Evaluation of Small Area Techniques for Applications in Official Statistics

Submitted in partial fulfilment of the requirements for the degree

Dr. rer. pol.

to the
Department IV
at the University of Trier
submitted by

Diplom Volkswirt Jan Pablo Burgard
Am Weidengraben 200, 54296 Trier
born 21.12.1983 in Lüneburg

Supervisors:
Prof. Dr. R. T. Münnich (Universität Trier)
Prof. Dr. P. Lahiri (University of Maryland)

Contents

Acknowledgment

I am profoundly grateful to my first supervisor Dr. Ralf Münnich for his guidance, trust in me and my work. His support, also in the career planning and scientific projects made a real difference.

To my second supervisor Dr. Partha Lahiri I am also very grateful indeed for supporting me and for many very fruitful discussions influencing my research.

For the extremely pleasant and productive working atmosphere and discussions I sincerely thank all my colleagues, especially Dr. Jan-Philipp Kolb, Dr. Martin Vogt, Thomas Zimmermann, and Stefan Zins.

Finally, I would like to warmly thank my family for helping me to overcome the tough times during the writing process.

Deutsche Zusammenfassung / German Summary

In Politik und Wirtschaft und damit in der amtlichen Statistik wird aktuell die präzise Schätzung von Kennzahlen für kleine Regionen oder Teile von Populationen, sogenannten Small Areas oder Domains, intensiv diskutiert. Die derzeit verwendeten designbasierten Schätzmethoden beruhen überwiegend auf asymptotischen Eigenschaften und sind somit bei großen Stichprobenumfängen zuverlässig. Bei kleinen Stichprobenumfängen hingegen greifen diese designbasierte Überlegungen oft nicht, weswegen für diesen Fall spezielle modellbasierte Schätzverfahren entwickelt wurden – Die Small Area-Verfahren. Diese können zwar Verzerrungen aufweisen, haben dafür aber häufig kleinere mittlere quadratische Fehler (MSE) der Schätzung als designbasierte Schätzer. In dieser Arbeit werden sowohl klassische, designbasierte Schätzmethoden, als auch modellbasierte Schätzmethoden vorgestellt und miteinander verglichen. Der Fokus liegt hierbei auf der Eignung der verschiedenen Methoden für einen Einsatz in der amtlichen Statistik. Hierzu werden zunächst Theorie und geeignete Algorithmen für die benötigten statistischen Modelle vorgestellt, die als Grundlage für die darauf folgenden modellbasierten Schätzer dienen. Anschließend werden für Small Area Anwendung entwickelte Stichprobendesigns vorgestellt. Auf diesen Grundlagen aufbauend werden sowohl designbasierte Schätzer und als auch modellbasierte-Schätzverfahren entwickelt. Besondere Berücksichtigung findet hierbei der area-level empirisch besten Prädiktor für binomiale Variablen. Für diesen analytisch nicht lösbaren Schätzer werden numerische und Monte-Carlo Schätzverfahren vorgeschlagen und verglichen. Weiterhin werden für ihn Methoden zur Schätzung seines MSEs herausgearbeitet.

Eine sehr beliebte und flexible Resampling-Methode, die im Bereich der Small Area Statistik viel Anwendung findet, ist der parametrische Bootstrap. Ein großer Nachteil des Verfahrens ist dessen hohe Computerintensivität. Um diesen Nachteil abzuschwächen, wird in dieser Arbeit erstmals eine Varianzreduktionsmethode für parametrische Bootstraps vorgeschlagen. Anhand von theoretischen Überlegungen wird das enorme Potential dieses Vorschlags nachgewiesen. Mit Hilfe einer Monte-Carlo Simulationsstudie wird gezeigt, wie starke Varianzreduktion mit dieser Methode in realistischen Szenarien erreicht werden kann. Diese kann bis zu 90% betragen. Dadurch wird tatsächlich die Nutzung vom parametrischen Bootstrap in Anwendungen in der amtlichen Statistik realisierbar.

Schließlich werden die vorgestellten Schätzmethoden in einer großen Monte-Carlo Simulationsstudie in einer konkreten Anwendung für die schweizerische Strukturerhebung hin untersucht. Dabei werden Fragestellungen erörtert, die gerade für die amtliche Statistik von hoher Relevanz sind. Insbesondere sind dies:

(a) Wie klein gegliedert dürfen Areas sein, ohne dass die Präzision der Schätzung ungeeignet wird?

(b) Sind die Genauigkeitsangaben für die Small Area Schätzer reliabel genug, um sie für die Veröffentlichung zu nutzen?

(c) Stören sehr kleine Areas bei der Modellierung der interessierenden Variablen? Und wird dadurch ein Verschlechterung der Schätzungen größerer und damit wichtigere Areas verursacht?

(d) Wie können Kovariablen, die in verschiedenen Aggregationsebenen vorliegen auf geeignete Weise zur Verbesserung der Schätzung herangezogen werden.

Als Datengrundlage dient die schweizerische Volkszählung von 2001. Die zentralen Ergebnisse sind, dass aus Sicht des Autors die Verwendung von Small Area Schätzern für die Produktion von Schätzwerten für Areas mit sehr geringen Stichprobenumfängen trotz des Modellierungsaufwandes ratsam ist. Die MSE-Schätzung bietet dabei ein brauchbares Maß der Präzision, erreicht aber nicht in allen Small Areas die Reliabilität wie die Varianzschätzung für designbasierte Schätzer.

List of Figures

List of Tables

List of Algorithms

Mathematical Symbols and Abbreviation

$\odot$	Element wise multiplication.
$\emptyset$	Empty set.
β	Vector of fixed-effects model parameters of length $p+1$ including the intercept.
$\text{Bern}(\theta)$	Bernoulli distribution with success probability θ.
$\text{Bin}(n,\theta)$	Binomial distribution with success probability θ and n trials.
e	Vector of residuals.
e	The Euler number $\lim_{r\to\infty}\left(1+\frac{1}{r}\right)^r \approx 2.7183$
$\text{E}[*]$	Expectation of $*$.
d,D	Running index $d = 1\ldots D$ with D as the total number of areas.
d_*	Deviance contribution of unit $*$.
δ_*	Sum of the squared rescaled weights of area $*$.
$\text{diag}(*)$	Diagonal matrix with elements $*$.
f,f	Arbitrary functions.
g	$N \times D$ matrix of the g-weights.
$g(\cdot)$	Arbitrary function.
g_1, g_2, g_3	Components of the Prasad-Rao like MSE-estimators.
γ_*	shrinkage-factor for area $*$.
h	Arbitrary function.
h,H	Running index $h = 1\ldots H$ strata.
H	Hessian matrix.
H	Functions for the parametric double bootstrap.
$I_{(*)}$	Identity matrix of dimension $* \times *$.
$\mathbb{I}_*$	Indicator function, if $*$ is true, then $\mathbb{I}_* := 1$ else $\mathbb{I}_* := 0$.
$\mathbb{I}_*(**)$	Indicator function, if $*$ is true, then $\mathbb{I}_*(**) := **$ else $\mathbb{I}_*(**) := 0$.
j,J	Running index $j = 1\ldots J$.
J	Jacobian.
k,K	Running index $k = 1\ldots K$.
κ	Arbitrary working variable.
L	Likelihood function.
l	Log-likelihood function.

l_*	Log-likelihood function for unit $*$.
log	The natural logarithm.
$\mathscr{L}$	Bootstrap distribution.
$logit(\theta)$	Is the function $log(\frac{\theta}{1-\theta})$ for $\theta \in (0,1)$.
m	Arbitrary variable.
M	Values used in the parametric double bootstrap.
μ, μ_*	Mean, mean of area $*$.
$\widehat{\mu}_{*,**}$	Estimator $**$ for the mean of area $*$.
$\mathrm{MSE}(*)$	Mean squared error (MSE) of $*$.
$\widehat{\mathrm{MSE}}(*)$	Mean squared error estimator for $*$.
n,n_*	Sample size, sample size in area $*$.
N,N_*	Population size, Population size of the area $*$
$N(\mu, \sigma^2)$	Normal distribution with expectation μ and variance σ^2
p	Number of fixed effects parameters excluding the intercept vector.
P	Projection matrix used for REML estimation.
$\mathrm{P}(*)$	Probability of $*$.
π, π_*	$n \times 1$ vector of the inclusion probabilities, inclusion probability of observation $*$.
ϕ	Is either a vector of variance components, e.g $\phi = (\sigma_e^2, \sigma_u^2)$, or the dispersion parameter of the exponential family of distributions.
ψ	Vector of unknown model parameters.
Ψ	An arbitrary Population parameter.
ql	Quasi (log-)likelihood function.
$\mathrm{Q}_*(\cdot)$	The $*$-th quantile of the distribution $\cdot$
r, R	Running index $r = 1 \ldots R$.
R	Arbitrary remainder term.
$\mathscr{S}, \mathscr{S}_*$	Set of sampled units, Set of sampled units in area $*$.
σ_*^2	Varianz of the variable $*$.
tan	Tangent function.
τ, τ_*	Total, total in area $*$
$\widehat{\tau}_{*,**}$	Total estimate from estimator $**$ in area $*$
θ	Success probability of the binomial distribution.
θ, θ_*	Ratio, ration in area $*$.
$\widehat{\theta}_{*,**}$	Ratio estimate from estimator $**$ for the ration in area $*$.
u_k	Random effects vector of length ν_k for the k-th random effect
u	Random effects vector of length ν
$\mathscr{U}, \mathscr{U}_*$	Universe or population, universe or population in area $*$
ν_k	Number of columns of the k-th random effects design matrix z_k and length of u_k.
ν	Number of columns of the combined random effects design matrix z and length of u.
v	Arbitrary variance function.
V	$n \times n$ variance covariance matrix on sample length.

$\mathrm{V}[*]$	Denotes the variance of $*$.
$\mathrm{VCOV}[*]$	Denotes the variance covariance matrix of $*$.
w, w_*	Design weight, design weight for unit $*$.
ω	Working weight.
W	Arbitrary weights matrix.
x	$n \times p+1$ matrix of the covariates including the preceding column of ones on sample length.
x_*	If $*=i$ or $*=id$ a $1 \times p+1$ row vector of covariate, if $*=d$ a $n_d \times p+1$ matrix of covariates for area d.
X	$N \times p+1$ matrix of the covariates including the preceding column of ones on universe length.
ξ	Arbitrary variable.
y	$n \times 1$ vector of the independent variable on sample length.
Y	$N \times 1$vector of the independent variable on universe length
z_k	$n \times \nu_k$ Design matrix of the k-th random effect on sample length.
z	$=(z_1,\ldots,z_K)$ Combined design matrix of the K random effects on sample length.
Z_k	$N \times \nu_k$ Design matrix of the k-th random effect on universe length.
Z	$=(Z_1,\ldots,Z_K)$ Combined design matrix of the K random effects on universe length.
ζ	Arbitrary working variable.

Chapter 1

The Need for Small-Scale Estimates in Official Statistics

For the 2010 census round, the Swiss Federal Statistical Office conducted a register assisted census for the first time. Instead of interviewing the full population (full census), only a sample was surveyed (Bundesamt für Statistik Schweiz, 2013b). This implies a change of paradigm in census methodology. In the full census, all the population values could be obtained by counting out the entire census record. In contrast, under a register based census, the population values produced are estimates. The classical estimators used by official statistics to produce these census estimates are design based estimators. The properties of these estimators rely mainly on asymptotics. That is, they should hold for very large sample sizes. For typical surveys, where only estimates on higher aggregation are needed, e.g. on a state or national level, the design based estimators perform well. However, if the domain of interest is rather small, the sample size allocated therein is in many cases also very small. Large sample asymptotic properties will not hold for these domains. For example, the cross classification of sex, age classes and nationality can be such a domain or also a county. As the census is the central information source for politics and the economy for the figures on population, households, family, housing, employment, mobility, education and religion, amongst others, it is crucial to pay close attention to the accuracy of these estimates. For many useful population figures, the classical design based estimators will provide poor quality estimates.

As the title suggests, the aim of this work is to evaluate small area techniques for applications in official statistics. These techniques comprise the point estimation on small areas, with special focus on proportions, the MSE estimation for these point estimates and the use of register data which might only be available on certain aggregation levels due to disclosure reasons. Further, the practicability of resampling methods for the MSE estimation of small area estimates is of major interest, where no analytical approximation exists. The evaluation is done with design based Monte Carlo studies, giving guidance on the usability of the different techniques at hand, especially for official statistics applications. The Swiss Census data set from 2010 is used as the exemplary data set. This data

set has been kindly provided by the Swiss Statistical Office within the project *Simulation der Strukturerhebung und Kleingebiet-Schäzungen* (see Münnich & Burgard, 2012b) with permission to use it for this thesis.

This work is organized into six chapters. Following this introduction, the fundamental methods for modelling the dependent variable by using covariates are presented. These methods consist of the linear and generalized linear regression model and the linear and generalized linear mixed models and are used in the following Chapter for the construction of most of the estimators. Besides the models, estimation methods and algorithms are also discussed.

In Chapter Three a brief introduction is given to sampling designs relevant in the context of small area estimation. On this basis, design based estimators are presented which are typically used by statistical offices to produce estimates for publishing. Further, a variety of small area estimation methods are presented. Besides the point estimates, precision estimates are discussed for both the design based estimators and the small area methods. Special focus is placed on the area level empirical best predictor, which is developed for binomially distributed variables of interest. For this estimator, analytically intractable integrals need to be approximated. Numerical and Monte Carlo integration techniques are discussed which may be apt for solving these integrals. Also, two resampling methods for the estimation of the MSE of this estimator will be presented. Last but not least in this chapter, a modelling approach is proposed, which can include register information on an intermediate level. This situation can arise when register information is only made available, e.g. for disclosure reasons, on a cross-classification of demographic variables in each area.

A major pitfall in the precision estimation for the small area methods is that some of them require resampling methods, such as the parametric bootstrap. In the event of a large number of areas, the parametric bootstrap is very time costly. Therefore, in Chapter Four, the use of variance reduction methods is proposed. The proposed methods are then analysed exemplary for the so-called Fay-Herriot model in a model based simulation study, in order to visualize in which populations this method will work. It will be shown that this approach can reduce the computational burden for parametric bootstrap MSE estimates by a vast amount, depending on the population at hand.

In Chapter Five, the different estimators are analysed under the aforementioned aspects. First, a brief introduction on the differences between model and design based Monte Carlo simulations is given and a systematization is proposed. Next, the classical measures used in Monte Carlo studies in the survey context for the evaluation of the estimators are presented. Further, an additional measure is proposed, which in Monte Carlo studies gives another view on the precision of the point estimates. Finally, two major design based simulations are performed. First, the question is tackled as to how small an area may be in order for the estimators to provide acceptably precise point estimates. Therefore, four scenarios are built and compared with each other. Also, the variance and MSE estimators are discussed for a selection of interesting point estimates and scenarios. Second, the

applicability of the method for using intermediate level register information proposed in Section Three is studied.

The last chapter points out the findings of this thesis and discusses them controversially. Also, an outlook is given on further and already ongoing research, as well as possible future research directions.

Chapter 2

Regression Models for Small Area Estimation

The basis of small area estimators are regression models. These models are used to explain the variability of the variable of interest using covariates. The basic model used for this is the general linear model, which is explained in the next section. An extension to the general linear model, the general mixed linear model allows for more complex data situations, including clustering and complex correlation structures and will be explained in Section 2.2. Often the variable of interest is not continuous, such that linear models may not be suitable. In this case generalized linear models (Section 2.3) and generalized linear mixed models (Section 2.4) may be applied.

