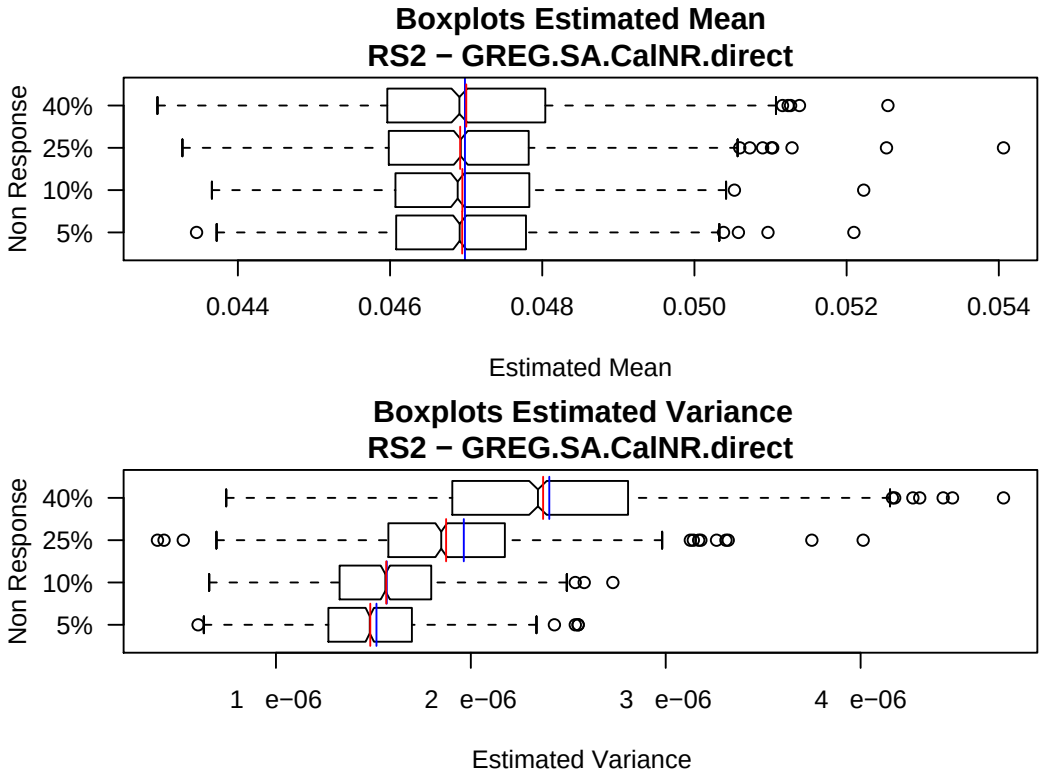

Figure 3.36: Point estimator GREG with non-response correction and its variance estimator for regional class 2.

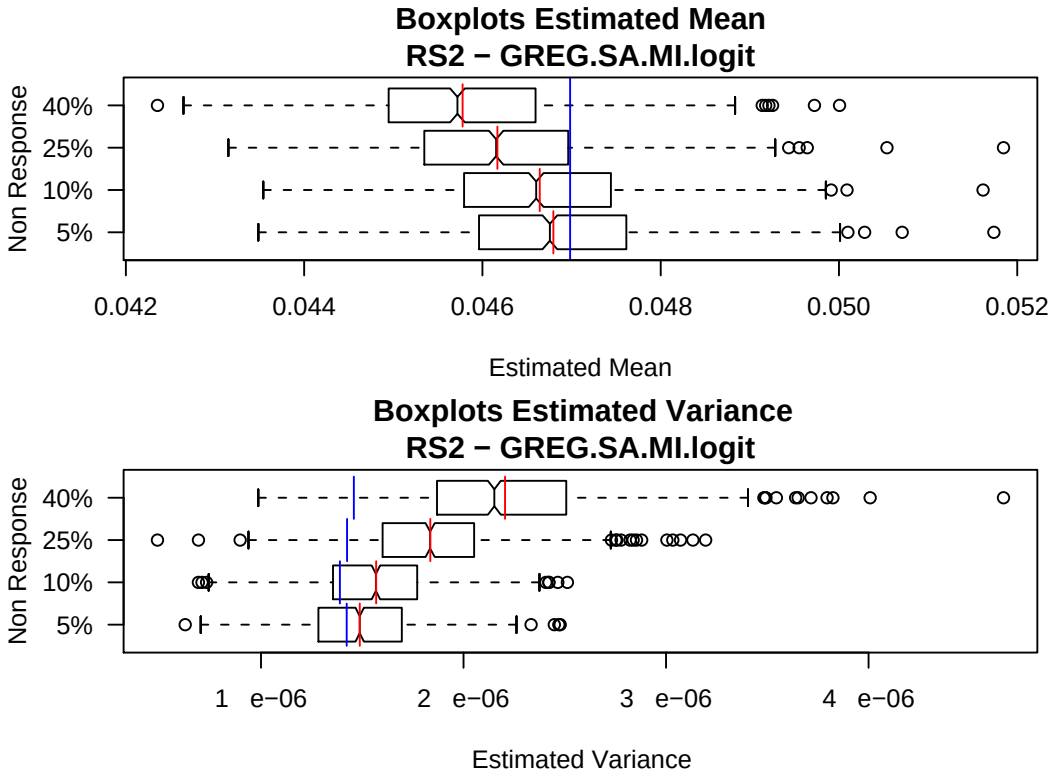

Figure 3.37: Point estimator GREG with MI and its variance estimator for regional class 2.

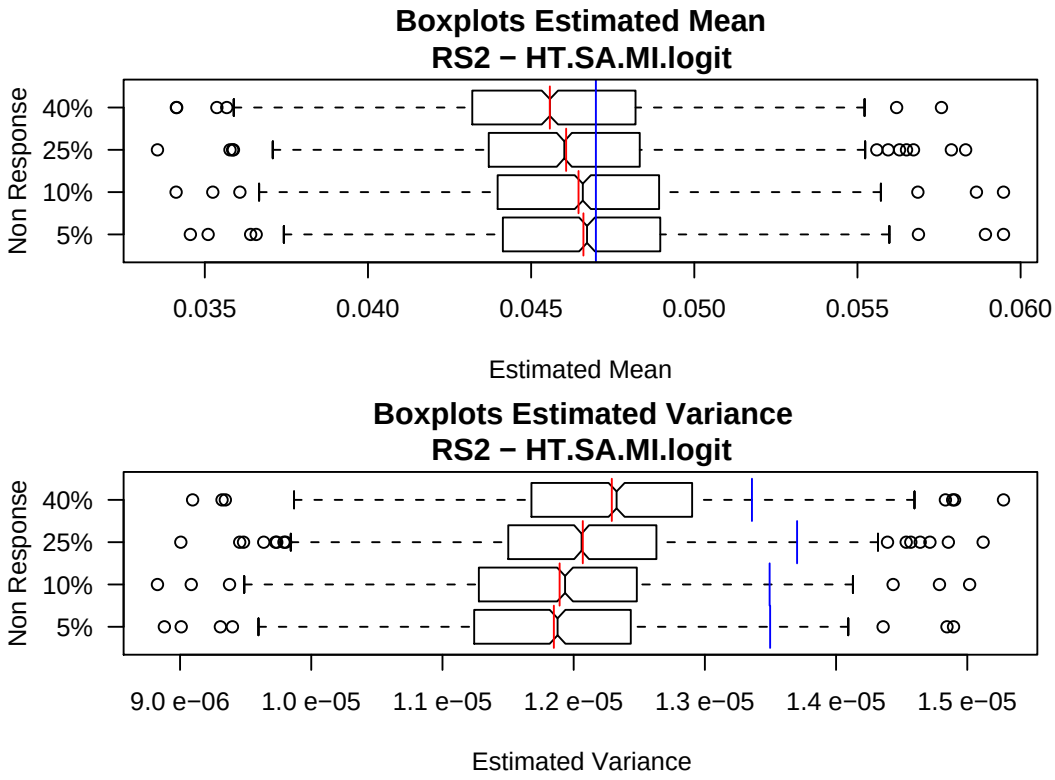

Figure 3.38: Point estimator HT with MI and its variance estimator for regional class 2.

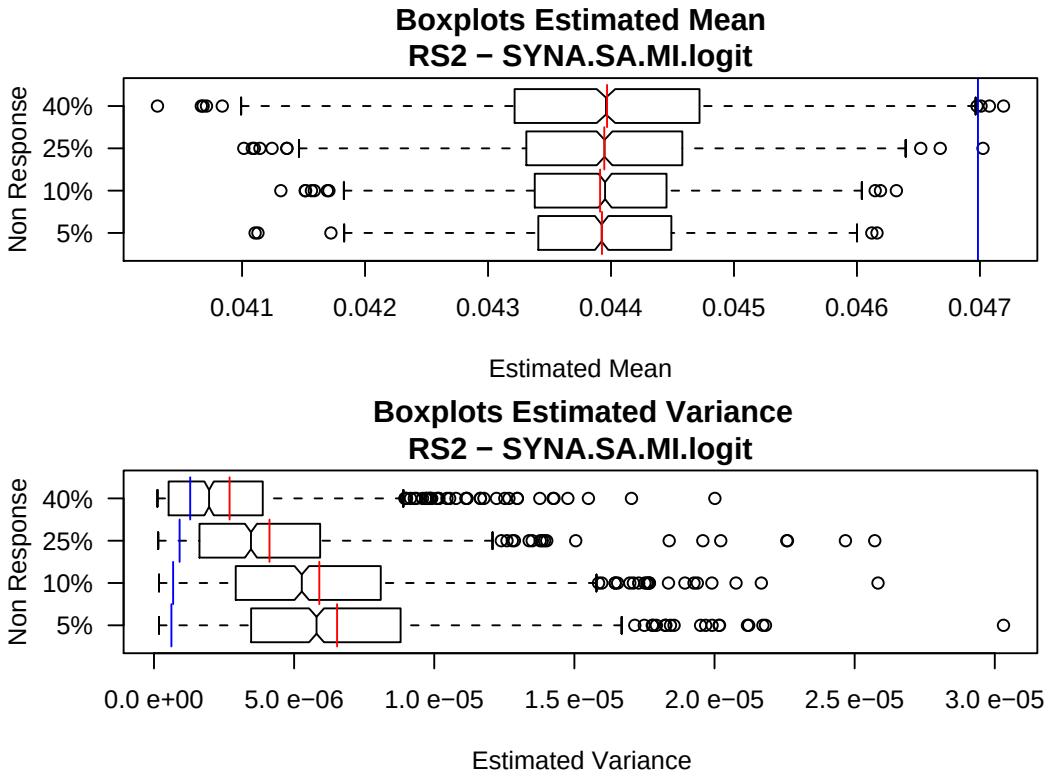

Figure 3.39: Point estimator Synthetic A with MI and its variance estimator for regional class 2.

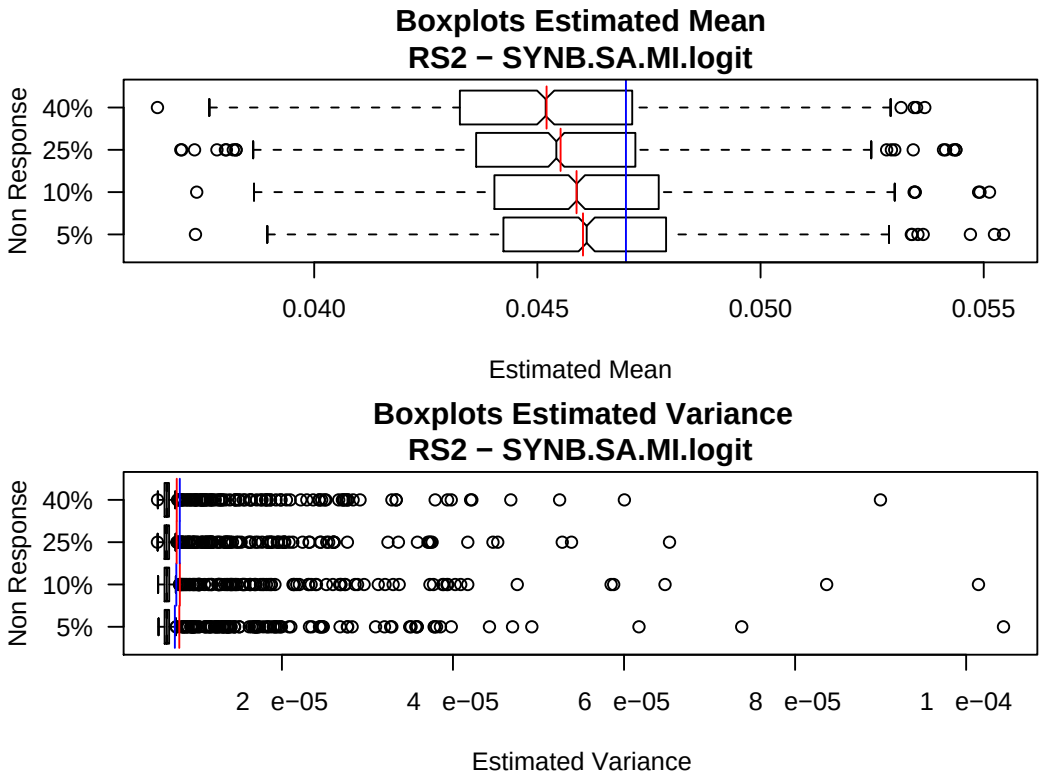

Figure 3.40: Point estimator Synthetic B with MI and its variance estimator for regional class 2.