2.1 The General Linear Model

The linear regression is a tool to determine the linear relationship between a dependent random variable y and one or more independent random variables x. This method dates back to the 19th century, where Legendre (1805) and Gauss (1809) developed the least squares method in order to determine the solar orbit of the planets. This linear relation is expressed as

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p + e = x\beta + e \quad .$$

where $x = (1, x_1, \ldots, x_p)$ is the $n \times p+1$ design matrix containing explanatory covariates, y is the dependent variable of interest, and e is a residual error term containing the variation

of y that is not explained by x. The least squares solution to this equation consists in finding the vector β, which produces the minimal sum of squares of the residuals e.

$$\widehat{\beta} = \underset{\beta=(\beta_0,\ldots,\beta_p)}{\operatorname{argmin}} \sum_{i=1}^{n} e_i^2 = \underset{\beta=(\beta_0,\ldots,\beta_p)}{\operatorname{argmin}} \; (y - x\beta)'(y - x\beta) \quad . \tag{2.1}$$

By multiplying out the part to be minimized,

$$(y - x\beta)'(y - x\beta) = -y'y - \beta'(x'x)\beta + 2y'x\beta \quad ,$$

and taking the derivative with respect to β, this yields

$$\frac{d}{d\beta} - y'y - \beta'(x'x)\beta + 2y'x\beta = -2(x'x)\beta + 2y'x = 2(y'x - (x'x)\beta) \quad .$$

Equating this derivative to zero and solving for β, the least squares estimate for β results in

$$\widehat{\beta} = (x'x)^{-1}x'y \quad . \tag{2.2}$$

The least squares estimate for β is at the same time the maximum likelihood estimate assuming normality of the error terms and the generalized method of moments estimator (Verbeek, 2005, p. 183). There exist many different representations of the model assumptions underlying the linear regression model. A few examples of different representations of the assumptions can be found in Poole and O'Farrell (1971), von Auer (2011, § 2) and Dougherty (2011, § 2.2). In particular, it is assumed that each observation is a realisation from a distribution with constant variance over all units in the population. That is, the error term has expectation $\mathrm{E}[e_i] = 0$ and variance $\sigma_{e_i}^2$, which is identical for all observations $\sigma_{e_i}^2 = \mathrm{E}\left[e_i^2\right] \equiv \sigma_e^2, i = 1, \ldots, n$. It is also assumed that the error term is independently distributed with no form of autocorrelation, that is $\mathrm{COV}\left[e_i, e_j\right] = 0, i \neq j$.

If the observations shall have a unit-specific importance for the estimation of the model parameters, weights may be introduced to this linear regression. This can be done by rewriting the Equation (2.1) into

$$\widehat{\beta}_w = \underset{\beta=(\beta_0,\ldots,\beta_p)}{\operatorname{argmin}} \sum_{i=1}^{n} w_i e_i^2 = \underset{\beta=(\beta_0,\ldots,\beta_p)}{\operatorname{argmin}} \; (y - x\beta)'W(y - x\beta) \quad , \tag{2.3}$$

with w_i being the weight attributed to observation i and W being an $n \times n$ diagonal matrix with the i-th diagonal element being w_i (Björck, 1996, § 4.4). The weighted least squares solution then is

$$\widehat{\beta}_w = (x'Wx)^{-1}x'Wy \quad . \tag{2.4}$$

For a more in-depth discussion on numerical issues concerning the weighted least squares estimate see Björck (1996, § 4.4) and the references therein.

2.2 The General Linear Mixed Model

The general linear model has some very restrictive assumptions. For example, in small area estimation applications, the assumption that all observations are independent is usually not reasonable. In practice, the sampling is done for each area separately, and usually it is assumed that the areas are different in some sense, e.g. with distinct mean and / or variability. Thus, for small area estimation, it is more reasonable to assume cluster structures in the correlation of the observations. The general linear mixed model allows for this and many other correlation structures in the variance covariance matrix and therefore is very useful for small area estimation.

The mixed model formulation is similar to the linear regression (Pinheiro & Bates, 2000, § 5.1)

$$y = x\beta + zu + e \tag{2.5}$$

The difference lies in an additional term zu, which is due to the K independent random effects u. $z = (z_1, \ldots, z_K)$ is a known matrix of size $n \times \sum_{k=1}^{K} \nu_k$ with z_k being of size $n \times \nu_k$ and of ranks ν_k. $u = (u_1', \ldots, u_K')'$ is a $\sum_{k=1}^{K} \nu_k \times 1$ vector having a multivariate normal distribution with means zero and non-singular $\sum_{k=1}^{K} \nu_k \times \sum_{k=1}^{K} \nu_k$ variance covariance matrix Σ_u.

$$u \sim \mathrm{N}(0, \Sigma_u) \tag{2.6}$$

And e is the $n \times 1$ residual vector having a multivariate normal distribution with means zero and non singular variance covariance matrix Σ_e of size $n \times n$:

$$e \sim \mathrm{N}(0, \Sigma_e) \quad . \tag{2.7}$$

Alternatively, the distribution of y may be expressed as

$$y \sim \mathrm{N}(x\beta + zu, \Sigma_e) \quad , \tag{2.8}$$

or

$$y \sim \mathrm{N}(x\beta, \Sigma_e + z\Sigma_u z') \quad . \tag{2.9}$$

Generally, the assumptions of normality may be dropped, but in the small area context normality is usually assumed.

By setting z as a null matrix, e.g. in equation (2.9), one can see that the linear regression is a special case of the mixed model with $\Sigma_e = \sigma_e^2 I_{(n)}$. Similarly, if Σ_u is, or is almost, a null matrix, the same applies. Therefore, in the case where there is no random effect or only a weak random effect is observed, the mixed model approaches the linear regression model.

Henderson (1950) shows, that if Σ_u and Σ_e are known, then the best linear unbiased estimator $\widehat{\beta}$ for β and the best linear unbiased predictor $\widehat{u}$ for u exist. Some different approaches to prove this can be found in Harville (1990); Jiang (1997) and Schmid (2011). This result can also be derived from a Bayesian perspective as can be seen in Dempfle (1977); Lindley and Smith (1972) and a comprehensive comparison of the frequentist and Bayesian approach is presented by Vogt (2007).

Henderson (1950) showed that $\widehat{\beta}$ and $\widehat{u}$ can be obtained by solving the following equation:

$$\begin{pmatrix} x'\Sigma_e^{-1}x & x'\Sigma_e^{-1}z \\ z'\Sigma_e^{-1}x & z'\Sigma_e^{-1}z+\Sigma_u^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{u} \end{pmatrix} = \begin{pmatrix} x'\Sigma_e^{-1}y \\ z'\Sigma_e^{-1}y \end{pmatrix} \quad . \tag{2.10}$$

The solution is given by the weighted least squares estimate of β

$$\widehat{\beta} = (x'V^{-1}x)x'V^{-1}y \quad , \tag{2.11}$$

and

$$\widehat{u} = \Sigma_u z'V^{-1}(y - x\widehat{\beta}) \tag{2.12}$$

(Henderson, 1963).

The resulting log-likelihood for the maximum likelihood and residual maximum likelihood are respectively

$$\text{ML:} \quad l_{ML}(\phi,\beta) = -\frac{1}{2}\Big(\log(2\pi) + \log|V| + e'V^{-1}e\Big) \quad , \tag{2.13}$$

$$\text{REML:} \quad l_{REML}(\phi,\beta) = -\frac{1}{2}\Big((n-p)\log(2\pi) + \log|V| + e'V^{-1}e + \log|x'V^{-1}x|\Big) \quad , \tag{2.14}$$

$V = V(\phi) = \Sigma_e + z\Sigma_u z'$ being the variance covariance matrix of y under the linear mixed model, ϕ is the vector of variance components $\phi = (\Sigma_e, \Sigma_u)$, and the residuals $e = y - (x\beta + zu)$.

As Σ_u and Σ_e are usually not known, they have to be estimated as well. By plugging the estimated $\widehat{\Sigma}_u$ and $\widehat{\Sigma}_e$ into equations (2.11) and (2.12), one obtains the *empirical* best linear unbiased predictor $\widehat{u}$ for u (Kackar & Harville, 1981). Kackar and Harville (1981) and Jiang (2000) also show that if the data is symmetrically distributed and $\widehat{\phi}$ are estimated

with an even and translation invariant estimator, the EBLUP remains unbiased. As Jiang and Lahiri (2006b) state, ML and REML estimates fulfil these requirements.

Skinner (1989) proposes a method for incorporating survey weights which is similar to the one in Equation (2.3). However, survey weighted mixed models for complex surveys are a broad subject that would go beyond the scope of this work. For some papers giving insight into the topic see e.g., Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998), Rabe-Hesketh and Skrondal (2006) and Carle (2009). The sampling design needed in the design based simulations in Chapter 5 remains simple enough, as to allow for using a simple weighting of the model, which is analogous to the one in Equation (2.3).

2.2.1 Review of General Linear Mixed Model Estimation Methods

Ample literature exists on different estimation methods for general linear mixed models. In this section, a brief overview of the core directions is presented, including some findings about the performance of these methods.

Henderson (1953) proposed three methods for estimating the variance components, of which the third is being extended by many different authors (Pérez, Peña, & Molina, 2011; Sarraj & Rosen, 2009). The three methods may be classified in the words of Djordjević and Lepojević (2003):

> Method 1 is simply an analogue of the analysis of variance method used with balanced data; Method 2 is designed to correct a deficiency of Method 1 that arises with mixed models; and Method 3 is based on the method of fitting constants so often used in fixed effects models.
>
> (see Djordjević & Lepojević, 2003, § 6, p. 63)

C. R. Rao (1970, 1971a, 1971b, 1972) proposed non-iterative variance components estimation methods called MINQUE (Minimum Norm Quadratic Unbiased Estimation) and MIVQUE (Minimum Variance Quadratic Unbiased Estimation of Variance Components). However, in practice they are often used iteratively (C.-t. Wu, Gumpertz, & Boos, 2001). In this case, Harville (1977) notes that MINQUE/MIVQUE give identical results to REML under the assumption of normality of u and e. Further, he argues that as MINQUE does not need any distributional assumptions, Gaussian REML could be used, even if the data is not normally distributed. P. S. R. S. Rao (1977) gives an overview of the MINQUE method and Kleffe and Seifert (1986) show how to handle it computationally.

Many authors, such as Giesbrecht and Burrows (1978); Swallow and Monahan (1984) and Harville (1977) compared the different methods. Essentially, they find that if REML is computational feasible it gives comparably good results. One problem of REML estimates is that they may result in zero-estimates. H. Li and Lahiri (2010) propose an

adjusted maximum likelihood estimator of the model variance. It maximizes an adjusted likelihood, which is defined as a product of the model variance and a standard likelihood function, e.g. a profile or a restricted maximum likelihood. The main advantage of this approach is that it yields strictly positive $\widehat{\sigma}_u$, which is necessary later on for the construction of the small area estimates. Further, H. Li and Lahiri (2010) state that this method has better small sample properties than REML, which in the case of small area estimation is an important point. As in the Swiss structural survey the sample size is 200,000 and zero-estimates for the variance components almost never occur, in the following mainly REML estimates will be used.

Lindstrom and Bates (1988) suggest applying the Newton-Raphson algorithm using a QR-decomposition approach. A well known drawback of Newton-based iteration algorithms is that convergence is not guaranteed. It is even possible to obtain parameter estimates which are out of the parameter space, such as negative variances. Harville (1977) also compares the Newton-Raphson algorithm with the Fisher Scoring algorithm proposed by Patterson and Thompson (1971) and speeded up by Longford (1987). He states, that as the expected Hesse matrix may be easier to compute than the observed one, the Fisher-Scoring algorithm may be faster than the Newton-Raphson algorithm, with the drawback that sometimes more iterations are necessary. Knight (2008) finds that Newton-type algorithms do a reasonably good job in cases where the starting values are already close to the final ML or REML estimates. However, if the starting values are poor, e.g. far away from the final ML or REML estimates, EM-algorithm based iteration methods will perform much better. The main advantage of the EM-algorithm based methods is that convergence is assured. However, since the EM-algorithm can only detect a local maximum, the convergence to the global maximum is not guaranteed.

In reply to the question of whether ML or REML is to be used, Searle, Casella, and McCulloch (1992) give the following advice:

> As to the question "ML or REML?" there is probably no hard and fast answer. Both have the same merits of being based on the maximum likelihood principle – and they have the same demerit of computability requirements. ML provides estimators of fixed effects, whereas REML, of itself, does not. But with balanced data REML solutions are identical to ANOVA estimators which have optimal minimum variance properties – and to many users this is a sufficiently comforting feature of REML that they prefer it over ML.
>
> (Searle et al., 1992, § 6.8, p. 255)

In the next sections, the Fisher-Scoring Algorithm for ML and REML estimation and the EM Algorithm for ML estimation of mixed model parameters are discussed.

2.2.2 Estimating the General Linear Mixed Model with the Fisher-Scoring Algorithm

One popular approach to obtain the variance components is to use the Fisher-Scoring algorithm (c.f. A.3). This algorithm needs the Jacobian and the Hessian matrix of the likelihood and is described in 2.1. First some starting values $\beta^{(0)}$ and $\phi^{(0)}$ have to be set. This point is crucial, as the starting values should be as near as possible to the true values. Otherwise convergence may be slow or the algorithm might not even converge at all (Knight, 2008). Cook (1982); Laird, Lange, and Stram (1987) propose to use the following starting values:

$$\beta^{(0)} = (x'x)^{-1}x'y \tag{2.15}$$

$$u_d^{(0)} = (z_d'z_d)^{-1}z_d'(y_d - x_d\beta^{(0)}) \tag{2.16}$$

$$\Sigma_e^{(0)} = \sigma_e^{2,(0)}I_{(n)} \tag{2.17}$$

$$\sigma_e^{2,(0)} = \frac{\sum_{d=1}^{D} y_d'y_d - \beta^{(0)}\sum_{d=1}^{D} x_d'y_d - \sum_{d=1}^{D} u_d'z_d'(y_d - x_d\beta^{(0)})}{n-(D-1)\nu - p - 1} \tag{2.18}$$

and

$$\Sigma_u^{(0)} = \frac{1}{D}\left(\sum_{d=1}^{D}(u_d^{(0)})'u_d^{(0)} - \sigma_e^{2,(0)}\sum_{d=1}^{D}(z_d'z_d)^{-1}\right) \tag{2.19}$$

Second, the variance covariance matrix V is built in order to, third, calculate the new fixed effects vector β.

As both ML and REML are considered, both log-likelihoods, ML and REML are considered as well. Given this new fixed effects vector β and the old variance components ϕ, the variance components are updated in step four. Steps two to four are repeated sequentially, until convergence of the parameters β and ϕ.

Algorithm 2.1 ML and REML Estimation of Variance Components

The elements of the Jacobian are

$$\frac{d\,l_{ML}}{d\,\phi_l} = J_{(l)}^{\mathrm{ML}}(\beta,\phi) = -\frac{1}{2}\left(\mathrm{tr}\left[V^{-1}V_{(l)}\right] + (y-x\beta)'V^{-1}V_{(l)}V^{-1}(y-x\beta)\right)$$

$$\frac{d\,l_{REML}}{d\,\phi_l} = J_{(l)}^{\mathrm{REML}}(\beta,\phi) = -\frac{1}{2}\left(\mathrm{tr}\left[PV_{(l)}\right] - y'PV_{(l)}Py\right)$$

with $P = V^{-1} - V^{-1}x(x'V^{-1}x)^{-1}x'V^{-1}$, $V_{(l)} = dV/d\,\phi_l$, and $\phi = (\Sigma_e, \Sigma_u)$

The elements of the Hessian are

$$\frac{d^2 l_{ML}}{d\phi_l d\phi_k} = H^{\mathrm{ML}}_{(l,k)}(\beta,\phi) = \frac{1}{2}\left(V^{-1}V_{(l)}V^{-1}V_{(k)}\right)$$

$$\frac{d^2 l_{REML}}{d\phi_l d\phi_k} = H^{\mathrm{REML}}_{(l,k)}(\beta,\phi) = \frac{1}{2}\left(PV_{(l)}PV_{(k)}\right)$$

1. Set starting values $j = 0$ and $\beta^{(0)}_{\mathrm{ML}}$, $\beta^{(0)}_{\mathrm{REML}}$, $\phi^{(0)}_{\mathrm{ML}}$, and $\phi^{(0)}_{\mathrm{REML}}$.
2. Increase j by one and build

   ML $V_{\mathrm{ML}}(\phi^{(j-1)}_{\mathrm{ML}}) = \Sigma^{\mathrm{ML}(j-1)}_e + z\Sigma^{\mathrm{ML}(j-1)}_u z'$.