Using the second small area classification – done by house size classes – leads to the problem of again only few, but additionally extremely heterogenous small areas. Especially the estimators of the fourth house size class, institutions, can hardly be interpreted. Looking at the performance of the estimators, e.g. for house size class 1 presented in Figures 3.41 to 3.45, the GREG estimator with non-response correction again shows nice performance, as do the GREG estimator with multiple imputation, the synthetic B estimator and the therefore the EBLUP B as well. However, the GREG estimators outperforms the model-based estimators with respect to the estimation of the variance.

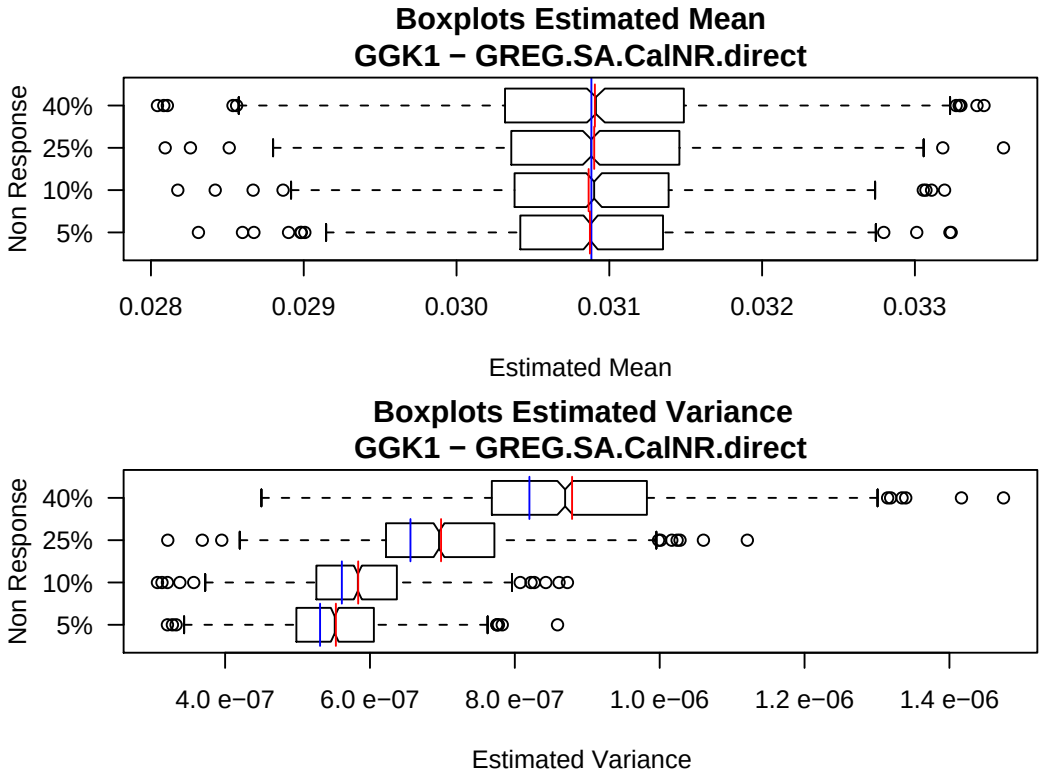

Figure 3.41: Point estimator GREG with non-response correction and its variance estimator for house size class 1.

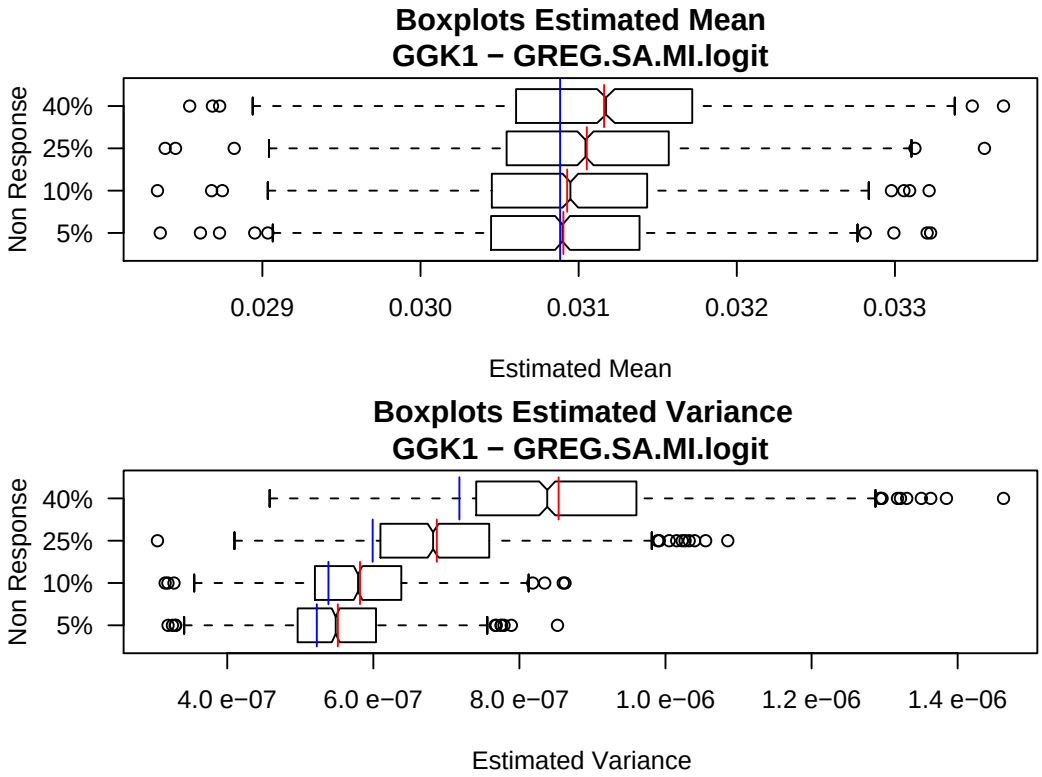

Figure 3.42: Point estimator GREG with MI and its variance estimator for house size class 1.

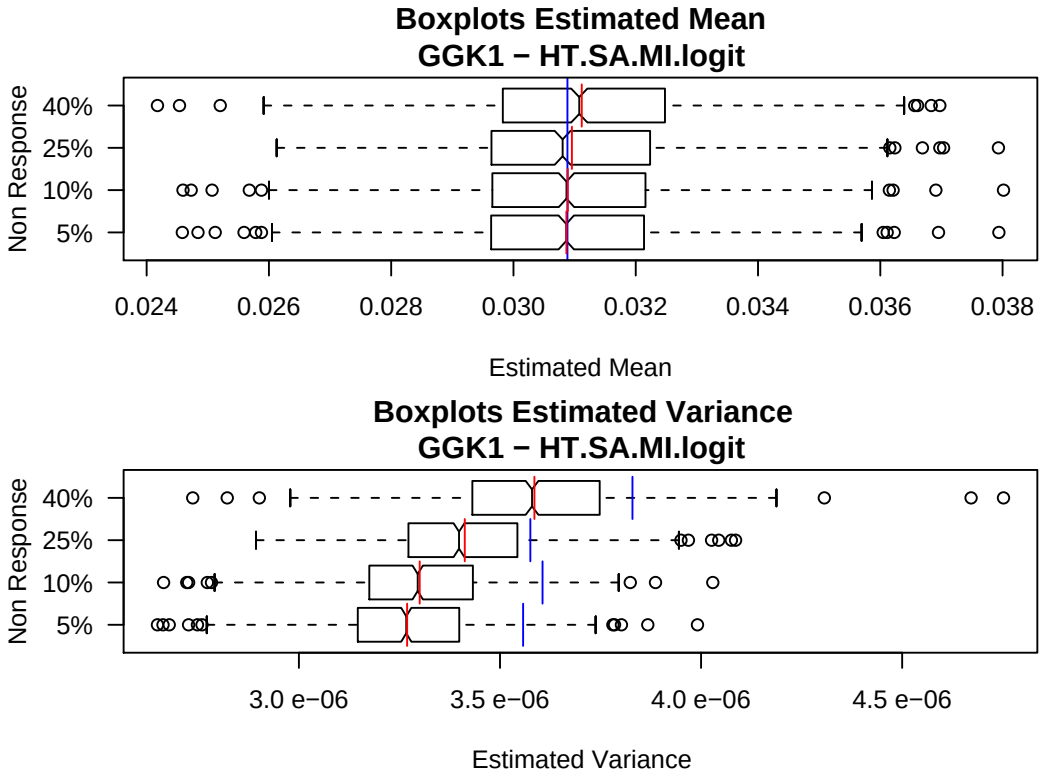

Figure 3.43: Point estimator HT with MI and its variance estimator for house size class 1.

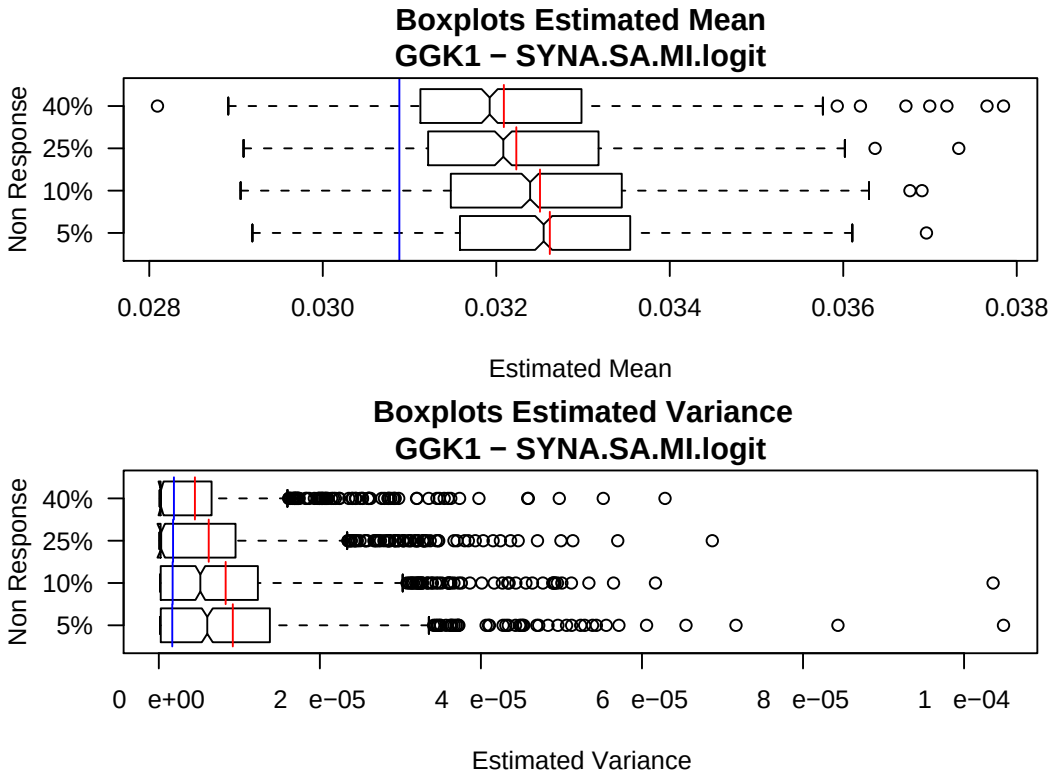

Figure 3.44: Point estimator Synthetic A with MI and its variance estimator for house size class 1.