   REML $V_{\mathrm{REML}}(\phi^{(j-1)}_{\mathrm{REML}}) = \Sigma^{\mathrm{REML}(j-1)}_e + z\Sigma^{\mathrm{REML}(j-1)}_u z'$.
3. Calculate new fixed effects vector β

   ML $\beta^{(j)}_{\mathrm{ML}} = (x'V^{-1(j-1)}_{\mathrm{ML}}x)^{-1}x'V^{-1(j-1)}_{\mathrm{ML}}y$

   REML $\beta^{(j)}_{\mathrm{REML}} = (x'V^{-1(j-1)}_{\mathrm{REML}}x)^{-1}x'V^{-1(j-1)}_{\mathrm{REML}}y$
4. Update the variance components vector ϕ

   ML $\phi^{(j)}_{\mathrm{ML}} = \phi^{(j-1)}_{\mathrm{ML}} + \left(H^{\mathrm{ML}}_{(l,k)}(\beta^{(j-1)}_{\mathrm{ML}},\phi^{(j-1)}_{\mathrm{ML}})\right)^{-1} J^{\mathrm{ML}}_{(l)}(\beta^{(j-1)}_{\mathrm{ML}},\phi^{(j-1)}_{\mathrm{ML}})$

   REML
   $\phi^{(j)}_{\mathrm{REML}} = \phi^{(j-1)}_{\mathrm{REML}} + \left(H^{\mathrm{REML}}_{(l,k)}(\beta^{(j-1)}_{\mathrm{REML}},\phi^{(j-1)}_{\mathrm{REML}})\right)^{-1} J^{\mathrm{REML}}_{(l)}(\beta^{(j-1)}_{\mathrm{REML}},\phi^{(j-1)}_{\mathrm{REML}})$
5. repeat steps 2-4 until convergence of $\phi^{(j)}$

(c.f. J. N. K. Rao, 2003, § 6.2.4 and Searle et al., 1992, § 6)

Two concurrent convergence criteria are presented in the following section. They may be applied as well for the Fisher-Scoring algorithm.

2.2.3 Estimating the General Linear Mixed Model with the EM-Algorithm

The first theoretical foundations for the EM-algorithm were formulated by Orchard and Woodbury (1972). Dempster, Laird, and Rubin (1977) generalize this approach, name it, and provide some general proof. The basic concept is, that in the presence of missing data, the likelihood is often impossible to maximize. Therefore, instead of maximizing the likelihood directly, an iterative procedure is applied (Dempster et al., 1977):

1. Estimate the expected likelihood Q given the estimated parameters of the likelihood of the preceding step.
2. Maximize the likelihood given the estimated missing data.

In Algorithm 2.2 the general formulation of the EM-algorithm is given.

Algorithm 2.2 The general EM-algorithm

1. E-step: Obtain a new $Q(\psi; \psi^{(t-1)})$

$$Q(\psi; \psi^{(t-1)}) = \mathrm{E}\left[L(\psi; y)|\psi^{(t-1)}\right]$$

2. M-step: Maximize the Likelihood by choosing new ψ^t

$$\psi^t = \underset{\psi}{\operatorname{argmax}} Q(\psi; \psi^{(t-1)})$$

   where the new ψ^t must be in the parameter space and $Q(\psi^t; \psi^{(t-1)}) \geq Q(\psi; \psi^{(t-1)})$ for any ψ in the parameter space.

3. Repeat steps 1-2 until convergence.

(Dempster et al., 1977)

The EM-algorithm remains rather general, as it does not provide a closed form equation to solve a problem. It can be seen much more as a class of algorithms. A generalization of the EM algorithm was also proposed by Dempster et al. (1977), which relaxes the M-step a bit and is presented in algorithm 2.3. In contrast to the algorithm 2.2 it is not necessary to find the optimal parameter vector given the expected likelihood, but it is sufficient to find a new parameter vector in the parameter space which yields a higher likelihood than the one of the step before.

Algorithm 2.3 The generalized EM-algorithm

1. E-step: Obtain a new $Q(\psi; \psi^{(t-1)})$

$$Q(\psi; \psi^{(t-1)}) = \mathrm{E}\left[L(\psi; y)|\psi^{(t-1)}\right]$$

2. M-step: Maximize the Likelihood by choosing new ψ^t with the condition that ψ^t must be in the parameter space and

$$Q(\psi^t;\psi^{(t-1)}) \geq Q(\psi^{(t-1)};\psi^{(t-1)}) \quad .$$

It is sufficient to find this ψ^t such that

$$L(\psi^t;y) \geq L(\psi^{(t-1)};y)$$

3. Repeat steps 1-2 until convergence.

(Dempster et al., 1977)

Missing data does not mean necessarily unit or item non-response. It is meant in a much broader sense, such that basically everything may be included. For example, one could assume the random effects u to be the missing data. By applying this approach, mixed models may also be estimated by the EM-algorithm. Let ψ be the vector of unknown parameters of the assumed probability density function and the missing data, in this case β and u respectively. L denotes the joint Likelihood of the observed and missing data. In the case of models with exponential family distributions, it is enough to compute the expected sufficient statistics in the E-step (Dempster et al., 1977).

There are different proposals for convergence criteria. McLachlan and Krishnan (2007) for example choose to stop when

$$|L(\psi^t;y) - L(\psi^{(t-1)};y)| < \varepsilon \quad , \tag{2.20}$$

for a predefined arbitrary small ε. Foulley and Van Dyk (2000) choose to use instead the convergence criteria based on the estimated parameters ψ over the change in the likelihood. Their stopping criteria checks whether the following norm of ψ is less than 10^{-8}. Let K be the number of elements of ψ, then the norm is defined as

$$\sqrt{\frac{\sum_{k=1}^{K}(\psi_k^t - \psi_k^{(t-1)})^2}{\sum_{k=1}^{K}(\psi_k^t)^2}} \quad . \tag{2.21}$$

Special EM-algorithm implementations or the estimation of mixed models were proposed by many authors. In the original paper by Dempster et al. (1977) there is one application for mixed model estimation, but also Foulley and Van Dyk (2000); Laird et al. (1987); Lindstrom and Bates (1988); C. Liu, Rubin, and Wu (1998); Meza, Jaffrézic, and Foulley

(2007); van Dyk (2000) and Knight (2008), amongst many others, propose specialized EM-algorithms, some increasing speed, others stability.

For simplicity a very basic EM-algorithm will be presented, following Knight (2008, pp. 48).

Recalling the mixed model in equation (2.5) we choose an alternative representation,

$$y = x\beta + \sum_{k=1}^{K} z_k u_k + e \quad , \tag{2.22}$$

where y is again the $n \times 1$ vector of the observed variable of interest, x is the $n \times p$-matrix of covariates, z_k is a known design matrix for the k-th, $k = 1, \ldots K$, random effect structure of dimension $n \times \nu_k$, and β is the fixed effects parameter vector of size $n \times p$. The unobservable, or alternatively called missing values u_k, are the $\nu_k \times 1$ random effects vectors.

In analogy to the equation (2.9) the distribution of y can be written as

$$y \sim N(x\beta, V) \quad ,$$

where V in this formulation is $V = \sum_{k=1}^{K} z_k z_k' \sigma_k^2 + \sigma_e^2 I_n$.

Now we can write the mixed model as a missing data problem by assuming the u as missing. Then the mixed model is:

$$\begin{pmatrix} y \\ u_1 \\ \vdots \\ u_K \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} x\beta \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma \right) \tag{2.23}$$

with

$$\Sigma = \begin{pmatrix} Z\Sigma_u Z' + \Sigma_e & Z\Sigma_u \\ \Sigma_u Z' & \Sigma_u \end{pmatrix} = \begin{pmatrix} V & \mathrm{diag}_{k=1}^{K}\left(\sigma_k^2 z_k\right) \\ \mathrm{diag}_{k=1}^{K}\left(\sigma_k^2 z_k'\right) & \mathrm{diag}_{k=1}^{K}\left(\sigma_k^2 I_{\nu_k}\right) \end{pmatrix} \tag{2.24}$$

The density function of (y, u) is therefore

$$f(y,u) = \frac{1}{\sqrt{2\pi}^q} |\Sigma|^{-0.5} e^{-0.5 \cdot \kappa' \Sigma^{-1} \kappa} \quad , \tag{2.25}$$

with the column vector $\kappa = (\kappa_0, \kappa_1, \ldots, \kappa_K)$, $\kappa_0 = y - z\beta$, $\kappa_k = u_k$, $q_0 = n$, and $q = \sum_{k=0}^{K} \nu_k$ being the length of the vector κ. The resulting complete data log-likelihood is then

$$L^{(c)} = -0.5q\log(2\pi) - 0.5\sum_{k=0}^{K} \nu_k \log \sigma_k^2 - 0.5\sum_{k=0}^{K} \frac{u_k' u_k}{\sigma_k^2} \quad , \tag{2.26}$$

with $u_0 = e = y - x\beta - \sum_{k=1}^{K} z_k u_k$. Technically, as the $u_k, k = 0, \ldots, K$ are not observed, the $L^{(c)}$ is a log-density function, but for ease it will be called the complete data log-likelihood. The sufficient statistics for this model are then $(y - \sum_{k=1}^{K} z_k u_k)$, and $u_k' u_k$, $k = 1, \ldots, K$ (Dempster et al., 1977, pp. 17).

The resulting maximum likelihood estimates derived from the complete data log-likelihood are then (Dempster et al., 1977, p. 18)

$$\widehat{\sigma}_k^2 = \frac{u_k' u_k}{\nu_k}, k = 0, 1 \ldots K \tag{2.27}$$

$$\widehat{\beta} = (x'x)^{-1} x'(y - \sum_{k=1}^{K} z_k u_k) \tag{2.28}$$

As the $u_k, k = 0, 1, \ldots K$, are not observed, for the E-step of the EM-algorithm first the expected values of $u_k' u_k$ and $(y - \sum_{k=1}^{K} z_k u_k)$ given the observed data y have to be found.

Following Searle et al. (1992, § 8.3.b, p. 298) the distribution of the $u_k|y$ is a multivariate normal distribution of dimension ν_k given by

$$u_k|y \sim \mathrm{N}\left(\sigma_k^2 z_k' V^{-1}(y - x\beta), \sigma_k^2 I_{(\nu_k)} - \sigma_k^4 z_k' V^{-1} z_k\right) \quad . \tag{2.29}$$

and the expectations of the conditional sufficient statistics for the complete data log-likelihood are given by

$$\mathrm{E}\left[u_k' u_k | y\right] = \sigma_k^4 (y - x\beta)' V^{-1} z_k z_k' V^{-1} (y - x\beta) + tr(\sigma_k^2 I_{(\nu_k)} - \sigma_k^4 z_k' V^{-1} z_k) \quad , \tag{2.30}$$

and

$$\mathrm{E}\left[(y - \sum_{k=1}^{K} z_k u_k)|y\right] = x\beta + \sigma_0^2 V^{-1}(y - x\beta) \quad . \tag{2.31}$$

For $k = 0$, the individual error term, the matrix z_k is the identity matrix $I_{(n)}$. Therefore, $\mathrm{E}\left[u_0' u_0 | y\right]$ reduces to

$$\mathrm{E}\left[u_0' u_0 | y\right] = \sigma_0^4 (y - x\beta)' V^{-1} V^{-1} (y - x\beta) + tr(\sigma_0^2 I_{(n)} - \sigma_0^4 V^{-1}) \quad . \tag{2.32}$$

This yields the EM algorithm 2.4 as a combination of the presented elements. First, the expectation of the sufficient statistics is computed, given the estimated parameters and the observed data (the E-step), then the maximum likelihood estimates of the complete data are calculated.

Algorithm 2.4 An Elementary EM-algorithm for Maximum Likelihood Estimation for the Linear Mixed Model

Set $t = 0$ and starting values $\beta^{(0)}, \sigma_k^{2,(0)}, k = 0, 1, \ldots K$.

1. E-step: Obtain the sufficient statistics which are sufficient for $Q = L^{(c)}$:

$$\begin{aligned} \widehat{s}_k^{(t)} &= \sigma_k^{4,(t)} (y - x\beta^{(t)})' V^{-1,(t)} z_k z_k' V^{-1,(t)} (y - x\beta^{(t)}) \\ &\quad + tr(\sigma_k^{2,(t)} I_{(\nu_k)} - \sigma_k^{4,(t)} z_k' V^{-1,(t)} z_k) \\ \widehat{\kappa}^{(t)} &= x\beta^{(t)} + \sigma_0^{2,(t)} V^{-1,(t)} (y - x\beta^{(t)}) \end{aligned}$$

   where $V^{(t)} = \sum_{k=1}^{K} z_k z_k' \sigma_k^{2,(t)} + \sigma_e^{2,(t)} I_{(n)}$

2. M-step: Compute the maximum likelihood estimates of the complete data log-Likelihood $L^{(c)}$ given the estimated sufficient statistics of the E-step:

$$\begin{aligned} \sigma_k^{2,(t+1)} &= \frac{\widehat{s}_k^{(t)}}{\nu_k} \\ \beta^{(t+1)} &= (x'x)^{-1} x' \widehat{\kappa}^{(t)} \end{aligned}$$

3. Repeat steps 1-2 until convergence.

(Searle et al., 1992, § 8.3.c, Hartley & Rao, 1967, and Dempster et al., 1977)

For a REML version of this EM algorithm see Searle et al. (1992, § 8.3.f). This EM-algorithm is an elementary one, on which many other EM-algorithms for the estimation in mixed models are based. Amongst them are essentially the

ECM The ECM algorithm proposed by Meng and Rubin (1993) is an extension for the generalized EM algorithm, for the case that the M-step is computationally difficult to perform. In this case, the M-step can be split into multiple conditional maximization steps (CM-steps), which generally are computationally simpler to solve. This algorithm often takes more iteration steps than the elementary EM algorithm. Nevertheless, depending on the complexity of the M-step, it may outperform the EM algorithm in terms of computation time due to the simpler CM-steps

ECME C. Liu and Rubin (1994) propose an extension to the ECM algorithm. They call it *Expectation/Conditional Maximisation Either (ECME)* algorithm. Instead of maximizing the conditional expected complete-data log-Likelihood function (CM-steps), the corresponding conditional *actual* likelihood function is maximized. They show that their ECME algorithm is still monotone convergent, like the EM and ECM algorithms. Further it converges faster with respect to computation time and iteration steps than both the EM and the ECM algorithms.

PX-EM C. Liu et al. (1998) proposed the *Parameter Expanded EM algorithm (PX-EM)*. By parameter expanded it is meant that the observed parameter likelihood is expanded by some parameters to form a new complete data log-likelihood. The additional parameters are then used in a manner of covariance adjustment. C. Liu et al. (1998) state that their PX-EM algorithm is as stable and simple as the EM-algorithm, but faster. Foulley and Van Dyk (2000), among others, apply the PX-EM algorithm in the mixed model context.

2.3 The Generalized Linear Model

In the case that the dependent variable is not continuous, the linear regression may produce nonsensical results. Say, the dependent variable is dichotomous with values 0 and 1, then using a linear regression easily results in predictions of values of over 1 or under 0. As these predictions are out of the range of the support of the dependent variable, they are neither meaningful nor acceptable.

The exponential family of distributions allows for differently scaled random variables. Some exponential families are the Binomial, the Poisson and normal distribution families, amongst many others.

The general form for the probability density function belonging to an exponential family distribution is

$$f(y|\theta,\phi) := \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \quad , \tag{2.33}$$

where $b(\theta)$ must be two times differentiable with respect to θ and it has to be possible to normalize $f(y|\theta)$. ϕ is called the dispersion parameter and stands for the variability in the random variable. One can show that $\mathrm{E}(y) = \mu = b'(\theta)$ and $\mathrm{V}(y) = a(\phi)b''(\theta)$. The choice of the functions a, b and c leads to the different distribution families which are exponential families. Early works on this family of distributions can be found in Darmois (1935), Pitman (1936) and Koopman (1936).

Nelder and Wedderburn (1972) propose a generalization of the linear regression allowing for a more flexible set of scales of the dependent variable. These models consist of three components

1. An assumption about the distribution of the dependent variable y.
2. A so called *systematic component* that describes the way the covariates combine to form the linear predictor.
3. A link function g: $\mu \mapsto \eta$ of the form $g(\mu) = \eta$. A special case is the canonical link function, for which $g(\mu) = \theta$ must hold.

The classical linear regression is a special case of the generalized linear model. By choosing $\theta = \mu, \phi = \sigma^2, a(\phi) = \phi, b(\theta) = \theta^2/2$ and $c(y,\phi) = -(y^2/\sigma^2 + \log(2\pi\sigma^2))/2$ one obtains that the normal distribution is an exponential family. This can be seen by substituting these θ, ϕ, a, b, c into equation (2.33).

$$\begin{aligned} f(y|\mu,\sigma^2) &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad . \end{aligned}$$

As the normal distribution is an exponential family, one can apply the generalized linear model as follows:

1. y is assumed to be normally distributed. (At least $y|x$ is normally distributed.)
2. The systematic component in the linear regression is $\mu = x\beta$.
3. The canonical link function is the identity function, as $g(\mu) = \theta = \mu$.

The model is then

$$\begin{aligned} \mathrm{E}[y|x] &= \mu = X\beta \quad , \\ \mathrm{V}[y|x] &= \sigma^2 \quad , \end{aligned}$$

or alternatively written

$$y|x \sim \mathrm{N}(x\beta, \sigma^2) \quad ,$$

which is the linear regression model.

In the case of a dichotomous dependent variable, the so called *logit model* may be used. The dependent variable is taken to be distributed according to a Bernoulli distribution, which is also an exponential family of distributions. Therefore, it fits into the theory developed by Nelder and Wedderburn (1972).

1. The dependent variable is assumed to be Binomial distributed

$$f(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, y \in \{0,1,\ldots,n\} \tag{2.34}$$

   As can be seen, the Binomial distribution is an exponential family:

$$f(y) = \binom{n}{y} \exp\left(y \log\left(\frac{\theta}{1-\theta}\right) - n\log(1-\theta)\right) \quad , \tag{2.35}$$

   where $\theta = \log(\theta/(1-\theta))$, $b(\theta) = n\log(1+\exp(\theta))$, $a(\phi) = 1$ and $c = \log\binom{n}{y}$. Further, $\mathrm{E}[y|x] = b'(\theta) = np$, $\mathrm{V}[y|x] = a(\phi)b''(\theta) = n\theta(1-\theta)$ holds, with $b'(\theta) = \exp(\theta)/(1+\exp(\theta))$ and $b''(\theta) = \exp(\theta)/(1+\exp(\theta))^2$.

2. The linear systematic component $\eta = x\beta$ is assumed:.

3. The canonical link is in this case the *logit*-function

$$g(\theta) = \mathrm{logit}(\theta) := \log\left(\frac{\theta}{1-\theta}\right) \quad , \tag{2.36}$$

   and its inverse function is

$$g^{-1}(\eta) = \mathrm{logit}^{-1}(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}} = \theta \quad . \tag{2.37}$$

The parameters of the generalized linear model may be estimated by maximum likelihood. The log-likelihood of the generalized linear model can be written in its general form as:

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta) \quad , \tag{2.38}$$

where

$$l_i(\beta) = \log(f(y_i|x)) = \frac{y_i\theta_i - b(\theta)}{a(\phi)} \tag{2.39}$$

denotes the likelihood contribution of the i'th observation. In this case, the $i = 1,\ldots,n$ observations are assumed to be independent. The term $c(y,\phi)$ is omitted in equation (2.39), because it is constant in θ, and thus also constant in β. Therefore, it plays no role in the maximization process of the likelihood.

In most generalized linear models there are no closed form solutions for the maximum likelihood parameter estimate. An exception here is the linear regression model (see section 2.1).

However, iterative solutions exist. The major statistical software, such as SAS, SPSS, Stata, S+ and R use ML estimation for obtaining the model parameters. This is done via the iteratively reweighted least squares, Newton-Raphson or Fisher-Scoring algorithm. For the Newton-Raphson and Fisher-Scoring algorithm see page 139. A comparison of these approaches is presented by Green (1984).

Simonoff (2003, § 5.1.2) presents a neat approach to the iteratively reweighted least squares algorithm. By taking the derivative of the log-likelihood in Equation (2.39), one obtains the score function. Setting the score function equal to zero yields the parameters β, which maximizes the log-likelihood.

$$\frac{\partial l(\beta)}{\partial \beta} = s(\beta) = \sum_{i=1}^{n} x_i \frac{d_i^2}{\sigma_i^2}(y_i - \mu_i) = 0 \quad , \tag{2.40}$$

with $\sigma_i^2 = \mathrm{V}[y_i|x_i]$ and $d_i = \partial \mu_i / \partial \eta_i$. In matrix form this equation can be written as

$$x' W((y-\mu) \odot J) = 0 \quad , \tag{2.41}$$

where $W = \mathrm{diag}\left(d_i^2/\sigma_i^2\right)$, $J = (\partial \eta_1/\partial \mu_1, \ldots, \partial \eta_n/\partial \mu_n)$, and $\odot$ is the element wise multiplication.

Now a neat trick is employed. On both sides of the equation the term $x'Wx\beta$ is added.

$$\begin{aligned} x'Wx\beta &= x'Wx\beta + x'W((y-\mu) \odot J) \quad , \\ <=> x'Wx\beta &= x'W\zeta \quad , \end{aligned}$$

with $\zeta = x\beta + ((y-\mu) \odot J)$. Therefore, the solution to the score equation can be seen as a weighted least squares problem with the solution

$$\widehat{\beta} = (x'Wx)^{-1} x'W\zeta \tag{2.42}$$

But as ζ and W are functions of β, the solution has to be found iteratively.

$$\widehat{\beta}^{(j+1)} = \left(x'W(\widehat{\beta}^{(j)})x\right)^{-1} x'W(\widehat{\beta}^{(j)})\zeta(\widehat{\beta}^{(j)}) \tag{2.43}$$

This leads to the iteratively reweighted least squares algorithm 2.5, which yields results identical to the Fisher-Scoring algorithm (Simonoff, 2003, p. 128).

Algorithm 2.5 Iteratively Reweighted Least Squares for Generalized Linear Models

1. Set starting values for $\beta^{(0)}$ and set $j = 1$.
2. Calculate $\eta^{(j)} = x\beta^{(j-1)}$ and $\mu^{(j)} = g^{-1}(\eta^{(j)})$.
3. Calculate $\sigma_i^{2,(j)} = \mathrm{V}[y_i|x_i]$, $d_i^{(j)} = \dfrac{\partial \mu_i^{(j)}}{\partial \eta_i^{(j)}}$, and $J^{(j)} = \left(\partial \eta_1^{(j)}/\partial \mu_1^{(j)}, \ldots, \partial \eta_n^{(j)}/\partial \mu_n^{(j)}\right)$
4. Compute $W^{(j)} = \mathrm{diag}\left(\dfrac{d_i^2}{\sigma_i^2}\right)$ and $\zeta^{(j)} = x\beta^{(j-1)} + ((y - \mu^{(j)}) \odot J^{(j)})$
5. Obtain new β: $\beta^{(j)} = \left(x'W^{(j)}x\right)^{-1} x'W^{(j)}\zeta^{(j-1)}$
6. Increase j by one.
7. Repeat steps 2-6 until convergence.

(Simonoff, 2003, § 5.1.2)

For many exponential families, the partial derivatives may be derived analytically. In this thesis, the Binomial generalized model with logit link is used and, hence, the derivatives will be given by:

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} \quad , \tag{2.44}$$

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i(1-\mu_i)} \quad , \tag{2.45}$$

and

$$\mathrm{V}[y_i|x_i] = np(1-p) \quad . \tag{2.46}$$

2.4 The Generalized Linear Mixed Model

In analogy to the pure fixed effects model, non-normal data may not be modelled meaningfully with the general linear mixed model approach. After the proposition of Nelder

and Wedderburn (1972), soon many authors extended the idea of generalized linear models to include random effects, especially in a longitudinal context (see e.g. Liang & Zeger, 1986; Zeger, Liang, & Albert, 1988). A more general framework was developed by Schall (1991) and Wolfinger and O'Connell (1993). As in the generalized linear model, three components of the generalized mixed model have to be chosen: (1) the assumption about the distribution of the dependent variable y; (2) the systematic component; and (3) the link function g. The difference with respect to the generalized linear model lies mainly in the systematic component, which is extended by the random effects. The distributional assumption and the link function remain basically the same.

For computational reasons, the notation of the generalized mixed model is often as follows (Schall, 1991):

$$y = \mu + e \quad , \tag{2.47}$$

where e is an error term and $\mathrm{E}(y) = \mu$. The link function g is used to map $\mu \to \mathbb{R}$. $\eta = g(\mu)$ is called the linear predictor, which is assumed to depend on x and z in the following form:

$$g(\mu) = \eta = x\beta + zu \quad . \tag{2.48}$$

Again, β is a vector of fixed effects parameters and u a vector of the random effects. e and u are assumed to be independent and independently distributed with $e \sim \mathrm{N}(0, \Sigma_e)$ and $u \sim \mathrm{N}(0, \Sigma_u)$. The conditional expectation of $y|u$ is

$$\mathrm{E}[y|u] = g^{-1}(x\beta + zu) \quad . \tag{2.49}$$

In the case of y being normally distributed and the link function is the identity function, the generalized mixed model equals the linear mixed model. However, for every other link the resulting likelihood involves an analytically intractable integral

$$L(\psi|y) = \int_{-\infty}^{+\infty} L(\psi|y,u) f(u) d\,u \tag{2.50}$$

$$L(\psi|y,u) = \prod_{i \in s} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y,\phi) \right) \tag{2.51}$$

and as $c(y,\phi)$ is independent from ψ

$$l(\psi|y,u) = \sum_{i \in s} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} \propto \log(L(\psi|y,u)) \tag{2.52}$$

by substitution equation (2.52) into (2.50) one obtains equation (2.53)

$$L(\psi|y) \propto \int_{-\infty}^{+\infty} \exp(l(\psi|y,u)) f(u) d\,u \quad . \tag{2.53}$$

$f(u)$ denotes the probability density function of the random effects, which can be univariate or multivariate. Usually, a normally distributed random effect is used as stated above. This approach may however be extended to non-normal random effects.

There exist a wide variety of methods for maximizing the likelihood (2.53) for generalized linear mixed models. Rabe-Hesketh, Skrondal, and Pickles (2002) and Bolker et al. (2009) state that the most commonly used methods are the penalized quasi-likelihood (PQL: see e.g. Laird, 1978; Stiratelli, Laird, & Ware, 1984) and marginal quasi-likelihood (MQL: see e.g. Goldstein, 1991).

With the rise of faster computers, Monte-Carlo-based methods have also become increasingly attractive. Examples of this are MCMC-EM algorithms (Booth & Hobert, 1999; Delyon, Lavielle, & Moulines, 1999; Kuhn & Lavielle, 2004), simulated maximum likelihood approaches, like the one proposed by Concordet and Nunez (2002) or the simulated method of moments developed by Jiang (1998). Here, the most commonly used method, PQL, is presented via the approach developed by Breslow and Clayton (1993). Their main idea is to apply the Laplace Approximation to a quasi-likelihood. The quasi-likelihood they choose is

$$\exp(\mathrm{ql}(\psi)) \propto |\Sigma_u|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i(u)) - \frac{1}{2} u' \Sigma_u^{-1} u \right) d\,u \quad , \tag{2.54}$$

with the d_i as the (scaled) deviance measure of fit of the form

$$d_i(y, \mu) := -2 \int_{y}^{\mu} \frac{y-u}{a_i \mathrm{v}(u)} d\,u \quad , \tag{2.55}$$

with $\mathrm{V}[y_i|u] = \phi a_i \mathrm{v}(u)$. If the $y|u$ are realisations of a linear exponential family with variance function $\mathrm{v}(\cdot)$, then this quasi-likelihood resembles the true likelihood. As stated, Breslow and Clayton (1993) propose to apply a Laplace approximation to equation (2.54). By rewriting equation (2.54) into

$$\exp(\mathrm{ql}(\psi)) \propto \int_{-\infty}^{+\infty} \exp(\kappa(u)) d\,u \quad , \tag{2.56}$$

the Laplace approximation to ql is

$$\mathrm{ql}(\psi) \approx -\frac{1}{2} \log|\Sigma_u| - \frac{1}{2} \log|\kappa''(\widetilde{u})| - \kappa(\widetilde{u}) \quad , \tag{2.57}$$

with $\widetilde{u} = \widetilde{u}(\psi)$ being the solution to

$$\kappa'(u) = -\sum_{i=1}^{n} \frac{(y_i - \mu_i(u)) z_i}{\phi a_i \mathrm{v}(\mu_i(u)) g'(\mu_i(u))} + \Sigma_u^{-1} u = 0 \quad . \tag{2.58}$$

Hence, $\widetilde{u}$ minimizes $\kappa(u)$ (Breslow & Clayton, 1993).

The second derivative of $\kappa(u)$ with respect to u is

$$\kappa''(u) = -\sum_{i=1}^{n} \frac{z_i z_i'}{\phi a_i \mathrm{v}(\mu_i(u))[g'(\mu_i(u))]^2} + \Sigma_u^{-1} + R = z'Wz + \Sigma_u^{-1} + R \quad , \tag{2.59}$$

where

$$R = -\sum_{i=1}^{n} (y_i - \mu_i(u)) z_i \frac{\partial}{\partial u} \left[\frac{1}{\phi a_i \mathrm{v}(\mu_i(u)) g'(\mu_i(u))} \right] \tag{2.60}$$

is a remainder term with expectation zero and equals zero in the case of a canonical link function (Breslow & Clayton, 1993). The $n \times n$ matrix W is a diagonal matrix with elements

$$w_{ii} = \frac{1}{\phi a_i \mathrm{v}(\mu_i(u))[g'(\mu_i(u))]^2} \quad . \tag{2.61}$$

Substituting equations (2.54) and (2.59) into equation (2.57) yields

$$\mathrm{ql}(\psi) \approx -\frac{1}{2} \log |I + z'Wz\Sigma_u| - \frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i(\widetilde{u})) - \frac{1}{2} \widetilde{u}' \Sigma_u^{-1} \widetilde{u} \quad , \tag{2.62}$$

with $\widetilde{u}$ being the vector that maximizes $-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i(u)) - \frac{1}{2} u' \Sigma_u^{-1} u$. Breslow and Clayton (1993) assume that the weights in W do not vary, or vary only very little with the change of the mean $\mu_i(u)$, and, therefore, neglect the first term in equation (2.62). By doing this, they obtain the penalized quasi-likelihood proposed by Green (1987) for semi-parametric regression models

$$\mathrm{ql}(\psi) \approx -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i(\widetilde{u})) - \frac{1}{2} \widetilde{u}' \Sigma_u^{-1} \widetilde{u} \quad . \tag{2.63}$$

Differentiating equation (2.63) with respect to β and u leads to the score equations

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i(u)) x_i}{\phi a_i \mathrm{v}(\mu_i(u)) g'(\mu_i(u))} = 0 \quad , \tag{2.64}$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i(u)) z_i}{\phi a_i \mathrm{v}(\mu_i(u)) g'(\mu_i(u))} = \Sigma_u^{-1} u \quad , \tag{2.65}$$

which can be solved with the Fisher scoring algorithm proposed by Green (1987). Breslow and Clayton (1993) prefer to base the iterative algorithm on the results of Harville (1977). First they define a *working variable* $\zeta_i = \eta_i(u) + (y_i - \mu_i(u)) g'(\mu_i(u))$, where $\eta_i(u) = X_i\beta + z_i u$. In analogy to Harville (1977) they propose to use the solution to the system of equations:

$$\begin{pmatrix} x'Wx & xWzD \\ z'Wx & I + z'WzD \end{pmatrix} \begin{pmatrix} \beta \\ \kappa \end{pmatrix} = \begin{pmatrix} x'Wy \\ z'Wy \end{pmatrix} \quad , \tag{2.66}$$

as estimators with $\widehat{\beta} = \beta$ and $u = \Sigma_u \kappa = \Sigma_u z' V^{-1}(y - x\widehat{\beta})$. V is the $n \times n$ variance covariance matrix $V = W^{-1} + z\Sigma_u z'$, where W is a diagonal matrix with elements defined in equation (2.61). Whilst Breslow and Clayton (1993) motivate the PQL method by a Laplace approximation and deviance and Pearson residuals, Wolfinger and O'Connell (1993) base a broader estimation method on a Taylor expansion around $g(\mu(u))$. As they state, their method encompasses not only the PQL by Breslow and Clayton (1993) but also other estimation techniques proposed by Zeger et al. (1988) and Engel and Keen (1994). The central aspect is the same as the presented method. The resulting Algorithm 2.6 consists of two steps. First, the new working variable $\zeta^{(j)}$ is computed, and second, the weighted linear mixed model is used to obtain the new parameters $\beta^{(j)}, \sigma_k^{2,(j)}, k = 0, \ldots K$.