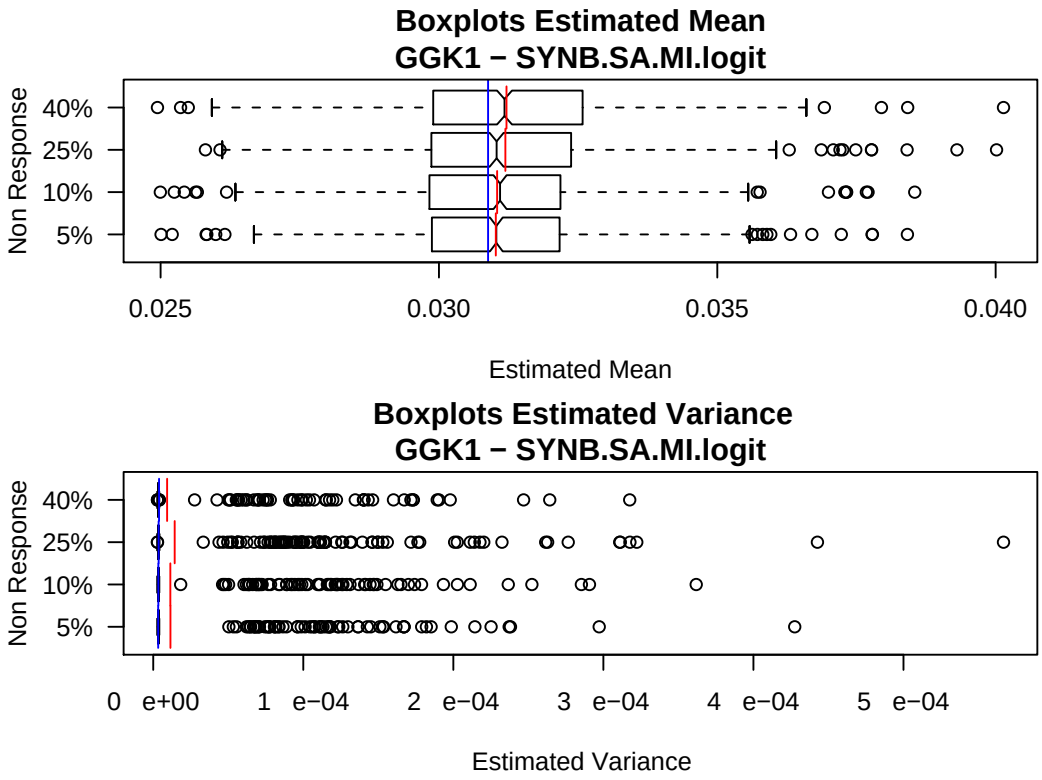

Figure 3.45: Point estimator Synthetic B with MI and its variance estimator for house size class 1.

This observations also work for the third small area classification, the combination of regional classes and house size classes. As in the two settings above, the GREG estimators, especially the GREG estimator with non-response correction outperform the model-based estimators with non-response. As can be seen in Figures 3.46 to 3.50 they may yield acceptable point estimators for all levels of non-response, but the variance estimators do produce problematic outcomes.

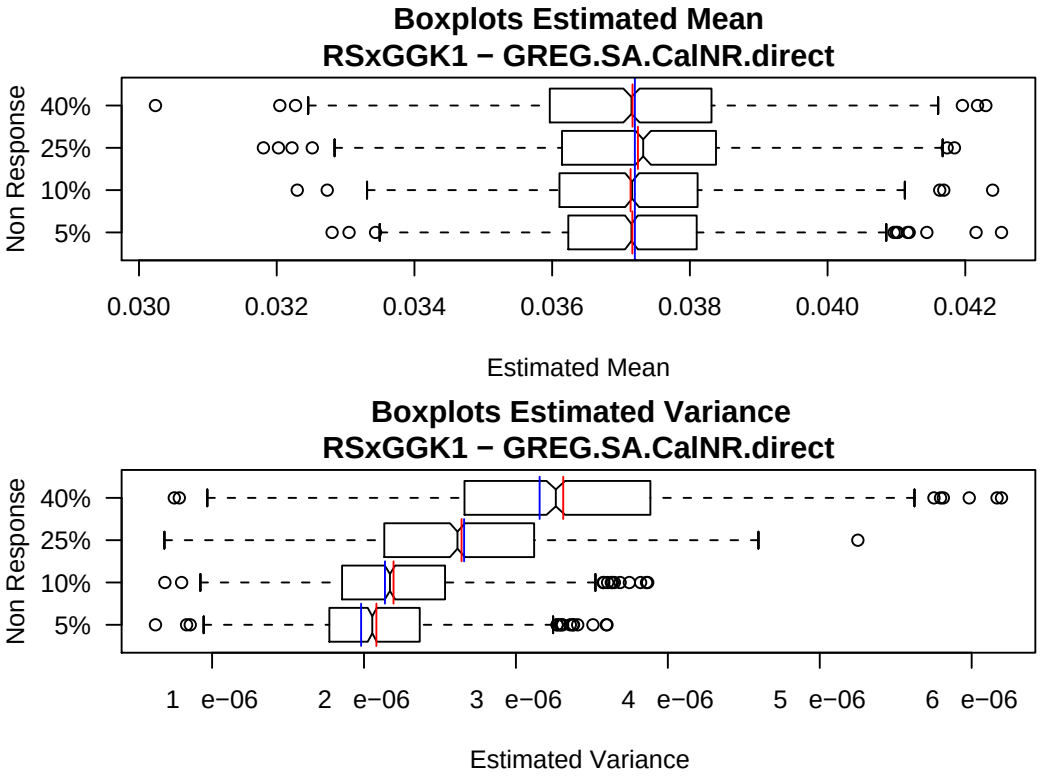

Figure 3.46: Point estimator GREG with non-response correction and its variance estimator for RSxGGK1.

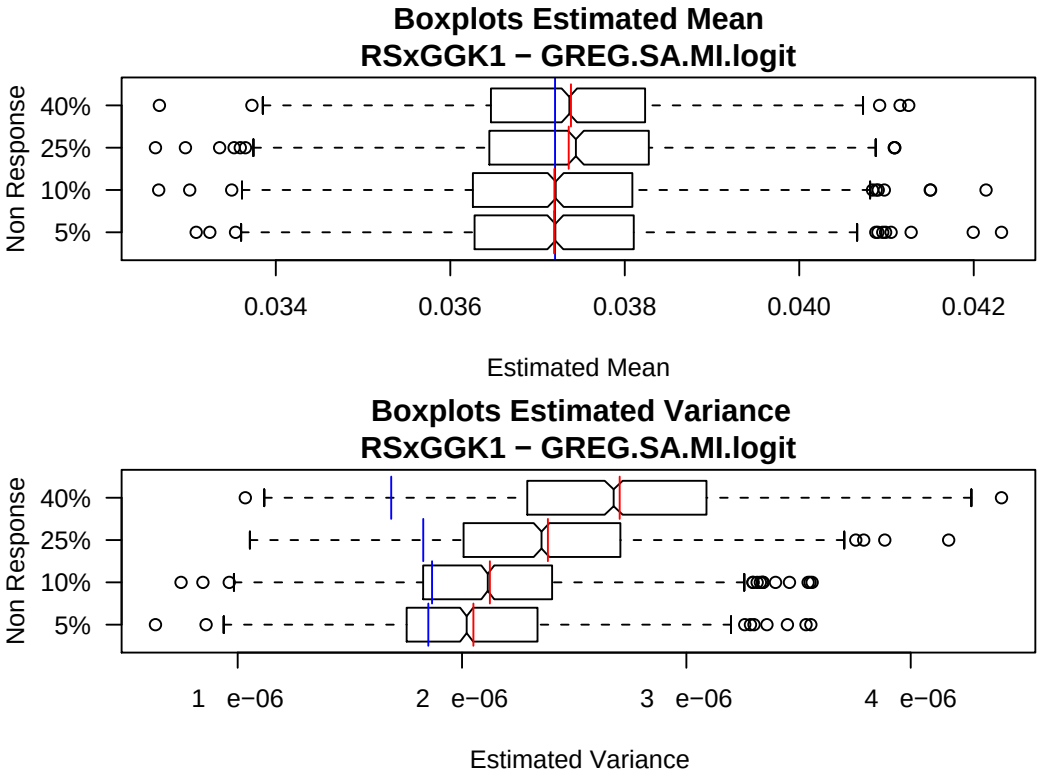

Figure 3.47: Point estimator GREG with MI and its variance estimator for RSxGGK1.

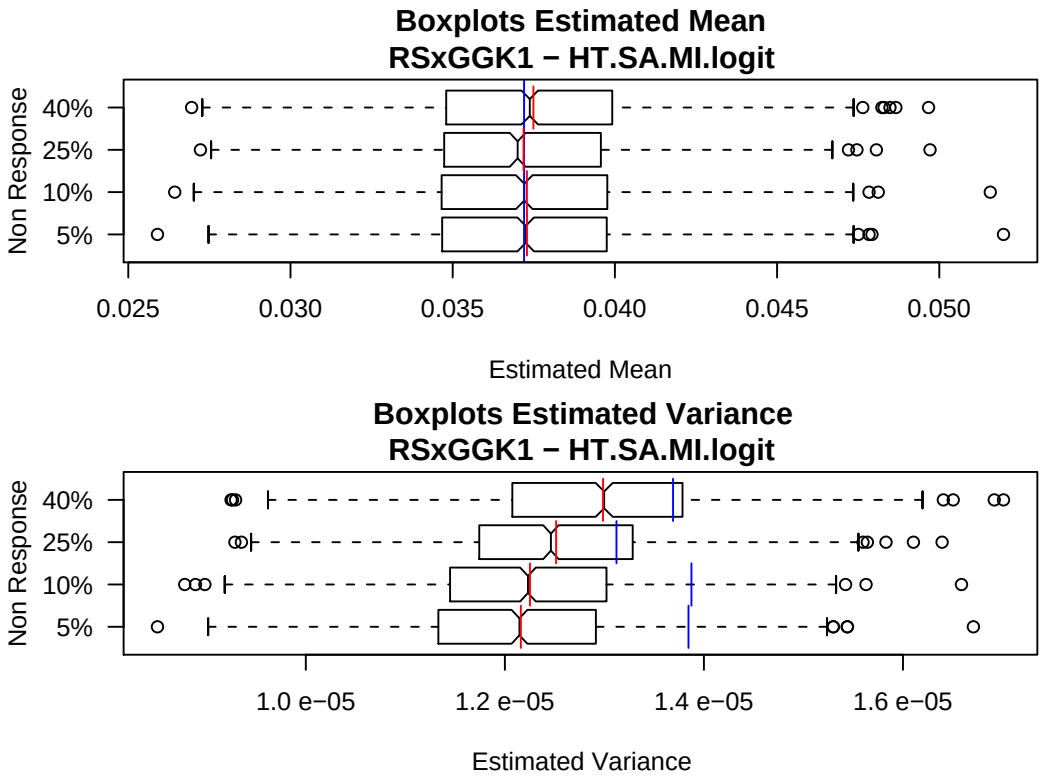

Figure 3.48: Point estimator HT with MI and its variance estimator for RSxGGK1.