Algorithm 2.6 PQL Estimation of the Parameters of a Generalized Mixed Model

Set $j = 0$ and starting values $\beta^{(0)}, \sigma_k^{2,(0)}, k = 0, \ldots K$.

1. $\zeta^{(j)} = \eta(u^{(j)}) + (y - \mu(u^{(t)}))g'(\mu(u^{(j)}))$
2. Obtain from a mixed model estimation method the parameters $\beta^{(j)}, \sigma_k^{2,(j)}, k = 0, 1, \ldots K$ for the weighted mixed model

$$\zeta^{(t)} = x\beta + zu + e$$

with weights $w_{ii}^{(j)} = \dfrac{1}{\phi a_i \mathrm{v}(\mu_i(u^{(j)}))[g'(\mu_i(u^{(j)}))]^2}$.

3. Repeat steps 1-2 until convergence.

(Breslow & Clayton, 1993 and Wolfinger & O'Connell, 1993)

Chapter 3

Small Area Estimation

In 2011, Switzerland moved from a 10-year based full census to the annual *Swiss Structural Survey* for measuring the main population parameters of interest. One of the most important figures in this context is the rate of *active population*. This rate is the proportion of persons who are, roughly speaking, either employed or unemployed, but part of the potentially working population aged between 15-74. For further details see Bundesamt für Statistik Schweiz (2013b).

As the Swiss Structural Survey is not a full census but a sample of the population, this rate and most other population parameters have to be estimated. The classical design based estimation methods, however, need relatively high sample sizes to deliver precise estimates. Thus, they will provide reliable and accurate estimates on high aggregation levels. On small aggregation levels, such as communities, the design based estimators may produce only inaccurate estimates not apt for publishing. The precision of the design based estimator is generally measured by a variance estimator, as they are at least asymptotically design unbiased.

Model based small area estimators may produce accurate estimators even on a relatively small aggregation level. This is achieved by using a model to explain the variation in the dependent variable, alias the variable of interest, with available covariates. The model is built depending on the aggregation level of these covariates. That is, if the variables available are aggregated on an *area level*, then an area level model is used. If they are available on a *unit level*, then a unit level model is used. Sometimes the covariates are available on multiple aggregation levels. Then, usually the highest aggregation level is used for the model. This may not be optimal as, in general, models based on lower aggregation levels allow for more precision for the estimation (for an exception see Vogt, 2007, § 4.2).

In contrast to the design based estimators, the model based small area estimators generally have a design bias in complex survey designs. Therefore, the variance of the point estimate is not a sufficiently good measure for its precision. Hence, instead of the variance

estimate, in general, an MSE estimate is reported. For some of the small area estimators, analytical approximations to the MSE are available. For the other small area estimators, resampling methods like the Jackknife and the Bootstrap are used. A basic introduction into small area estimation can be found in Münnich, Burgard, and Vogt (2013) and the comprehensive standard textbook is J. N. K. Rao (2003).

The lack of design unbiasedness is one of the major hurdles for official statistics to adopt these model based small area estimates. E.g., it is difficult to communicate to a Mayor that the total population figure produced may be negatively biased systematically. He would not accept such a number at a first glance. However, the unbiasedness is only half of the story. In Figure 3.1 this fact is visualized exemplary.

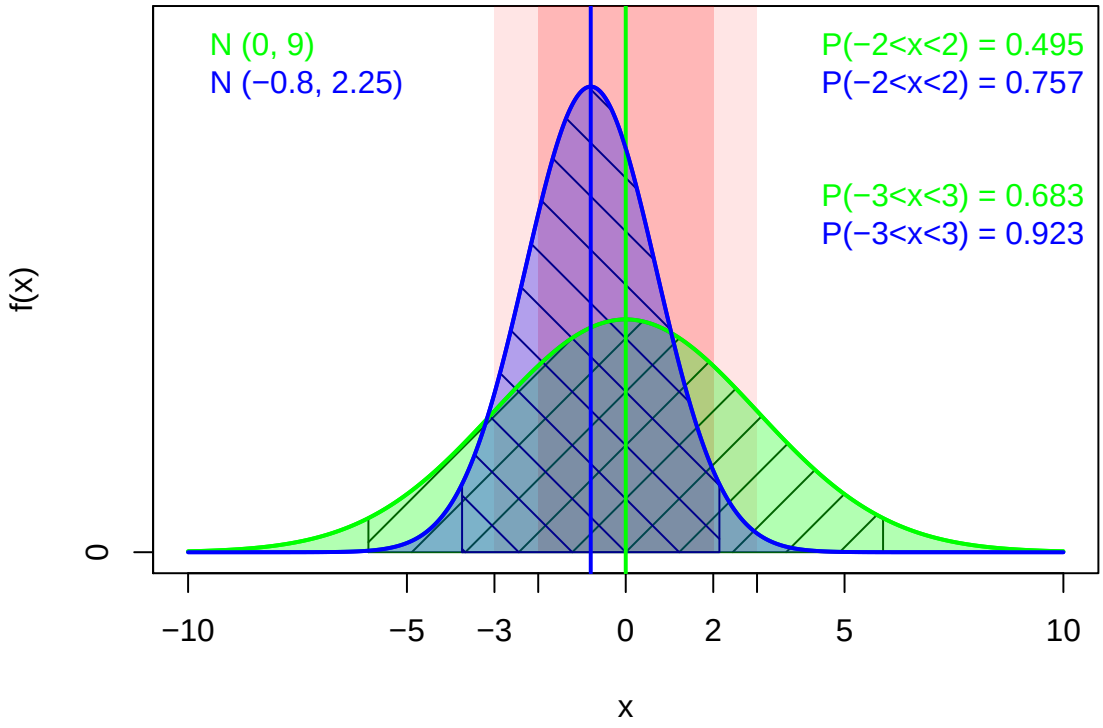

Figure 3.1: Distribution of design based (blue) versus model based (green) point estimates

The green line represents the distribution of a typical design based point estimate. It is unbiased, that is, its mean equals the true value, which in this example is zero. Further, its variability is relatively high. In contrast the blue line demonstrates the distribution of a model based estimate, which is design biased but has a considerably lower variance. For this example, without loss of generality, a negative bias is chosen. The central 95% of each distribution is shaded. In this situation the whole central 95% of the model based distribution lies within the central 95% of the design based distribution. This is due to the lower variability and an absolute bias that is not very high. If the bias or the variability is higher, this may cause the distribution not to overlap the true value. However, in this case, the design based estimates often have such a large variance that, in terms of the MSE, the model based estimate is still preferable.

Another way to compare these two methods is to predefine a certain interval, into which the estimate should fall. In this example, this interval is chosen to be $(-2,2)$ and $(-3,3)$, highlighted in red. Now, the probability of obtaining a point estimate that falls into this interval may be computed from the distribution of the estimates. As can be seen, even though the model based estimate has a bias, the probability in this case is higher for the model based than for the design based point estimate to fall into the interval. Of course, again, if the absolute bias had been higher, this picture would have been quite different. All in all, it seems more appropriate to use a measure for the precision that takes into account both the variability and the bias of the distribution of the point estimates. Consequently, small area estimates should be considered by official statistics, even if they are not design unbiased.

Coming back to the mayor, the answer one could give him is that he should base his policy on the 95% confidence interval. Because, in the ex ante perspective of 95% of samples drawn, the confidence interval will overlap the true value, regardless of whether or not the point estimate is biased. This is, of course, only true if the confidence interval is built correctly. The point estimate on its own has, evidently, only a little information, as it may fall within a wide range without further knowledge as to where. The advantage of small area estimates is the usually narrower confidence intervals and, therefore, more accurate information about the parameter of interest.

This chapter is organized as follows. First, a brief introduction to sampling designs developed for small area estimation purposes is given. Second, the classical design based estimators are discussed, including point and variance estimates. Third, the different model based small area point estimators and the related MSE estimators are presented. And fourth, an approach to include the model covariates on an intermediate aggregation level is proposed.

3.1 Sampling Designs for Small Area Estimation

Let $\mathscr{U} = \{1,\dots,N\}$ be a finite set of identifiers for a population of size N. Further, let $\mathscr{S}$ be a subset of $\mathscr{U}$. The inclusion probability π_i is the probability that unit i is sampled. i.e. $\mathrm{P}(i \in \mathscr{S})$ (c.f. Fuller, 2009, § 1, Lohr, 2010, and Cochran, 1977). The *design weights* w_i are called the *inverse inclusion probabilities*, that is, $w_i := 1/\pi_i$. These design weights are necessary for the estimation of design based estimates (see Section 3.2).

In register-assisted censuses, usually designs without replacement are used, e.g. a unit can only be observed once in the sample. For simplilcity, in the following only this case is considered. If a sample of size n is drawn from a population $\mathscr{U}$ of size N without further restrictions, than the design is called *simple random sampling without replacement* (SRS). The inclusion probabilities are then $\pi_i = \dfrac{n}{N}\ \forall i \in \mathscr{S}$.

An extension to the SRS design is the *stratified random sampling without replacement* (StrRS). In the case of StrRS, the population $\mathscr{U}$ is partitioned into H strata such that $\mathscr{U} = \bigcup_{h=1}^{H} \mathscr{U}_h$ and at the same time $\emptyset = \bigcap_{h=1}^{H} \mathscr{U}_h$. It follows directly that $N = |\mathscr{U}| = \sum_{h=1}^{H} |\mathscr{U}_h| = \sum_{h=1}^{H} N_h$. Within the strata SRS is applied. As such, the inclusion probabilities are $\pi_i = n_h/N_h$ for $i \in \mathscr{U}_h$. There are many different ways of allocating the sample size n to the H different strata. One possibility is to do an *equal allocation* where $n_h = n/H$. Another way is to do a *proportional allocation* where $n_h = n \cdot N_h/N$. For the cases where the resulting $n_h \notin \mathbb{N}$, they have to be rounded under the constraint to sum up to n.

To minimize the total variance of the estimation of variable Y, the *optimal Neyman allocation* further considers the standard deviation of Y in the different strata and leads to $n_h = \frac{N_h s_h}{\sum_{k=1}^{H} N_k s_k} \cdot n$. This kind of allocation of the sample size is known as Neyman allocation, after the famous paper of Neyman (c.f. 1934). As Žarković (c.f. 1956, 1962) shows, Kowalsky (1924) and Tschuprow (1923a, 1923b) anticipated the Neyman-allocation.

The variance reduction of the estimation is greater, the more homogeneous the strata are concerning the variable of interest Y (Pokropp, 1996, p. 5), the higher the correlation between the stratification variable and Y, respectively, and the greater the variation of Y between the strata is (Singh & Mangat, 1996, pp 133). This is due to the fact that the total variance of a point estimate in StrRS is the sum of the variances of the point estimates within the strata. Whereas in SRS, the total variance of a point estimate is the sum of the within strata variance and the between strata variance. Hence, if the between strata variance is greater than zero, then the stratified point estimate has a lower variance than the SRS point estimate.

The interplay between designs and small area estimation is part of recent research (see e.g. Münnich & Burgard, 2012a; Münnich, Burgard, & Zimmermann, 2012). Costa, Satorra, and Ventura (2004) proposed to use a design that mixes a StrRS with proportional allocation with a StrRS with equal allocation. Hereby, a tuning constant, $c \in [0,1]$, is used to determine how much of the proportional allocation or the equal allocation is used. Let n_h^{prop} denote the sample size allocated to stratum h by the proportional allocation and n_h^{equal} denote the sample size allocated to stratum h by the equal allocation. The sample size allocated to stratum h by the Costa et al. (2004) allocation with parameter c is

$$n_h^{\text{Costa}} = c\,\frac{nN_h}{N} + (1-c)\,\frac{n}{H} \quad . \tag{3.1}$$

The main idea behind this approach is to combine the advantages of both allocations. On the one hand, the population estimate should be of certain precision, which would favour the use of proportional allocation; on the other hand, the area estimates should give reasonably precise results, which would go in the direction of equal allocation. By using the convex combination in equation (3.1) a compromise between these two goals is achieved.

A similar idea is pursued by Longford (2006). His allocation is the solution to the minimization of the weighted sum of sampling variances in the areas. Assuming that the sampling variance V_d in area d is known, the set of n_d is searched which minimizes

$$\min_{n_1,..,n_D} \sum_{d=1}^{D} \zeta_d V_d \tag{3.2}$$

maintaining $n = \sum_{d=1}^{D} n_d$ fixed. He proposes to use the Lagrange multipliers approach for this minimization. The coefficients ζ_d denote the relative importance one wishes to give to the precision of the estimate in area d. This importance may be derived from e.g. political goals or simply by a function of the area sizes N_d.

Longford (2006) proposes e.g. to take $\zeta_d = N_d^{c_1}$, $0 \leq c_1 \leq 2$ where a lower value of c_1 would denote equal importance for all areas and the higher c_1 is the greater importance attributed to the larger areas. The sample sizes have to be derived depending on the sampling design applied in the areas. If $V_d = \sigma_d^2/n_d$ (e.g. in SRS), Longford (2006) shows that the optimal allocation under the minimization of equation (3.2) is

$$n_d^{\text{Longford,SRS}} = n \frac{\sigma_d \sqrt{\zeta_d}}{\sum_{d=1}^{D} \sigma_d \sqrt{\zeta_d}}. \tag{3.3}$$

In order to include the importance of the precision of a national estimate $\widehat{\Psi}$, Longford (2006) extends his approach by minimizing

$$\min_{n_1,..,n_D} \sum_{d=1}^{D} \zeta_d V_d \zeta_\circ \mathrm{V}\left[\widehat{\Psi}\right] \quad , \tag{3.4}$$

with $\zeta_\circ := c_2 \sum_{d=1}^{D} \zeta_d$. The higher the scalar c_2 the more importance is given to the precision of the population wide estimate $\widehat{\Psi}$. Again no general solution for the sample allocation can be obtained. In case of simple random sampling within the areas, and the use of the Horvitz-Thompson estimator (see Section 3.2.1) for the population-wide estimate $\widehat{\Psi}$, the optimal sample allocation under equation (3.4) is

$$n_d^{\text{Longford 2,SRS}} = n \frac{\sigma_d \sqrt{\zeta'_d}}{\sum_{d=1}^{D} \sigma_d \sqrt{\zeta'_d}} \tag{3.5}$$

where $\zeta'_d = \zeta_d + \zeta_\circ N_d^2/N^2$.

Choudhry, Rao, and Hidiroglou (2011, 2012) propose a method for finding the allocation with a minimal necessary sample size needed for stratified sampling under some constraints. It consists of finding the

$$\min_{n_1..n_D} \sum_{d=1}^{D} n_d \quad , \tag{3.6}$$

while controlling for some imposed constraints. They propose imposing a maximum tolerance of the coefficient of variation (cv) for each area estimate and for the population-wide estimate. The allocation can be found by using nonlinear programming methods (Choudhry et al., 2012).

In the German Census 2011 (DESTATIS, 2012) an allocation proposed by Gabler, Ganninger, and Münnich (2012) is implemented. The idea arose from a warning by Gelman (2007) who stated that a too high ratio between the highest and lowest design weight is problematic for model estimation. This allocation in principle is an optimal allocation with box constraints on the inclusion probabilities. These box constraints impose a maximum and a minimum inclusion probability for all units in a stratum. By doing this, the range of the design weights can be monitored explicitly and, therefore, so can the ratio between the highest and lowest design-weight. For the precise specification of this allocation see Gabler et al. (2012) and for a fast numerical algorithm to find the allocation see Münnich, Sachs, and Wagner (2012).

3.2 Design-Based Estimates

3.2.1 The Horvitz-Thompson-Estimator

One of the most popular estimators in official statistics is the so-called Horvitz-Thompson estimator proposed by Narain (1951) and Horvitz and Thompson (1952). For the estimation of the area totals it only uses information from the respective area. Therefore it is called a *direct estimator*.

Denoting the inclusion probability with π_i and the design weight is defined as $w_i := 1/\pi_i$ for unit i. Then the Horvitz-Thompson estimator (HT) for the total τ_d of the value of the dependent variable y in area d is defined as follows

$$\widehat{\tau}_{d,\mathrm{HT}} := \sum_{i \in \mathscr{S}_d} w_i y_i, \tag{3.7}$$

where $\mathscr{S}_d$ denotes the set of sampled units in area d.

This estimator is a design unbiased estimator for the total τ_d in area d (c.f.. Särndal, Swensson, & Wretman, 1992, S. 42ff). The precision of this estimator is estimated by the following variance estimator

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{d,\mathrm{HT}}\right) = \sum_{i \in \mathscr{S}_d} \sum_{j \in \mathscr{S}_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{1}{\pi_i} y_i \frac{1}{\pi_j} y_j \qquad , \tag{3.8}$$

(c.f. Särndal et al., 1992, S. 42). In the case of simple random sampling without replacement equation (3.8) can be simplified to:

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{d,\mathrm{HT}}\right) = N_d^2\left(\frac{1}{n_d} - \frac{1}{N_d}\right) \sum_{i \in \mathscr{S}_d} \frac{\left(y_i - n_d^{-1} \sum\limits_{j \in \mathscr{S}_d} y_j\right)^2}{n_d - 1} \quad . \tag{3.9}$$

If the mean is of interest instead of the total the following applies

$$\begin{aligned} \widehat{\mu}_{\mathrm{HT}} &= \frac{\widehat{\tau}_{\mathrm{HT}}}{N} \quad , \quad \text{and} \\ \widehat{V}\left(\widehat{\mu}_{d,\mathrm{HT}}\right) &= \frac{\widehat{V}\left(\widehat{\mu}_{d,\mathrm{HT}}\right)}{N_d^2} \quad . \end{aligned} \tag{3.10}$$

3.2.2 The Generalized Regression Estimator

In contrast to the HT, the Generalized Regression Estimator (GREG) is able to incorporate additional information besides the sole dependent variable and the survey weights. In the small area context there is often additional information about the population of an area. On the one hand, explanatory variables may exist which can help to stabilize the prediction. On the other hand, one can assume that the other areas in a population may behave similarly to the area of interest. In that case, the sample information from the other areas may be used as additional data that can be used for estimating more stable model parameters.

The basic idea behind the GREG is to use information from a linear regression model to correct the HT. The regression model expresses the relationship between the auxiliary information x and the dependent variable y that is to be estimated. The estimates for the regression coefficients $\widehat{\beta}$ are from a least squares estimation and asymptotically unbiased for β due to the inclusion of the design weights w_i. The HT-estimate for y is adapted for the difference between the total of the auxiliary information in the population τ_x, which is known for example from a register, and $\widehat{\tau}_x^{\mathrm{HT}}$, which corresponds with the design-based HT-estimate of y in the regression model. This approach assumes that the model is true and the difference is due to the error of the design-based estimation Münnich et al. (2013).

$$\widehat{\tau}_y^{\mathrm{GREG}} = \widehat{\tau}_y^{\mathrm{HT}} + (\tau_x - \widehat{\tau}_x^{\mathrm{HT}})\widehat{\beta}, \tag{3.11}$$

where $\widehat{\beta}$ is defined in (2.2).

If only the information within the area of interest may be used, then the estimation of the model parameters and the prediction of the estimate are done separately for each area. This estimator will be called SEP-GREG.

$$\widehat{\tau}_{d,\text{SEP-GREG}} = \sum_{i \in \mathscr{S}_d} w_i y_i + \Big(\sum_{i \in \mathscr{U}_d} X_i - \sum_{i \in \mathscr{S}_d} w_i x_i \Big) \widehat{\beta}_{d,\text{LR}} \qquad . \tag{3.12}$$

If one assumes that some other areas have the same relationship between the dependent variable and the covariates, then a grouped GREG (GRP-GREG) can be applied. The GRP-GREG estimates the model parameters over the whole sample of the group g. These parameters are then used for the prediction within all areas d which belong to the group g.

$$\widehat{\tau}_{d,\text{GRP-GREG}} = \sum_{i \in \mathscr{S}_d} w_i y_i + \Big(\sum_{i \in \mathscr{U}_d} X_i - \sum_{i \in \mathscr{S}_d} w_i x_i \Big) \widehat{\beta}_{g,\text{LR}} \qquad . \tag{3.13}$$

Another representation of the GREG estimator can be obtained by some algebraic transformations.

$$\widehat{\tau}_{d,\text{GREG}} = \sum_{i \in \mathscr{U}_d} X_i \widehat{\beta} + \sum_{i \in \mathscr{S}_d} w_i e_i \qquad , \tag{3.14}$$

Lehtonen and Veijanen (1998) expanded the idea of the generalized regression estimator by assuming a nonlinear function of $\widehat{\beta}_*$ (e.g., via a logit-link). Following them, the equation 3.14 may be expanded further by

$$\widehat{\tau}_{\text{GREG,Mod}} = \sum_{i \in \mathscr{U}_d} g_i^{-1}(\widehat{\psi}_{\text{Mod}}) + \sum_{i \in \mathscr{S}_d} (w_i (y_i - g_i^{-1}(\widehat{\psi}_{\text{Mod}}))) \quad , \tag{3.15}$$

where $\widehat{\psi}_{\text{Mod}}$ is a set of parameters obtained from an arbitrary generalized linear (mixed) model with an objective link function $g : \mathbb{R} \to M \subseteq \mathbb{R}$ (see 2.3). In the case of a fixed effects logit-model ψ_{FL} contains only the fixed effects parameters β_{FL} thus for this model it is $g_i^{-1}(\widehat{\psi}_{\text{FL}}) : \dfrac{e^{X_i \widehat{\beta}_{\text{FL}}}}{1 + e^{X_i \widehat{\beta}_{\text{FL}}}} = \widehat{y}_{i,\text{FL}}$. In analogy many different GREG Estimators may be defined.

In the classical GREG the vector $\widehat{\beta}$ is the least squares estimate (2.2) and is asymptotically unbiased for β, if the design weights w_i are used to consider the sample design. If the classical assumptions of the regression model are satisfied, the estimation of the population parameter τ_Y is unbiased. Even in the worst case of not fulfilling the assumptions of the linear regression model, at least it is asymptotically design unbiased (c.f. Särndal

et al., 1992, § 6.4). Thus, the MSE of the point estimate equals (at least asymptotically) its variance.

Särndal et al. (1992, S. 401) write the GREG by using so called g-weights:

$$\widehat{\tau}_{d,\text{GREGg}} = \sum_{i \in \mathscr{S}} g_{di} w_i y_i \quad , \tag{3.16}$$

where,

$$g_{di} = \mathbb{I}_{i \in \mathscr{U}_d} + \Big(\sum_{i \in \mathscr{U}_d} x_i - \frac{N_d}{\widehat{N}_d} \sum_{i \in \mathscr{S}_d} w_i x_i \Big)' \big(x' W x \big)^{-1} x' \quad . \tag{3.17}$$

$\mathbb{I}_{i \in \mathscr{U}_d}$ is an indicator function being one if unit i is in Area d, else zero, and $W = \text{diag}(w_1, \ldots, w_n)$. The variance estimator of the GREG may be expressed using the g-weights

$$\widehat{\text{V}}\big(\widehat{\tau}_{d,\text{GREGg}}\big) = \sum_{i \in \mathscr{S}} \sum_{j \in \mathscr{S}} \frac{\pi_{ji} - \pi_i \pi_j}{\pi_{ji}} g_{dj} w_j e_j \; g_{di} w_i e_i. \tag{3.18}$$

where π_{ij} are the second order inclusion probabilities. If j and i are independently drawn and not in the same area, then $\pi_i \pi_j = \pi_{ij}$ and the right-hand side of equation (3.18) vanishes. This is the case of the sampling design considered in this work. Further, it leads to a simplification of the variance formula

$$\widehat{\text{V}}\big(\widehat{\tau}_{d,\text{GREGg}}\big) = \sum_{h=1}^{D} N_h^2 \big(\frac{1}{n_h} - \frac{1}{N_h}\big) \sum_{i \in \mathscr{S}_h} \frac{\big(g_{dhi} e_{hi} - n_h^{-1} \sum_{j \in \mathscr{S}_h} g_{dhj} e_{hj}\big)^2}{n_h - 1} \quad . \tag{3.19}$$

A proof of the residual variance estimator can be found in Särndal et al. (c.f. 1992, § 6.6), and for separate groups for each stratum in Münnich (1997, § 2).

By Taylor expansion, one can obtain a different and computational less demanding variance formula (c.f. Särndal et al., 1992, § 6.6)

$$\widehat{\text{V}}\big(\widehat{\tau}_{d,\text{GREG}}\big) = N_d^2 \big(\frac{1}{n_d} - \frac{1}{N_d}\big) \sum_{i \in \mathscr{S}_d} \frac{\big(e_i - n_d^{-1} \sum_{j \in \mathscr{S}_d} e_j\big)^2}{n_d (n_d - 1)} \quad . \tag{3.20}$$

3.3 Model Based Prediction

3.3.1 The Fay-Herriot Estimator

Fay and Herriot (1979) proposed the so-called Fay-Herriot estimator (FH) for the estimation of the mean population income in a small area setting. They assume that covariates may only be available at aggregate level, such as communities, and not on unit-level. As Jiang and Lahiri (2006b) state, this can be seen as a special case of a mixed model where for every area there is only one observation. Hence, the matrix $z = Z$ is just the identity

matrix of size $D \times D$. As dependent variable Fay and Herriot (1979) used direct estimates obtained from the sample. This direct estimate $\widehat{\mu}_{d,\text{direct}}$ may be e.g. a HT or a GREG estimate. The covariates are true population parameters, e.g. population means $\overline{X}$. The model underlying the FH is then

$$\widehat{\mu}_{d,\text{direct}} = \overline{X}\beta + u_i + e_i \quad . \tag{3.21}$$

The FH is the prediction from this mixed model and is given by

$$\widehat{\mu}_{d,\text{FH}} = \overline{X}_d\widehat{\beta} + \widehat{u}_d \quad , \tag{3.22}$$

with

$$\widehat{u}_d = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \sigma_{e,d}^2}(\widehat{\mu}_{d,\text{direct}} - \overline{X}\widehat{\beta}) \quad .$$

Whilst the $\widehat{\sigma}_u^2$ and $\widehat{\beta}$ are estimates, the $\sigma_{e,d}^2, d = 1,\ldots,D$, are assumed to be known. Jiang and Lahiri (2006b) state that if all areas have sample allocation and simple random sampling is applied within the area, one can use the variance estimate for a direct survey estimate for area d as $\sigma_{e,d}^2, d = 1..D$. This could be the variance estimate (3.10). Assuming σ_u is known, then β can be estimated in a similar way, as in the mixed model, via a weighted least squares approach

$$\widehat{\beta}(\sigma_u) = (\overline{X}'V^{-1}\overline{X})^{-1}\overline{X}'V^{-1}\widehat{\mu}_{d,\text{direct}} \quad . \tag{3.23}$$

As there is only one observation per area, the matrix V^{-1} is easily computed by $V^{-1} = \sigma_e^{-2} \odot I_{(D)}\sigma_u^{-2}$, σ_e in this case being the $D \times 1$ vector of $\sigma_{e,d}$. There exist several methods for the estimation of σ_u when $\sigma_{e,d}$ are assumed to be known. Fay and Herriot (1979) proposed an estimation equation approach using the weighted least squares residuals. It consists of solving the equation

$$f(\sigma_u) = \sum_{d=1}^{D} \frac{\left(\widehat{\mu}_{d,\text{direct}} - \overline{X}_d\widehat{\beta}(\sigma_u)\right)^2}{\sigma_u^2 + \sigma_{e,d}^2} - (D-p) = 0 \quad , \tag{3.24}$$

and equation (3.23) iteratively. This can be done via the Newton-Raphson algorithm (c.f. Section A.3).

Algorithm 3.1 Estimation of σ_u for the Fay-Herriot estimator: The Fay-Herriot version.

An approximation to the first derivative of $f(\sigma_u^2)$ is

$$f'(\sigma_u^2) = \frac{df}{d\sigma_u^2} \approx -\sum_{d=1}^{D} \frac{\left(\widehat{\overline{y}}_d - \overline{X}_d \widehat{\beta}(\sigma_u^2)\right)^2}{\left(\sigma_u^2 + \sigma_{e,d}^2\right)^2}.$$

1. Set j=0 and the starting value $\sigma_u^{2(0)} = 0$.
2. Increase j by one and compute $\sigma_u^{2(j)} = \sigma_u^{2(j-1)} - \frac{f(\sigma_u^{2(j-1)})}{f'(\sigma_u^{2(j-1)})}$
3. Repeat step 2 until convergence of $\sigma_u^{2(j)}$ ($|\sigma_u^{2(j)} - \sigma_u^{2(j-1)}| < \varepsilon$ arbitrary small).

(Fay & Herriot, 1979)

As can be seen from equation (3.22), the relevance of the precision of σ_u depends on the scale of $\sigma_{e,d}$. If $\sigma_{e,d} \gg \sigma_u$, $d = 1..D$, then a small change in σ_u will not change much in the prediction and thus, ε does not have to be too small. But if some $\sigma_{e,d}$ and σ_u are almost of the same size or even some of the $\sigma_{e,d} \ll \sigma_u$, $d = 1..D$, then it is very important to choose a sufficiently small ε in order to obtain a precise prediction.

Another estimator is proposed by Prasad and Rao (1990). It is an estimator using Henderson's method 3 (Henderson, 1953), a fitting of constants estimator,

$$\widehat{\sigma}_u = \frac{\sum_{d=1}^{D} \left(\widehat{\overline{y}}_d - \overline{X}_d (\overline{X}'\overline{X})^{-1} \overline{X}' \widehat{\overline{y}}\right)^2 - \sum_{d=1}^{D} \sigma_{e,d} (1 - \overline{X}_d (\overline{X}'\overline{X})^{-1} \overline{X}'_d)}{D - p} \quad . \tag{3.25}$$

This estimator may become negative, and if this happens the $\widehat{\sigma}_u$ is set to zero. Plugging this variance into Equation (3.23) yields the corresponding β-estimate.

The σ_u can be estimated as well by ML or REML. The algorithm is a simplification of Algorithm 2.1, as only one variance component has to be estimated.

Algorithm 3.2 Estimation of σ_u for the Fay-Herriot estimator: The ML and REML version.