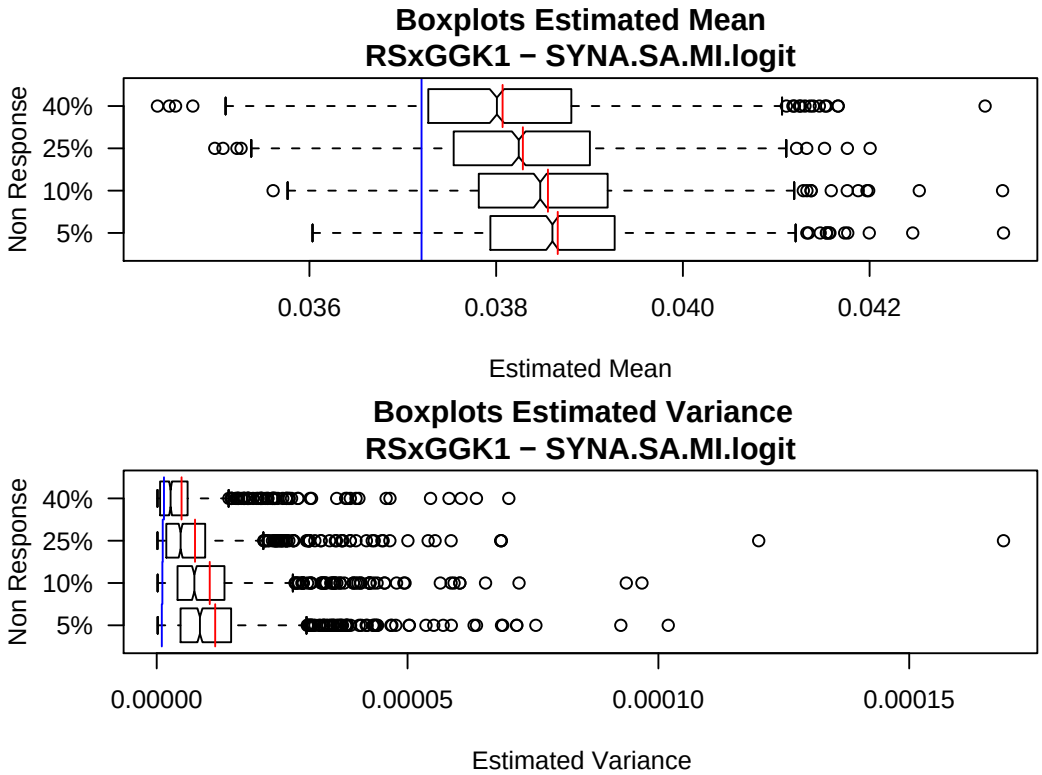

Figure 3.49: Point estimator Synthetic A with MI and its variance estimator for RSxGGK1.

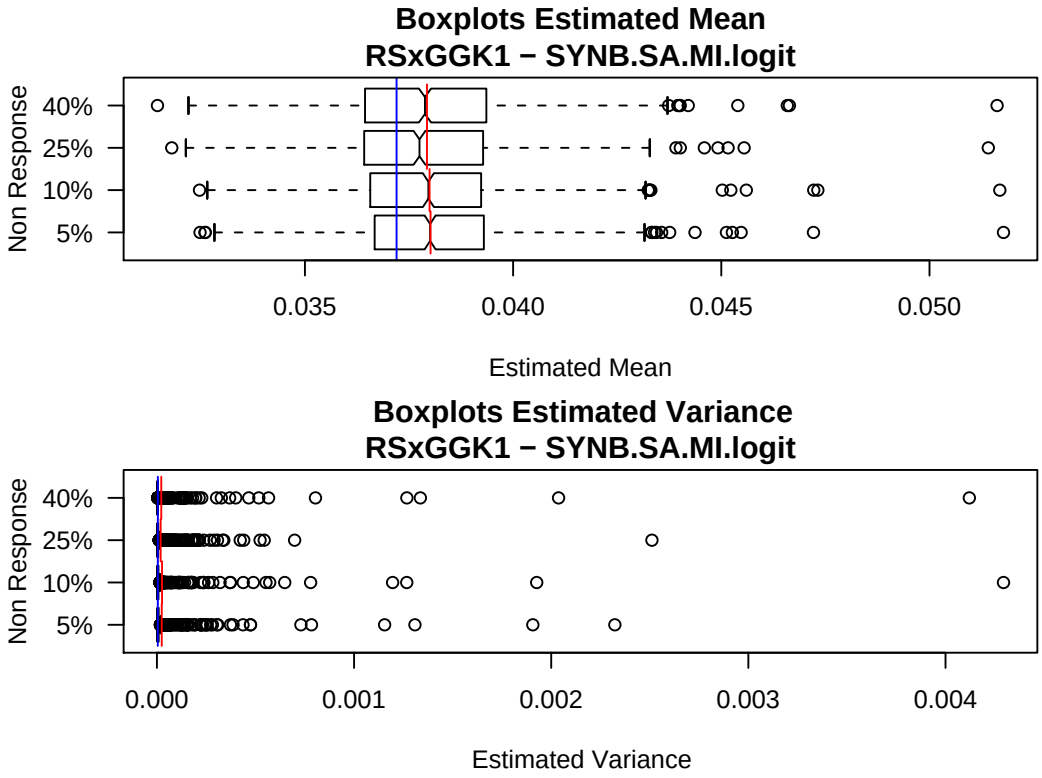

Figure 3.50: Point estimator Synthetic B with MI and its variance estimator for RSxGGK1.

Small area classification (iv), based on regional classes divided by eight, leads to many relatively small but rather homogenous small areas. Nevertheless, the problems of the second and third regional class mentioned above also show impact on this classification. Again the GREG estimator with non-response correction and with multiple imputation yields the best results. But alike the simulations with regional classes, the performance of the synthetic A estimators and therefore the EBLUP estimator as well, depend on the respective small area. The point estimates work well for the first small area which is based on the first regional class, although the variance estimate do not satisfy. In small area 14 however – as well as in all small areas based on regional class 2 or 3 – the problems of suitable estimation these estimators showed then occur again. For illustration see Figures 3.51 to 3.62.

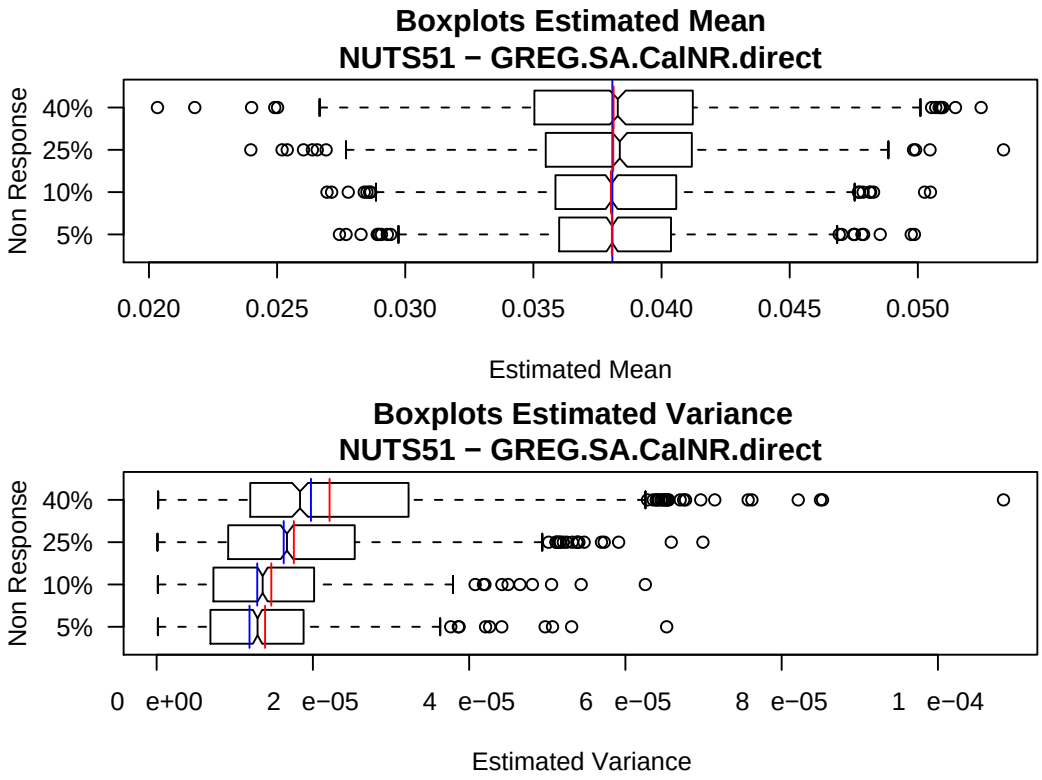

Figure 3.51: Point estimator GREG with non-response correction and its variance estimator for NUTS5/1.

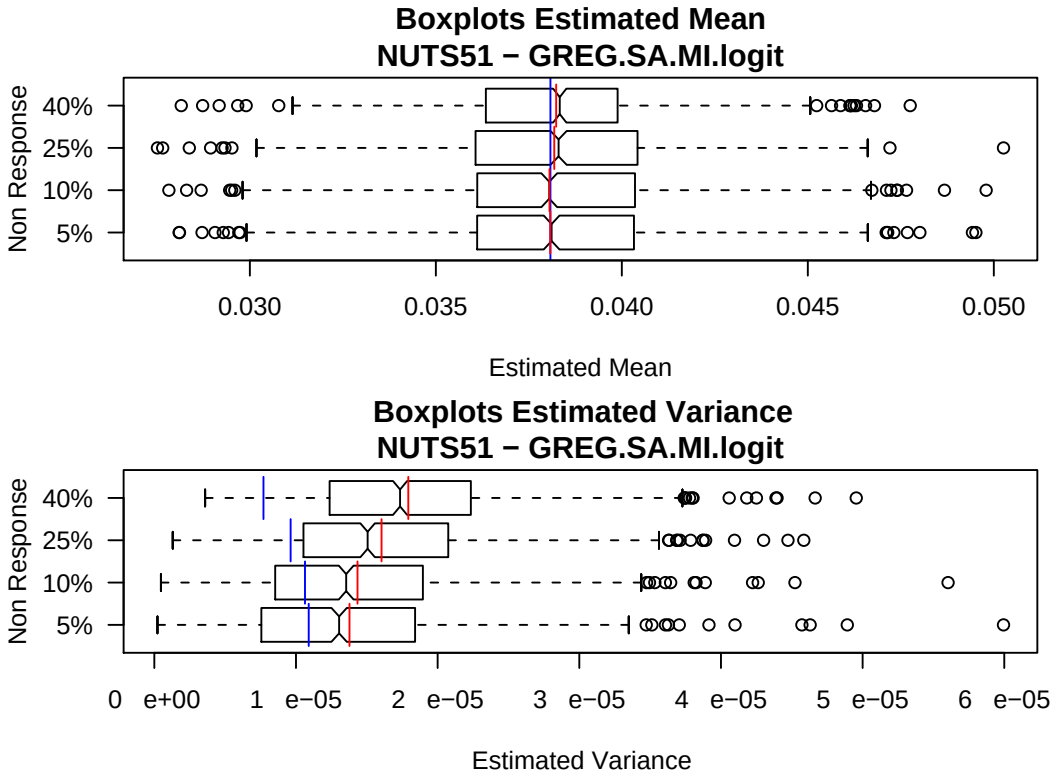

Figure 3.52: Point estimator GREG with MI and its variance estimator for NUTS5/1.

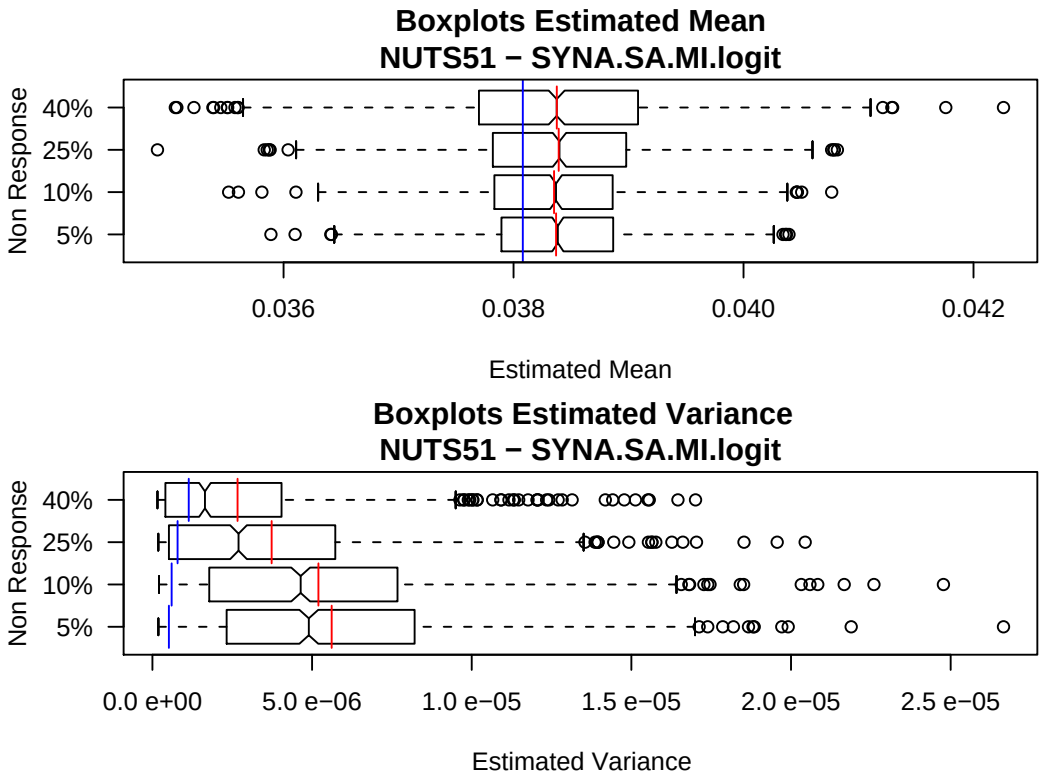

Figure 3.53: Point estimator synthetic estimator A with MI and its variance estimator for NUTS5/1.

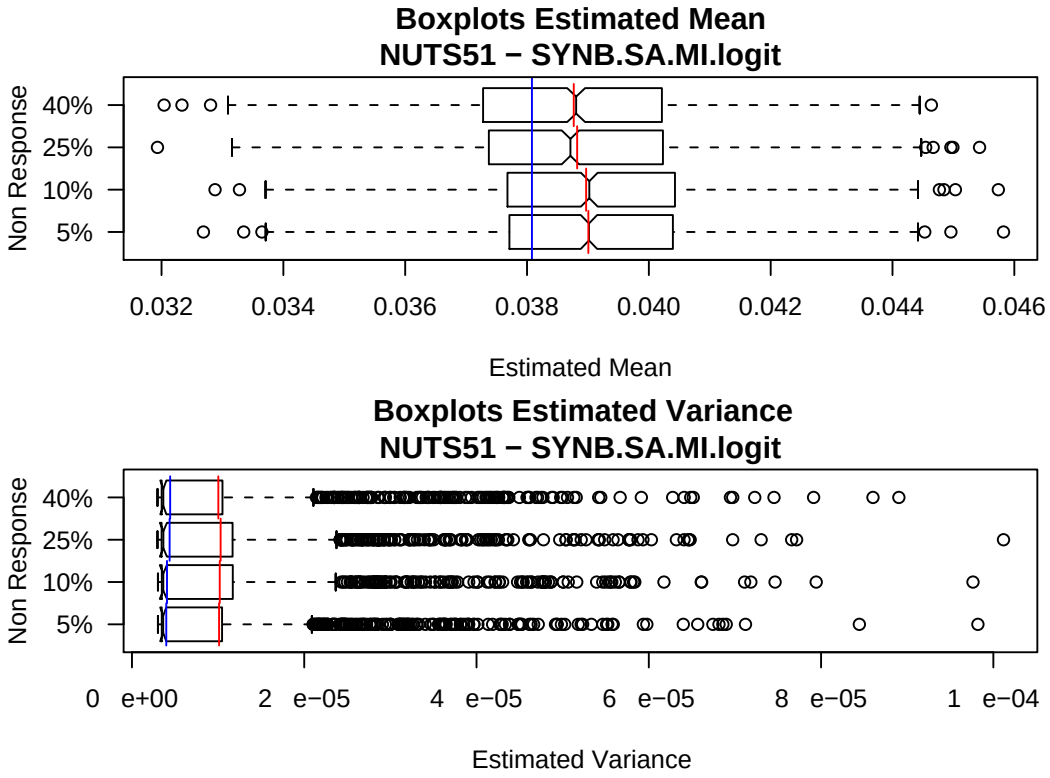

Figure 3.54: Point estimator synthetic estimator B with MI and its variance estimator for NUTS5/1.

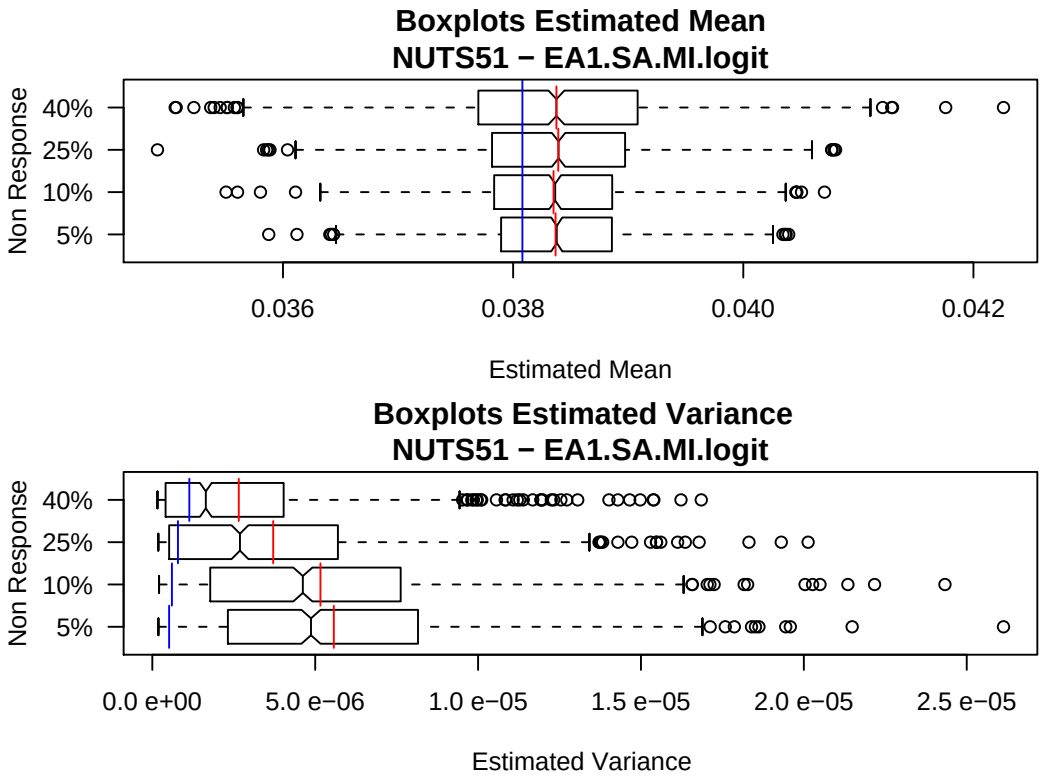

Figure 3.55: Point estimator EBLUP A estimator with MI and its variance estimator for NUTS5/1.

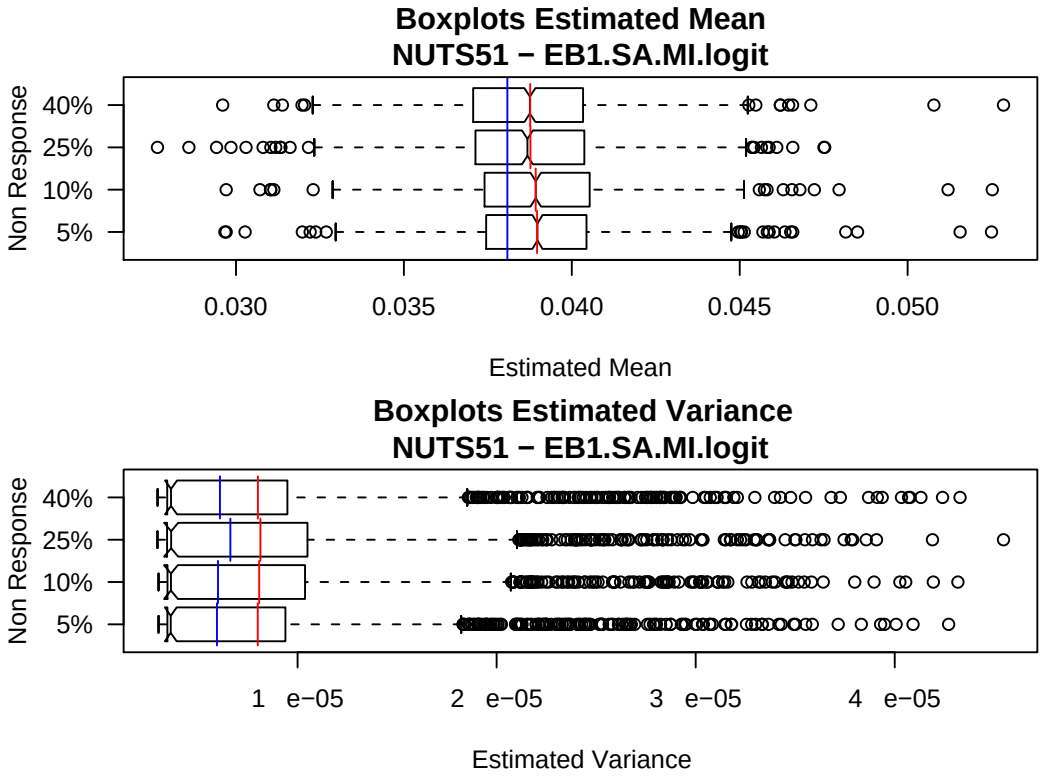

Figure 3.56: Point estimator EBLUP A estimator with MI and its variance estimator for NUTS5/1.

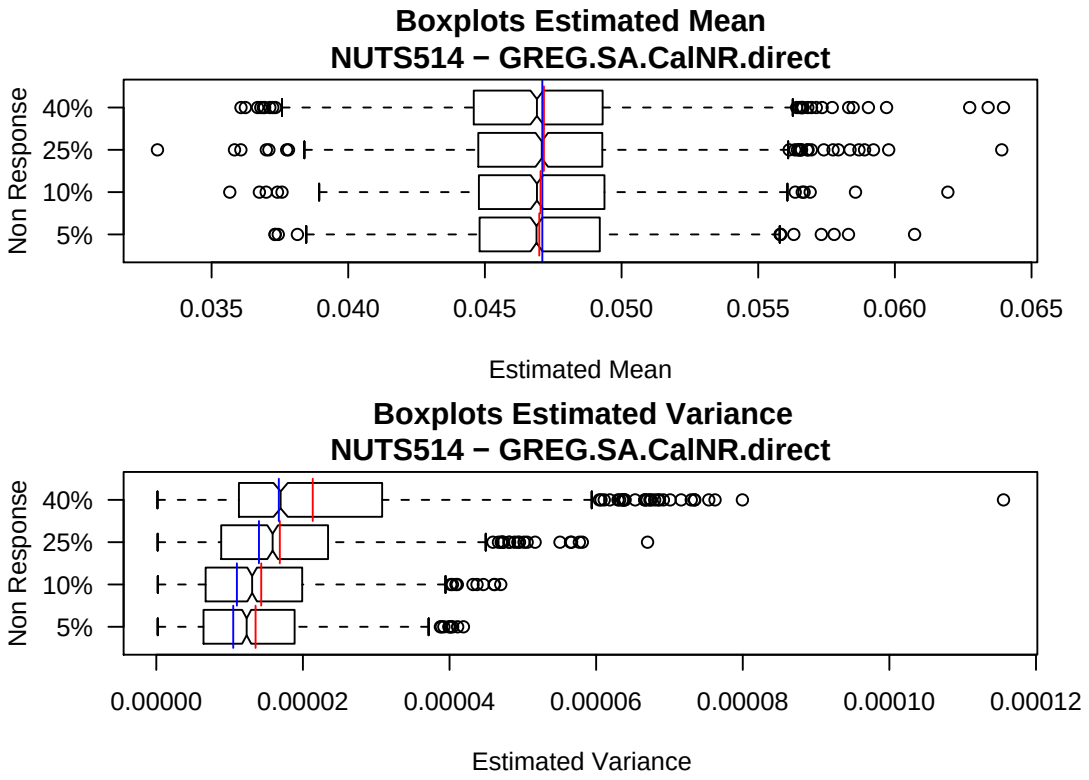

Figure 3.57: Point estimator GREG with non-response and its variance estimator for NUTS5/14.