The Jacobian is

ML $J^{\mathrm{ML}}(\beta, \sigma_u) = -\frac{1}{2}\Big(\sum_{d=1}^{D} \frac{1}{\sigma_u^2 + \sigma_{e,d}^2} - \sum_{d=1}^{D} \frac{(\widehat{\mu}_{d,\mathrm{direct}} - \overline{X}\beta)^2}{(\sigma_u^2 + \sigma_{e,d}^2)^2} \Big)$

REML $J^{\text{REML}}(\beta,\sigma_u) = -\frac{1}{2}\left(\text{tr}\,[PP] - \text{tr}\left[\widehat{\mu}'_{d,\text{direct}} P \widehat{\mu}_{d,\text{direct}}\right]\right)$

Obtain the Hessian matrix of the log-likelihood

ML $H^{\text{ML}}(\beta,\phi) = \frac{1}{2}\sum_{d=1}^{D}\frac{1}{(\sigma_u^2+\sigma_{e,d}^2)^2}$

REML $H^{\text{REML}}(\beta,\phi) = \frac{1}{2}\,\text{tr}\,[PP]$

1. Set j=0 and the starting value $\sigma_{u,\text{ML}}^{2(0)} = \sigma_{u,\text{REML}}^{2(0)} = 0$.
2. Increase j by one and compute

   ML $\beta_{\text{ML}}^{(j)} = \left(\overline{X}'(\sigma_{u,\text{ML}}^{2(j-1)} I_{(n)}\sigma_e^2)^{-1}\overline{X}\right)^{-1}\overline{X}'\widehat{\mu}_{d,\text{direct}}$

   REML $\beta_{\text{REML}}^{(j)} = \left(\overline{X}'(\sigma_{u,\text{REML}}^{2(j-1)} I_{(n)}\sigma_e^2)^{-1}\overline{X}\right)^{-1}\overline{X}'\widehat{\mu}_{d,\text{direct}}$
3. Update σ_u^2

   ML $\sigma_{u,\text{ML}}^{2(j)} = \sigma_{u,\text{ML}}^{2(j-1)} - \left(H^{\text{ML}}(\beta_{\text{ML}}^{(j-1)}\sigma_{u,\text{ML}}^{2(j-1)})\right)^{-1} J^{\text{ML}}(\beta_{\text{ML}}^{(j-1)}\sigma_{u,\text{ML}}^{2(j-1)})$

   REML $\sigma_{u,\text{REML}}^{2(j)} = \sigma_{u,\text{REML}}^{2(j-1)} - \left(H^{\text{REML}}(\beta_{\text{REML}}^{(j)}\sigma_{u,\text{REML}}^{2(j-1)})\right)^{-1} J^{\text{REML}}(\beta_{\text{REML}}^{(j)}\sigma_{u,\text{REML}}^{2(j-1)})$
4. Repeat step 2 until convergence of $\sigma_u^{2(j)}$ ($|\sigma_u^{2(j)} - \sigma_u^{2(j-1)}| < \varepsilon$ arbitrary small).

3.3.2 The Battese-Harter-Fuller Estimator

For the case that unit-level covariates are available, Battese, Harter, and Fuller (1988) proposed the so-called Battes-Harter-Fuller estimator (BHF). This estimator, like the FH, can be constructed by using the mixed model framework, as Moura and Holt (1999) and Jiang and Lahiri (2006b) show.

The mixed model used, is a random intercept model

$$y = x\beta + zu + e \quad , \tag{3.26}$$

with the matrix z being of dimension $n \times D$ with elements

$$z_{id} = \begin{cases} 1, & \text{if unit } i \text{ is in area } d, \\ 0, & \text{else} \end{cases} \quad . \tag{3.27}$$

The estimation of the model parameters is described in Section 2.2.

The BHF is the prediction from the linear mixed model (3.26) with the estimated model parameters $\widehat{\beta}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2$ and the national mean of the covariates $\overline{X}$:

$$\widehat{\mu}_{\text{BHF}} = \overline{X}\widehat{\beta} + \widehat{u} \quad , \tag{3.28}$$
$$\widehat{u} = \gamma_d(\overline{y} - \overline{x}\widehat{\beta}) = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_e^2/n_d}(\overline{y} - \overline{x}\widehat{\beta}) \quad .$$

This estimator is an empirical best linear unbiased predictor (EBLUP) (Battese et al., 1988).

3.3.3 MSE Estimation for the Nested Error Regression Models

As both, the FH and the BHF, rely on a nested error regression model, this can generally be written as

$$\mu(\phi) = \overline{X}\widehat{\beta} + \gamma(\overline{y} - \overline{x}\widehat{\beta}) \quad . \tag{3.29}$$

They are BLUP if the variance components are known (Henderson, 1975). Kackar and Harville (1984) showed that under certain conditions (cf. Kackar & Harville, 1984, pp. 853), the MSE of the BLUP can be decomposed into

$$\text{MSE}\left[\mu(\widehat{\phi})\right] = \text{MSE}\left[\mu(\phi)\right] + \text{E}\left[\mu(\widehat{\phi}) - \mu(\phi)\right]^2 \quad . \tag{3.30}$$

As Prasad and Rao (1990) state, the second term in (3.30) is only tractable in special cases, as done by Peixoto and Harville (1986) and Peixoto (1988) for the balanced one-way analysis of variance model. Kackar and Harville (1984) proposed a Taylor series approximation to this term

$$\text{E}\left[\mu(\widehat{\phi}) + \mu(\phi)\right]^2 \doteq \text{E}\left[\left(\frac{\partial \mu \phi}{\partial \phi}\right)'(\widehat{\phi} - \phi)\right]^2 \quad . \tag{3.31}$$

The proposed approximation is

$$\text{E}\left[\left(\frac{\partial \mu \phi}{\partial \phi}\right)'(\widehat{\phi} - \phi)\right]^2 \doteq \text{tr}\left[\text{VCOV}\left[\frac{\partial \mu \phi}{\partial \phi}\right]\text{E}\left[(\widehat{\phi} - \phi)(\widehat{\phi} - \phi)'\right]\right] \quad . \tag{3.32}$$

Prasad and Rao (1990) propose a further approximation, valid for their estimator (3.25) for the variance component.

$$\mathrm{E}\left[\left(\frac{\partial \mu\phi}{\partial \phi}\right)'(\widehat{\phi}-\phi)\right]^2 \doteq \mathrm{tr}\left[(J_b)V(J_b')\,\mathrm{E}\left[(\widehat{\phi}-\phi)(\widehat{\phi}-\phi)'\right]\right] \quad , \tag{3.33}$$

where $J_b = \mathrm{col}_{1\leq j\leq p}\left(\frac{\partial DGZ'V^{-1}}{\partial \phi_j}\right)$. In case of ML or REML estimation of $\widehat{\phi}$, Datta and Lahiri (2000) show that the approximation is slightly different

$$\mathrm{E}\left[\left(\frac{\partial \mu\phi}{\partial \phi}\right)'(\widehat{\phi}-\phi)\right]^2 \doteq \frac{1}{D}\,\mathrm{tr}\left[(J_b)V(J_b')\,\mathrm{VCOV}^{-1}[\phi]\right] \quad , \tag{3.34}$$

which in case of the FH leads to a different MSE estimator. Das, Jiang, and Rao (2004); Datta and Lahiri (2000); Datta, Rao, and Smith (2005); Prasad and Rao (1990) provide deeper discussions on the conditions necessary for this approximations to hold. Lahiri and Rao (1995) show, that the MSE estimator of Prasad and Rao (1990) is robust against non-normality of the $\mu_{d,\mathrm{BHF}}$, being correct to terms the order $O(D^{-1})$. An alternative approximation is given by Chen and Lahiri (2008). They obtain a Taylor approximation to the jackknife MSE estimate for the EBLUP of a general linear mixed model. As the MSE estimator by Prasad and Rao (1990) works well in the applications considered in Chapter 5, the focus lies on it, and the approximation by Chen and Lahiri (2008) will not be considered further.

3.3.3.1 Second Order Approximation to the MSE of BHF

Prasad and Rao (1990) split the MSE into three components

$$\mathrm{MSE}[\mu_d(\phi)] = g_{1,d}(\phi) + g_{2,d}(\phi) + g_{3,d}(\phi) \quad , \tag{3.35}$$

where

$$\begin{aligned} \mathrm{MSE}[\mu_d(\phi)] &= g_{1,d}(\phi) + g_{2,d}(\phi) \quad , \\ g_{3,d}(\phi) &= \mathrm{tr}\left[(J_b)V(J_b')\,\mathrm{E}\left[(\widehat{\phi}-\phi)(\widehat{\phi}-\phi)'\right]\right] \quad . \end{aligned}$$

They also derive explicit formulas for these three components for the MSE of the BHF estimator:

$$g_{1,d}(\phi) = (1-\gamma_d)\sigma_u^2 \quad , \tag{3.36}$$

$$g_{2,d}(\phi) = (\overline{X}_d - \gamma_d \bar{x}_d)'(x'v^{-1}x)^{-1}(\overline{X}_d - \gamma_d \bar{x}_d) \quad , \tag{3.37}$$

$$g_{3,d}(\phi) = \frac{n_d}{(\sigma_e^2 + n_d\sigma_u^2)^3}\,\mathrm{V}\left[\widehat{\sigma}_e^2\sigma_u^2 - \widehat{\sigma}_u^2\sigma_e^2\right] \tag{3.38}$$

$$= \frac{n_d}{(\sigma_e^2 + n_d\sigma_u^2)^3}\left[\sigma_e^4\,\mathrm{V}\left[\widehat{\sigma}_u^2\right] + \sigma_u^4\,\mathrm{V}\left[\widehat{\sigma}_e^2\right] - 2\sigma_e^2\sigma_u^2\,\mathrm{COV}\left[\widehat{\sigma}_e^2, \widehat{\sigma}_u^2\right]\right] \quad . \tag{3.39}$$

Prasad and Rao (1990) show, that $g_{1,d}(\phi)$ is approximated with order $\mathcal{O}(1)$, $g_{1,d}(\phi)$ with order $\mathcal{O}(D^{-1})$, and $g_{3,d}(\phi)$ with order $o(D^{-1})$. Thus the leading terms are $g_{1,d}(\phi)$ and $g_{2,d}(\phi)$.

According to Münnich and Burgard (2012a), these three components stand for

$g_{1,d}$ The error of the estimator, when all parameters are known.

$g_{2,d}$ The error arising from the estimation of the β vector.

$g_{3,d}$ The error due to the estimation of the γ_d, which depends on the estimation method used (c.f. Datta et al., 2005).

3.3.3.2 Second Order Approximation to the MSE of FH

The approach for the approximation to the MSE of the FH is identical to the one for the BHF. The decomposition of the MSE is again

$$\mathrm{MSE}\left[\mu_d(\phi)\right] = g_{1,d}(\phi) + g_{2,d}(\phi) + g_{3,d}(\phi)$$

$$g_{1,d}(\phi) = \frac{\sigma_u^2 \sigma_{e,d}^2}{\sigma_u^2 + \sigma_{e,d}^2}$$

$$g_{2,d}(\phi) = \frac{\sigma_{e,d}^4}{(\sigma_u^2 + \sigma_{e,d}^2)^2} x_d' (x' v^{-1} x)^{-1} x_d$$

$$g_{3,d}(\phi) = \frac{\sigma_{e,d}^4}{(\sigma_u^2 + \sigma_{e,d}^2)^3} \mathrm{V}\left[\widehat{\sigma}_u^2\right]$$

By assuming that u_d and e_d are normally distributed, $\mathrm{V}\left[\widehat{\sigma}_u^2\right]$ can be derived depending on the estimation methods used for $\widehat{\phi}$. If ϕ is estimated via the originally proposed method by Fay and Herriot (1979), which is a combination of the method of moments and a weighted residual sum of squares, then $g_{3,d}(\phi)$ is (Datta et al., 2005)

$$g_{3,d}(\phi) = \frac{2D\sigma_{e,d}^4}{(\sigma_u^2 + \sigma_{e,d}^2)^3 (\sum_{j=1}^{D} (\sigma_u^2 + \sigma_{e,j})^{-1})^2} \quad . \tag{3.40}$$

Prasad and Rao (1990) use the method of fitting constants to obtain $\widehat{\phi}$. For $g_{3,d}(\phi)$ they derive

$$g_{3,d}(\phi) = \frac{2\sigma_{e,d}^4}{(\sigma_u^2 + \sigma_{e,d}^2)^3 D^2} \sum_{j=1}^{D} (\sigma_u^2 + \sigma_{e,j})^2 \quad . \tag{3.41}$$

Datta and Lahiri (2000) showed, that if $\widehat{\phi}$ is obtained by an ML or REML estimator of ϕ, then $g_{3,d}(\phi)$ is given by

$$g_{3,d}(\phi) = \frac{2\sigma_{e,d}^4}{(\sigma_u^2 + \sigma_{e,d}^2)^3 \sum_{j=1}^{D} (\sigma_u^2 + \sigma_{e,j})^{-2}} \quad . \tag{3.42}$$

Datta et al. (2005) state that the relation between these $g_{3,d}(\phi)$ terms is (3.42) $\leq$ (3.40) $\leq$ (3.41), with equality if and only if $\sigma_{e,i} = \sigma_{e,k} \forall j,k = 1..D$.

3.3.4 Pseudo EBLUP Estimators

Neither the FH nor the BHF account explicitly for design weights. As such, they assume that the sample is a simple random sample from the population. In most cases, however, the design is not simple random sampling, but much more complex. The result is that not all units in the population have the same probability to be in the sample. In order to account for this, the design weights should be incorporated into the estimation process. Estimators coping for the sampling design were mainly proposed by Pfeffermann et al. (1998); Prasad and Rao (1999) and Pfeffermann and Sverchkov (2007). Münnich and Burgard (2012a) study the effect of a large variety of designs on small area estimates within a large Monte Carlo simulation study. The general result is that the design is critical. Especially in the case of stratified sampling with optimal allocation and cluster sampling, the small area models incorporating design weights outperform the somewhat *naïve* estimators which assume a simple random sample.

A very interesting approach for the incorporation of the design into hierarchical Bayes estimators is proposed by Lahiri and Mukherjee (2007). They suggest a correction of the hierarchical Bayes estimator for achieving design consistency, and propose also uncertainty measures for this estimator. As the hierarchical Bayes method is not part of this work this approach is beyond the scope.

3.3.4.1 Pseudo EBLUP Area-Level Estimator

Prasad and Rao (1999) developed an area-level pseudo EBLUP which is design consistent and incorporates weights. It is a special case of the estimator by Fay and Herriot (1979) without covariates, where the area observations are obtained from the sample by a weighted area mean. The weights used are the inverse inclusion probabilities w_{id} for person i in area d.

$$\begin{aligned}\widehat{\bar{y}}_d &= \frac{1}{\sum\limits_{i\in\mathscr{S}_d} w_{id}} \sum_{i\in\mathscr{S}_d} w_{id} y_{id} = \sum_{i\in\mathscr{S}_d} \widetilde{w}_{id}(\beta_0 + u_d + e_{id}) \qquad (3.43)\\ &= \beta_0 + u_d + \sum_{i\in\mathscr{S}_d} \widetilde{w}_{id} e_{id} \quad ,\\ \widetilde{w}_{id} &= \frac{w_{id}}{\sum\limits_{j\in\mathscr{S}_d} w_{jd}} \quad \text{with unit } i \in \mathscr{S}_d \quad . \qquad (3.44)\end{aligned}$$

Prasad and Rao (1999) derive the best linear unbiased estimator for the area mean under model (3.43) as

$$\widehat{\mu}_{d,\mathrm{PR1}} = \widehat{\beta}_0 + \widehat{u}_d \quad , \qquad (3.45)$$

with

$$\begin{aligned}\widehat{\beta}_0 &= \frac{\sum\limits_{d=1}^{D} \widehat{\gamma}_{d,\mathrm{PR}} \widehat{\bar{y}}_d}{\sum\limits_{d=1}^{D} \widehat{\gamma}_{d,\mathrm{PR}}} \quad , \qquad \widehat{u}_d = \widehat{\gamma}_d(\widehat{\bar{y}}_d - \widehat{\beta}_0) \quad ,\\ \widehat{\gamma}_{d,\mathrm{PR}} &= \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_e^2 \delta_d} \quad , \qquad \delta_d = \sum_{i\in\mathscr{S}_d} \widetilde{w}_{id}^2 \quad .\end{aligned}$$

As (σ_e, σ_u) are usually not known, Prasad and Rao (1999) use $(\widehat{\sigma}_e^2, \widehat{\sigma}_u^2)$ as plugin estimators in equations (3.45). These are given by

$$\widehat{\sigma}_e^2 = \frac{\sum\limits_{i=1}^{n_d} \sum\limits_{d=1}^{D} (y_{id} - \bar{y}_d)^2}{n - D} \quad , \qquad (3.46)$$

and as $\widehat{\sigma}_u^2$ may not be negative,

$$\widehat{\sigma}_u^2 = max(0, \widehat{\widetilde{\sigma}}_u^2) \quad , \quad \widehat{\widetilde{\sigma}}_u^2 = \frac{\sum\limits_{d=1}^{D} n_d(\widehat{\bar{y}}_d - \widehat{\bar{y}})^2 - (D-1)\widehat{\sigma}_e}{n - \frac{1}{n}\sum\limits_{d=1}^{D} n_d^2} \quad . \qquad (3.47)$$

As the estimator $\widehat{\mu}_{\text{PR}}$ incorporates the design weights, Prasad and Rao (1999) call it a pseudo EBLUP area-level estimator.