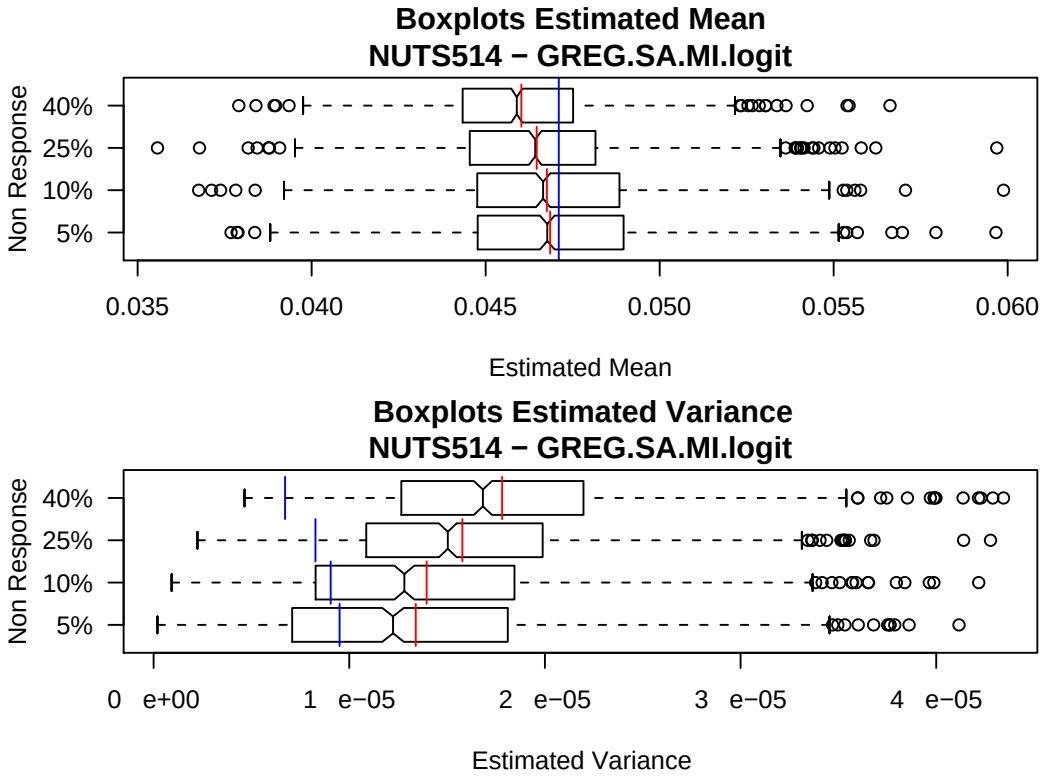

Figure 3.58: Point estimator GREG with MI and its variance estimator for NUTS5/14.

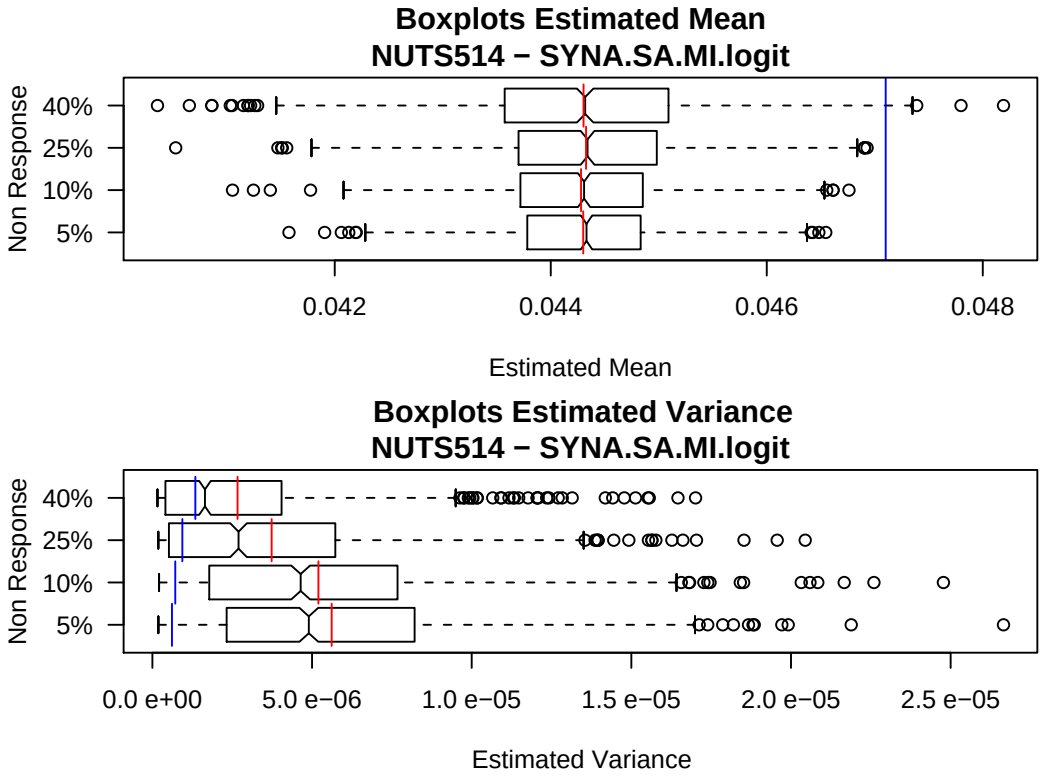

Figure 3.59: Point estimator synthetic estimator A with MI and its variance estimator for NUTS5/14.

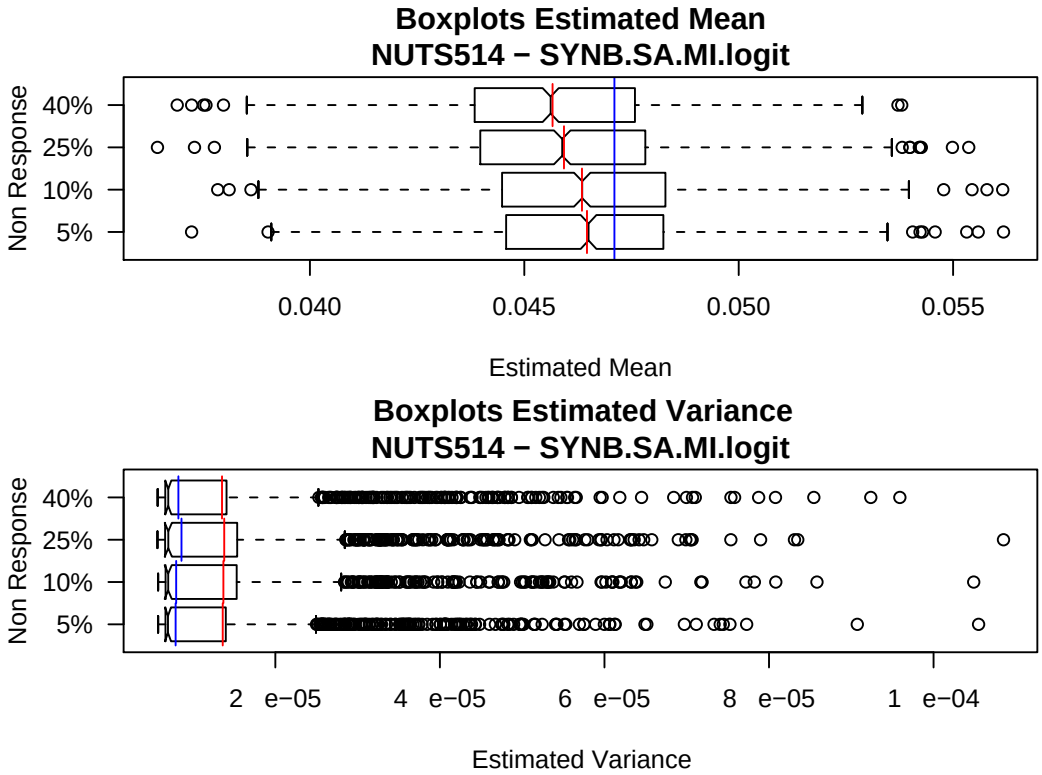

Figure 3.60: Point estimator synthetic estimator B with MI and its variance estimator for NUTS5/14.

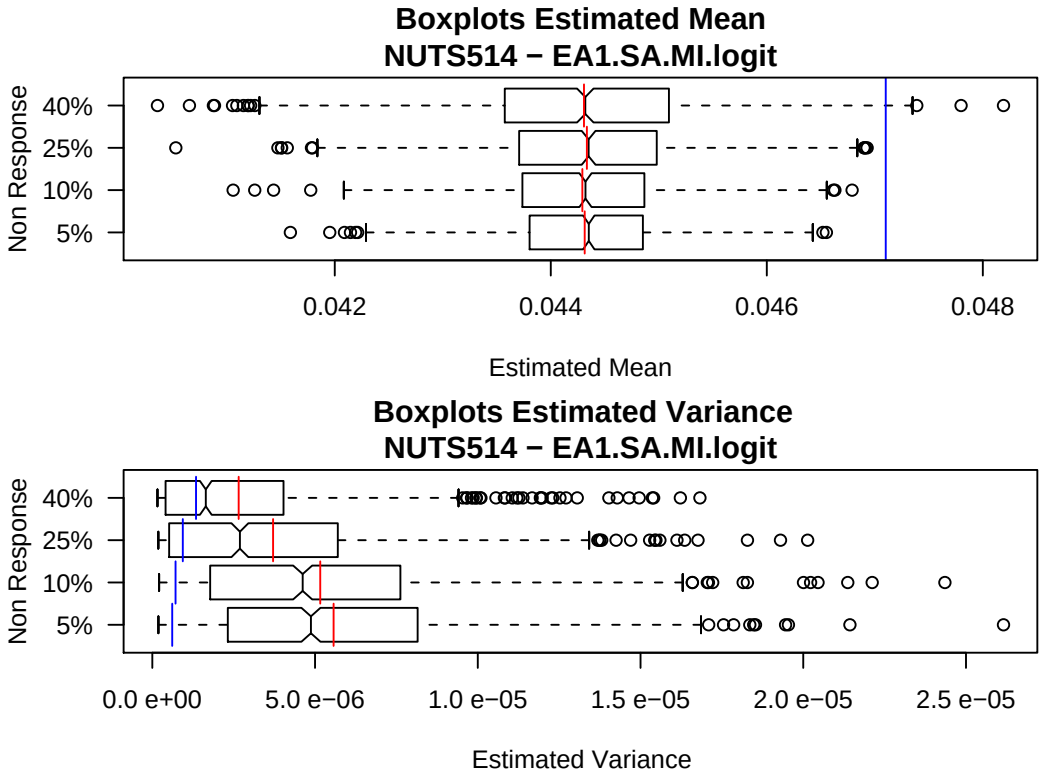

Figure 3.61: Point estimator EBLUP A estimator with MI and its variance estimator for NUTS5/14.

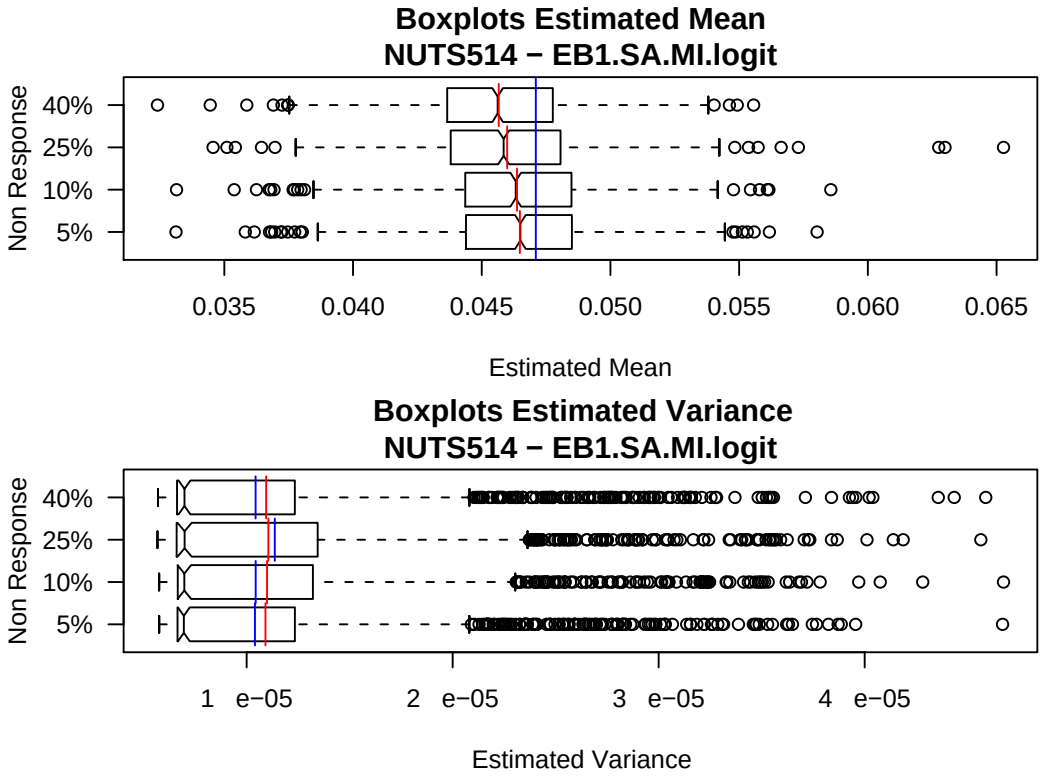

Figure 3.62: Point estimator EBLUP A estimator with MI and its variance estimator for NUTS5/14.

Besides the different small area classifications, the level of non-response is an important factor influencing the performance of the small area estimators. Therefore, simulations with varying levels were conducted. Observing the figures in this subsection, one can see that the level of non-response primarily influences the estimators of variance. Rising level of non-response leads to an increase in the variance as well. Another observation that can be made from Figures 3.43 to 3.62 is that the impact of the rising level of non-response is bigger, when multiple imputation is used to correct for non-response. Due to its structure the GREG estimator with non-response correction seems to be able to deal with even relatively high levels.

Nevertheless, one should keep in mind that no specific area information was used for the imputation model. It seems likely that additional variables on area-level help to stabilize the estimation results under multiple imputation. Further, one should not forget that the calibration and GREG estimators can hardly be applied in the presence of item non-response adequately.

3.2 The Finnish Register Data

3.2.1 Introduction

The simulations of the standard estimators described in Chapter 2 were performed in two steps. The first step included simulations of three target variables

- disposable income,
- unemployment, and
- household composition,

based on the entire Finnish population. Estimates were calculated for NUTS3 and NUTS4 regions using 500 samples with sample size 12,000. The common set of auxiliary information for all project partners was used for the estimators. The simulation results are described in the EURAREA project documentation (OFFICE of NATIONAL STATISTICS, 2003).

The second set of simulations was conducted to compare the performance of enhanced estimators (EBLUP estimators with time and area effects). The variable of interest here was the continuous variable disposable income on NUTS4 level of regions in Western Finland. The results of these simulations are described in this chapter.

3.2.2 Universe

The simulation study rested on the Finnish register-based employment statistics database. This database includes about 5.7 millions records, thereof about 2.0 millions in Western Finland, for the years between 1987 and 1998. The population on which the simulation study was based on was the Western Finland data. For standard estimators the data of

1998 was used, for the EBLUP estimator with correlated time effects, the data of the years 1994 until 1998 were made use of. Every record consists of about 60 variables per year including personal information (age, sex, education, language, marital status), household type information (dwelling, place of residence), data about working life, income and taxes.

The population in Western Finland in 1998 consisted of

1,472,184 individuals 15 years and older and

793,642 households.

The population size of the NUTS4 regions ranged from 4,142 to 131,610 households. The mean disposable income in Western Finland by NUTS4 regions was between 66,800 and 84,800 FIM in 1998 (see Figure 3.63). Generally income was lower in smaller rural regions and higher in large cities.

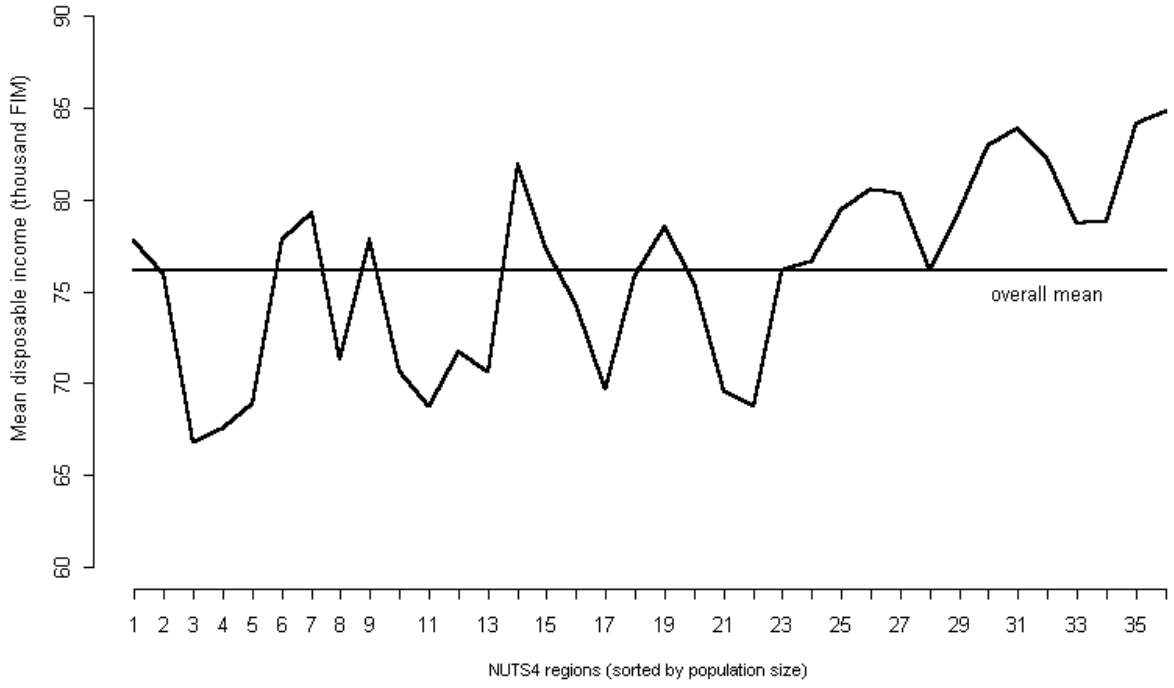

Figure 3.63: True mean disposable income in Western Finland by population size.

3.2.3 Sampling Design and Data Issues

From the population described in the last section, 1000 independent longitudinal cross-sectional samples with sample size 2,000 were drawn by stratified simple random sampling.

The stratification of the sample by socio-economic groups can be seen in Table 3.5. Farmers and other entrepreneurs with unstable income were oversampled to get the sampling design close to real design used in national surveys in Statistics Finland.

Table 3.5: Stratification by socio-economic groups with total sample size 2,000 and average inclusion probability 0.00227

Strata	n	Inclusion probability
wage and salary earners	830	0.00138
farmers	270	0.00618
other entrepreneurs	270	0.00448
pensioners	330	0.00086
other socio-economic categories	250	0.00083
not specified (mainly children)	50	0.00063

In the case of household samples the inclusion probabilities depend on the number of eligible individuals and, thus, the sampling design was a πPS. The sample database included all sampled individuals with individual and household level variables, sample indicator, sampling weights and regional indicators (NUTS4). All results calculated in the simulation study were analysed in two groups depending on the average sample size. Regions containing less than 40 household were considered as small regions, regions with 40 households or more as large regions. There were 21 regions in Western Finland containing less than 40 households. The complete distribution of the regions by sample size can be seen in Table 3.6.

Table 3.6: Number of NUTS 4 regions by sample size in Western Finland

Sample size	Number of NUTS4 regions	regions clustered by size
12-19	5	21 small regions (12-39)
20-29	10	
30-39	6	
40-49	4	15 large regions (40-284)
50-100	7	
120-284	4	
Total	36	

3.2.4 Simulation

To estimate the mean disposable income of the 36 NUTS4 regions in Western Finland the following six standard estimators were used:

– direct estimator,

- GREG estimator,
- synthetic estimator A,
- synthetic estimator B,
- EBLUP A estimator, and
- EBLUP B estimator.

All estimators (except for the direct which does not use any auxiliary information) use the same two auxiliary variables:

- the number of persons having tertiary education in the household and
- the total number of months of all household members being in employment per year.

The model-based standard estimators Synthetic A and EBLUP A with an underlying unit-level linear mixed model do not use design weights. The model-based estimators with an area-level model Synthetic B and EBLUP B do use design weights in the estimation of the model parameters, because the stratified sampling design causes significant overestimates (by 10-15%) in the unweighted case.

3.2.5 Performance of Estimators

Figures 3.64 and 3.65 show the relative bias (RB) and the RRMSE of the six calculated standard estimators for NUTS4 regional level. Note that the 36 regions are sorted by sample size.

As far as the bias is concerned, the GREG estimator outperforms the direct estimator for all 36 regions. The model-based standard estimators – the synthetic and the EBLUP estimators – perform quite similar. They all show a tendency to overestimate the mean disposable income in the smaller regions and to underestimate it in the 15 regions with a larger number of households. As the smaller regions tend to feature lower average income and the larger regions to feature higher average income (see Figure 3.63), this performance comes along with overestimating lower and underestimating higher mean values of income. The estimators EBLUP B and Synthetic B give almost equal results because the ML method of estimation model parameters gives more weight of synthetic than direct estimator part of EBLUP estimator. On the other hand, the EBLUP A estimator outperforms the synthetic estimator A due to the influence of the GREG estimator.

Using the RRMSE as a comparison criterion, the synthetic estimators perform best followed by the EBLUP estimators. The direct estimator and the GREG estimator come off clearly worse. Compared to the other standard estimators, for many regions their RRMSE are about two to four times the RRMSE of the synthetic or EBLUP estimators. This applies especially to the smaller regions containing less than 40 households and showing lower average income. The eye-catching large value of estimated RRMSE of the EBLUP A estimator in the second regions is caused by large outliers in six replicates.

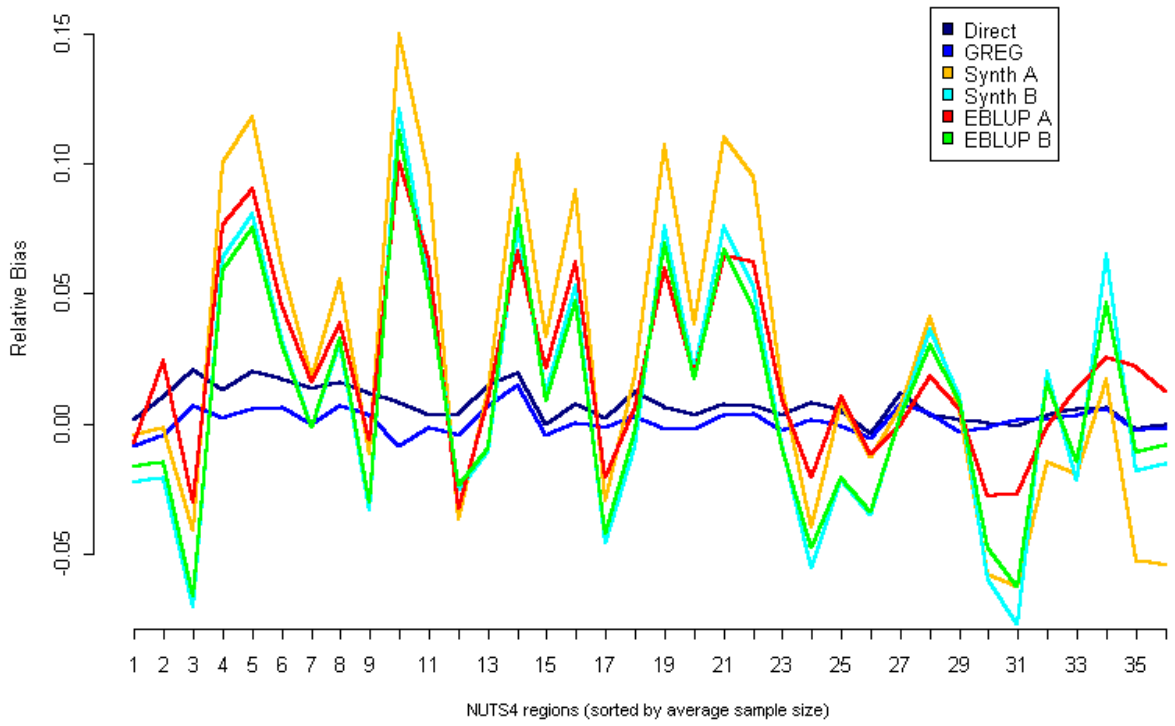

Figure 3.64: RB of standard estimators by NUTS4 regions

Figure 3.65 and the Figures 3.66 and 3.67 show that the precision of the direct and the GREG estimators depend significantly on the sample size. The precision of model-based estimators (especially the synthetic estimators) for small areas is about 20-30% higher than in large areas.

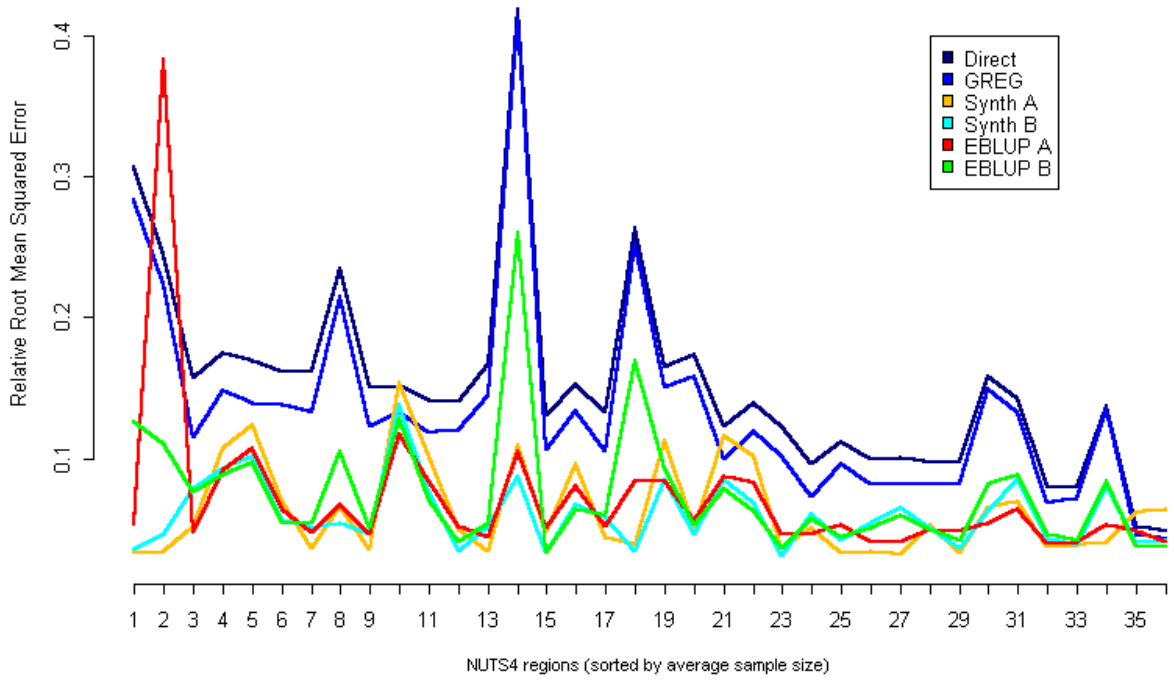

Figure 3.65: RRMSE of standard estimators by NUTS4 regions

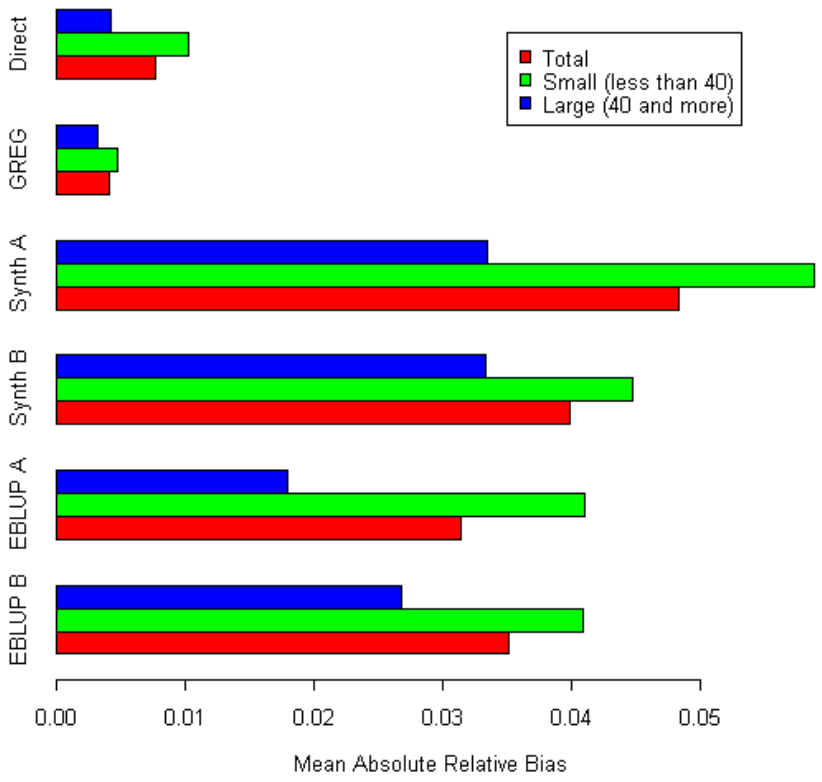

Figure 3.66: Mean absolute RB by average sample size

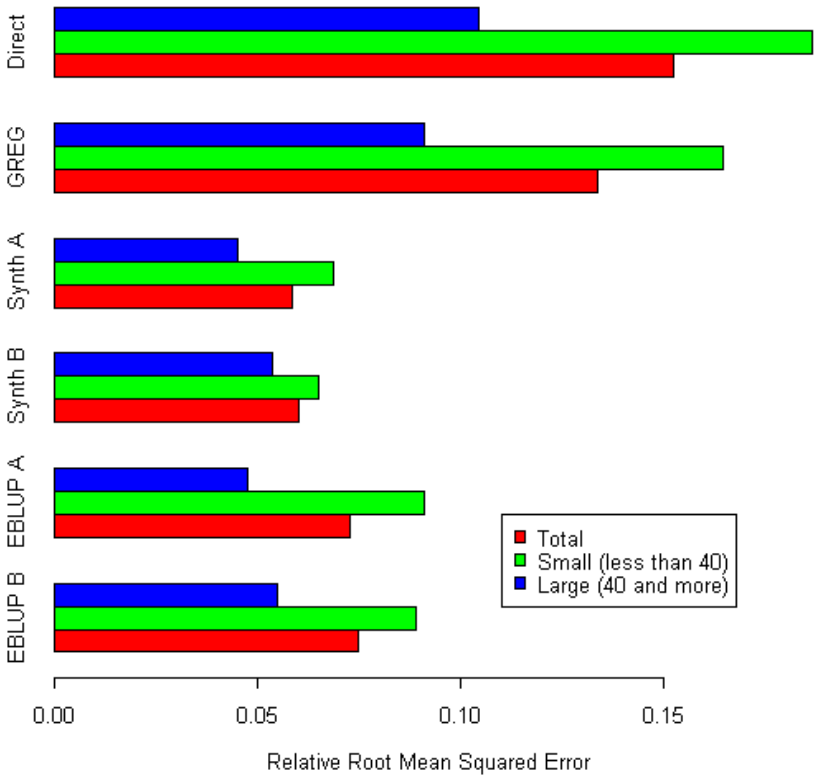

Figure 3.67: RRMSE by average sample size

Table 3.7 and Figure 3.68 show the performance of the 95% confidence intervals for the various estimators. For the direct estimator the average coverage rate of the confidence intervals is well below the 95 percent value. This applies especially for small regions with a sample size of less than 40. According to (OFFICE of NATIONAL STATISTICS, 2003) this problem is likely to be caused by ignoring the variance estimators' own variance. The problems arising in the context of estimating within-area variance are probably the reason for confidence intervals of the synthetic estimator B and EBLUP B estimator, which are too tight. The coverage rates of the model-based estimators vary conspicuously by regions. For a large number of regions the coverage rate is close to 100%, but for some regions it is significantly lower.

Table 3.7: Coverage rates of standard estimators of 95% confidence interval

	Simple mean			Minimum	Maximum
	Total	Small (< 40)	Large (≥ 40)		
Direct	91.8	90.4	93.8	84.5	95.0
GREG	95.6	96.5	94.3	88.2	98.4
Synthetic A	96.5	95.2	98.4	75.4	99.8
Synthetic B	90.3	91.7	88.4	56.9	99.4
EBLUP A	95.2	94.7	95.9	78.9	99.3
EBLUP B	89.2	91.2	86.4	53.0	99.4

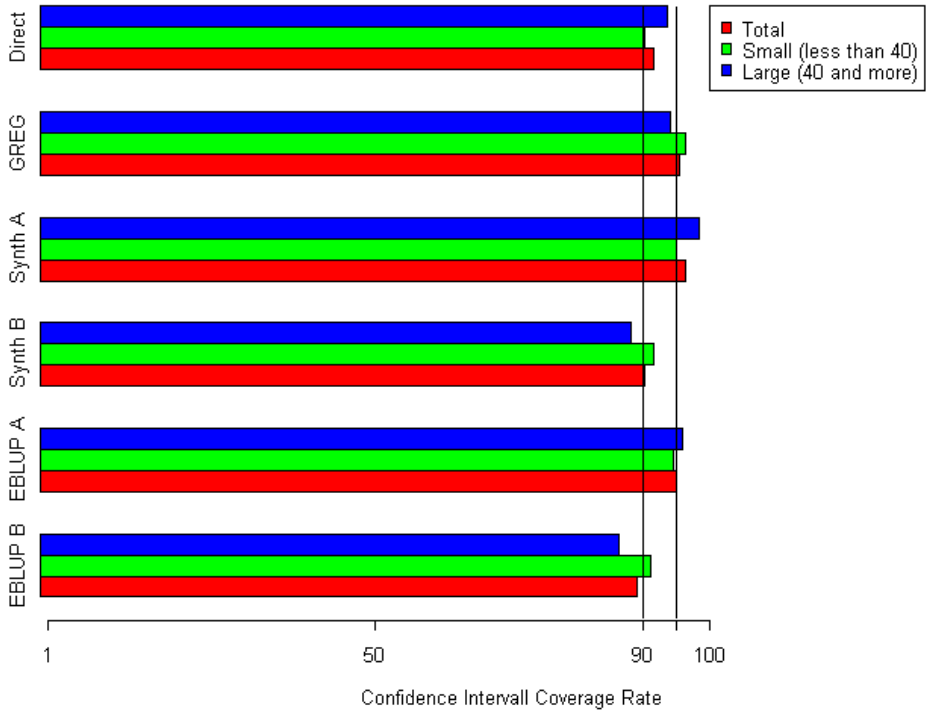

Figure 3.68: Confidence interval coverage rates

3.2.6 Summary

The experience with standard estimator simulations based on real income data illustrated some theoretical and practical points, which have to be taken into account applying standard estimators into practice.

– Design-based estimators direct and GREG outperform model-based estimators according to bias, but accuracy of direct and GREG estimators in regions with small sample size are poor.

- Model-based estimators synthetic and EBLUP are less variable in small regions, but tend to *miss the target* in regions that differ significantly from overall mean.
- Outliers in combination with small sample size can cause significant problems in the estimation of model parameters and MSE.
- The informative sampling where the inclusion probability is correlated with target variable can cause heavily biased estimates using area-level model (EBLUP B and synthetic B estimators).
- The choice of auxiliary information has great importance because the estimation results can be significantly improved.

Chapter 4

Conclusions

The large simulation study on the German data-sets, in general provided equivalent recommendations to the EURAREA project. However, the special stratification in the GMC with regards to the house size classes gave additional ideas.

This may lead to the conclusion that deeper knowledge on the data and their regional peculiarities as well as their inclusion into the models is needed in order to adequately apply the small area methods. Therefore, these methods are more demanding in applications than the classical methods like the Horvitz-Thompson and the GREG estimator but may lead to a non-negligible gain in efficiency.

Within the workpackage report, a sufficient overview to selected tasks of the small area simulation study was given. The results of the study were presented as a ceteris paribus analysis. The entire results can be drawn from the electronic recommended practice manual in deliverable D12.2.

In addition to the general study on the GMC, a non-response study was performed. First results give some impression that a comparison of the methodology seems a little more sophisticated. The main problem is the presence of adequate auxiliary information on area-level in connection with a considerable number of small areas. However, an interpretation of the results on specific areas seems inadequate in terms of the methodology, whereas the end-user may want to do this investigation. Nevertheless, further research on adequate variance estimation methods for small area estimators in the presence of non-response seems to be needed.

References

Münnich, R. and Rässler, S. (2004): Variance estimation unter multiple imputation. Proceedings of the Q2004 conference.

Office of National Statistics (2003): Report on the performance of the ßtandardëstimators. P. Heady and R. Lehtonen. Version1. 0/300503. EURAREA working paper IST-2000-26290.

Rao, J. N. K. (2003): *Small area estimation*. John Wiley & Sons. Wiley Series in Survey Methodology.

Särndal, C. E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*. New York: Springer-Verlag.

The EUAREA Consortium (2003): Draft report on the performance of the „Standard" Estimators. `http://www.statistics.gov.uk/methods_quality/eurarea/`.