The MSE estimation is done in analogy to the MSE estimation for the Fay-Herriot estimator. The MSE of $\widehat{\mu}_{\text{PR}}$ can be decomposed into

$$\text{MSE}\left[\widehat{\mu}_{\text{d,PR}}\right] \approx g_{1,d} + g_{2,d} + 2g_{3,d}$$

with

$$\begin{aligned} g_{1,d} &= (1-\widehat{\gamma}_{d,\text{PR}})\widehat{\sigma}_u^2 \\ g_{2,d} &= \frac{((1-\widehat{\gamma}_{d,\text{PR}})\widehat{\sigma}_u)^2}{\sum_{d=1}^{D}\widehat{\gamma}_{d,\text{PR}}} \\ g_{3,d} &= \frac{\widehat{\gamma}_{d,\text{PR}}(1-\widehat{\gamma}_{d,\text{PR}})^2}{\widehat{\sigma}_u^2}\left(\text{V}\left[\widehat{\widetilde{\sigma}}_u^2\right] - 2\frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_e^2}\text{COV}\left[\widehat{\widetilde{\sigma}}_u^2, \widehat{\sigma}_e^2\right] + \frac{\widehat{\sigma}_u^4}{\widehat{\sigma}_e^4}\text{V}\left[\widehat{\widetilde{\sigma}}_u^2\right]\right) \end{aligned} \tag{3.48}$$

If u_d, e_{id} are normally distributed then the variance and covariance in $g_{3,d}$ are given by (Prasad & Rao, 1999)

$$\begin{aligned} \text{V}\left[\widehat{\sigma}_e\right] &= \frac{2\sigma_e}{n-D} \\ \text{V}\left[\widehat{\widetilde{\sigma}}_u\right] &= \frac{2}{(n-\frac{1}{n}\sum_{d=1}^{D}n_d^2)^2}\Big(\frac{\sigma_e^4(D-1)(n-1)}{(n-D)} + \\ &\qquad 2(n-\frac{1}{n}\sum_{d=1}^{D}n_d^2)\sigma_e^2\sigma_u^2 + \{\sum_{d=1}^{D}n_d^2 - 2\frac{\sum_{d=1}^{D}n_d^3}{n} + \frac{(\sum_{d=1}^{D}n_d^2)^2}{n^2}\}\sigma_u^4\Big) \\ \text{COV}\left[\widehat{\sigma}_e, \widehat{\widetilde{\sigma}}_u\right] &= \frac{-(D-1)}{n-\frac{1}{n}\sum_{d=1}^{D}n_d^2}\text{V}\left[\widehat{\sigma}_e\right] \end{aligned} \tag{3.49}$$

Prasad and Rao (1999) propose an extension of this pseudo EBLUP to the nested error regression model case, where instead of using only a common intercept for all areas, some covariates x are introduced into the model. These x are known from the sample and the population mean $\overline{X}$ is assumed to be known. The area-level pseudo EBLUP with covariates they propose is

$$\widehat{\mu}_{\text{d,PR}} = \overline{X}_d\widehat{\beta}_{\text{PR}} + \widehat{u}_d \quad , \tag{3.50}$$

where $\widehat{\beta}_{\text{PR}}$ is the weighted least square estimate

$$\widehat{\beta}_{\text{PR}} = \left(\sum_{d=1}^{D}\gamma_{d,\text{PR}}\overline{x}'_{d,\widetilde{w}}\overline{x}_{d,\widetilde{w}}\right)^{-1}\left(\sum_{d=1}^{D}\gamma_{d,\text{PR}}\overline{x}'_{d,\widetilde{w}}\widehat{\overline{y}}_d\right) \quad , \tag{3.51}$$

with

$$\overline{x}_{d,\widetilde{w}} = \sum_{i \in \mathscr{S}_d} x_i \widetilde{w}_i \quad , \tag{3.52}$$

and

$$\widehat{u}_d = \widehat{\gamma}_{d,\mathrm{PR}} (\widehat{\overline{y}}_d - \overline{x}_{d,\widetilde{w}} \widehat{\beta}_{\mathrm{PR}}) \quad . \tag{3.53}$$

The MSE estimation is performed analogously to the MSE estimation by Prasad and Rao (1990).

3.3.4.2 Pseudo EBLUP Unit-Level Estimator

You and Rao (2002) extended this area-level pseudo EBLUP to obtain a unit-level pseudo EBLUP. In analogy to Prasad and Rao (1999), they first rescale the weights as shown in equation (3.44). Additionally, they estimate the model parameters on unit-level, in such way that they obtain, as they call it, an *automatic benchmarking property*. This means that the aggregation of the area estimates equals a GREG estimate on the whole population level.

Similar to the other EBLUPs the estimator by You and Rao (2002) is defined as

$$\widehat{\mu}_{d,\mathrm{YR}} = \overline{X}_d \widehat{\beta}_{\mathrm{YR}} + \widehat{u}_{d,\mathrm{YR}} \quad , \tag{3.54}$$

where $\widehat{u}_d = \widehat{\gamma}_{d,\mathrm{PR}} (\widehat{\overline{y}}_d - \overline{x}_{d,\widetilde{w}} \widehat{\beta}_{\mathrm{YR}})$, and

$$\widehat{\beta}_{\mathrm{YR}} = \left(\sum_{d=1}^{D} \sum_{i \in \mathscr{S}} w_i x_i (x_i - \gamma_{d,\mathrm{PR}} \overline{x}_{d,\mathrm{PR}}) \right)^{-1} \left(\sum_{d=1}^{D} \sum_{i \in \mathscr{S}} w_i y_i (x_i - \gamma_{d,\mathrm{PR}} \overline{x}_{d,\mathrm{PR}}) \right) \quad . \tag{3.55}$$

For $\overline{x}_{d,\widetilde{w}}$ see Equation (3.52), and $\widehat{\gamma}_{\mathrm{PR}}$ is the vector of the $\widehat{\gamma}_{d,\mathrm{PR}}$ (see Equation (3.45)).

In order to estimate the MSE for this pseudo EBLUP, You and Rao (2002) adapt straightforward the approach of Prasad and Rao (1999), who extend the MSE estimator of Prasad and Rao (1990) in order to account for the use of the sampling weights. Basically, $g_{1,d}$ and $g_{3,d}$ remain as in equation (3.48). The term $g_{2,d}$ is changed to account for the weighted β estimation by using

$$g_{2,d} = \left(\overline{X}_d - \gamma_{d,\mathrm{PR}} \overline{x}_{d,\widetilde{w}}\right)' \left[\sigma_e (x'\zeta)^{-1} \zeta' \zeta \left((x'\zeta)^{-1} \right)' + \sigma_u (x'\zeta)^{-1} \kappa \left((x'\zeta)^{-1} \right)' \right] \left(\overline{X}_d - \gamma_{d,\mathrm{PR}} \overline{x}_{d,\widetilde{w}}\right) \quad , \tag{3.56}$$

where $\zeta = (\zeta_1, \ldots, \zeta_n)$, $\zeta_{id} = (\widetilde{w}_{id}(x_{id} - \gamma_{d,\mathrm{PR}}\bar{x}_{d,\widetilde{w}}))$ and $\kappa = \sum_{d=1}^{D} \zeta_d' \zeta_d$ with ζ_d being the part of ζ belonging to area d.

As Jiang and Lahiri (2006a) state, You and Rao (2002) did not consider all cross-product terms in the MSE. These cross-product terms may be omitted if all the weights within an area are identical (Torabi & Rao, 2010, § 3.1). Using the regularity conditions and results of Jiang and Lahiri (2006a), Torabi and Rao (2010) apply a Taylor-linearisation to approximate these cross-product terms, yielding a new composition of the MSE

$$\mathrm{MSE}\left[\mu_d(\phi)\right] = g_{1,d}(\phi) + g_{2,d}(\phi) + g_{3,d}(\phi) + C_{1,d}(\phi) + C_{2,d}(\phi) \quad . \tag{3.57}$$

The resulting approximation formula for $C_{2,d}(\phi)$, however, is very extensive. Before it can be applied to surveys of the size of the Swiss Structural Survey it has to be eased computationally. However, as the design weights in this setting are generally constant within the areas, the cross product terms will be very small, and thus, negligible. Alternatively, Torabi and Rao (2010) propose to use a parametric bootstrap for the MSE estimation. In order to include the cross-product terms, they choose to use a parametric double bootstrap following Hall and Maiti (2006). Again, as the approximation to the MSE proposed by You and Rao (2002) will have a bias of low order in the setting of the Swiss Structural Survey, for computational reasons this approach will not be considered. For an application of the parametric double bootstrap method to another estimator see Section 3.3.6.3.

3.3.5 Empirical Best Predictor for Binary Variables

The linear methods previously described are useful for continuous variables. In case of binary variables, they may produce nonsense-estimates; for instance, estimated proportions lying over one or under zero. This is especially the case if (a) the true proportion of the areas cannot be explained sufficiently by the model, (b) the true proportions lie near zero or one, and (c) the true proportions vary considerably between the areas. Thus, an estimator is needed that can handle the special structure of binary variables.

Malec, Sedransk, Moriarity, and LeClere (1997) proposed a hierarchical Bayes approach for the estimation of proportions for the National Health Interview Survey. They state that their method works well, but that it is computationally very demanding. Ghosh, Natarajan, Stroud, and Carlin (1998) propose a broader hierarchical base approach, in that their method is not only apt for binary variables, but also for every variable that can be modelled via a generalized linear (mixed) model. Nevertheless, this approach is also computationally extremely demanding.

Easing the computational burden, Jiang and Lahiri (2001) propose, as they call it,

> a frequentist alternative to the already existing hierarchical Bayes methods (Jiang & Lahiri, 2001, p. 218).

They use a mixed logistic model to model the binary variable

$$y_{id}|\theta_{id} \overset{\text{iid}}{\sim} \text{Bern}(\theta_{id}) \quad , \tag{3.58}$$
$$\text{logit}(\theta_{id}) = x_{id}\beta + u_d \quad . \tag{3.59}$$

Jiang and Lahiri (2001) show that the best predictor (BP) $\widehat{\theta}$ for the area proportion under this model is

$$\text{E}[\theta_d|y] = \text{E}[\theta_d|y_d] = \frac{\text{E}\left[h_d(u_d,\phi)e^{l_d(y_d,u_d,\phi)}\right]}{\text{E}\left[e^{l_d(y_d,u_d,\phi)}\right]} \quad , \tag{3.60}$$

with $h_d(u_d,\phi) = \text{logit}^{-1}(X_d\beta + u_d)$. $l_d(y_d,u_d,\phi)$ is the log-likelihood function for area d under the model (3.58) and $\text{E}\left[e^{l_d(y_d,u_d,\phi)}\right]$ is the marginal likelihood.

B. Liu (2009) extends this estimator by assuming the more general exponential distribution on the u_d instead of the normal distribution.

Here the estimator of Jiang and Lahiri (2001) is presented for the case that a Binomial distribution is assumed on y

$$y_d|\theta_d \overset{\text{iid}}{\sim} \text{Bin}(n_d,\theta_d) \quad . \tag{3.61}$$

The resulting log-likelihood function for area d is

$$l_d(y_d,u,\phi) = \log\left(\frac{1}{\sigma_u\sqrt{2\pi}}\binom{n_d}{y_d}\left[\frac{e^{\overline{X}_d\beta+u}}{1+e^{\overline{X}_d\beta+u}}\right]^{y_d}\left[1-\frac{e^{\overline{X}_d\beta+u}}{1+e^{\overline{X}_d\beta+u}}\right]^{n_d-y_d} e^{-\frac{u^2}{2\sigma_u^2}}\right) \quad . \tag{3.62}$$

Following Jiang and Lahiri (2001), the area-level BP (ABP) for binomial variables can then be obtained by

$$\widehat{\theta}_{d,\text{ABP}} = \frac{\int_{-\infty}^{\infty} h_d(u,\phi)l_d(y_d,u,\phi)du}{\int_{-\infty}^{\infty} l_d(y_d,u,\phi)du} \quad . \tag{3.63}$$

In most cases ϕ is not known. Thus, β and σ_u have to be estimated. By using $\widehat{\phi} = (\widehat{\beta}, \widehat{\sigma}_u)$ instead of ϕ the area-level *empirical* best predictor is obtained (AEBP).

$$\widehat{\theta}_{d,\text{AEBP}} = \frac{\int_{-\infty}^{\infty} h_d(u, \widehat{\phi}) l_d(y_d, u, \widehat{\phi}) du}{\int_{-\infty}^{\infty} l_d(y_d, u, \widehat{\phi}) du} \tag{3.64}$$

However, equation (3.64) still involves solving two integrals for each area. In Figure 3.2 the function that is to be integrated is plotted for certain constellations of parameters and in Figure 3.3 the same is plotted on the log-scale. As can be seen, this function has a high peak maximum. The higher n the steeper the function becomes, and thus the narrower is the interval where the function is notably above zero. A smaller σ_u again makes the function narrow, but additionally shifts the peak towards zero. Whilst the magnitude of n and σ_u has mainly an impact on the steepness of the function, the observed mean $\frac{y}{n}$ and the logit of estimated synthetic mean $\text{logit}(\mu) = \eta$ notably shift the graph on the x-axis. Negative values of η, i.e. $\mu < 0.5$ shift the graph to the right. In contrast, for an observed mean $\frac{y}{n}$ lower than 0.5, the graph is shifted to the left and vice versa.

3.3.5.1 Computation of the AEBP

As seen in Figures 3.2 and 3.3 the calculation of the integrals can be extremely problematic, especially in larger areas, i.e. $n \gg 1000$. Therefore different approaches will be presented to obtain an approximation to the integrals.

Integration via Monte-Carlo

The most straightforward way is to do a Monte-Carlo integration. As one component of the integral is the pdf of the normal distribution, one can use normally distributed random variables $u^{(r)} \sim \text{N}(0, \sigma_u^2), r = 1 \ldots R$ for the integration. The approximation would then be

$$\widehat{\theta}^{\text{MC}}_{d,\text{AEBP}} = \frac{\sum_{r=1}^{R} \left[\frac{e^{\overline{X}_d\beta + u^{(r)}}}{1 + e^{\overline{X}_d\beta + u^{(r)}}}\right]^1 \binom{n_d}{y_d} \left[\frac{e^{\overline{X}_d\beta + u^{(r)}}}{1 + e^{\overline{X}_d\beta + u^{(r)}}}\right]^{y_d} \left[1 - \frac{e^{\overline{X}_d\beta + u^{(r)}}}{1 + e^{\overline{X}_d\beta + u^{(r)}}}\right]^{n_d - y_d}}{\sum_{r=1}^{R} \binom{n_d}{y_d} \left[\frac{e^{\overline{X}_d\beta + u^{(r)}}}{1 + e^{\overline{X}_d\beta + u^{(r)}}}\right]^{y_d} \left[1 - \frac{e^{\overline{X}_d\beta + u^{(r)}}}{1 + e^{\overline{X}_d\beta + u^{(r)}}}\right]^{n_d - y_d}} \tag{3.65}$$

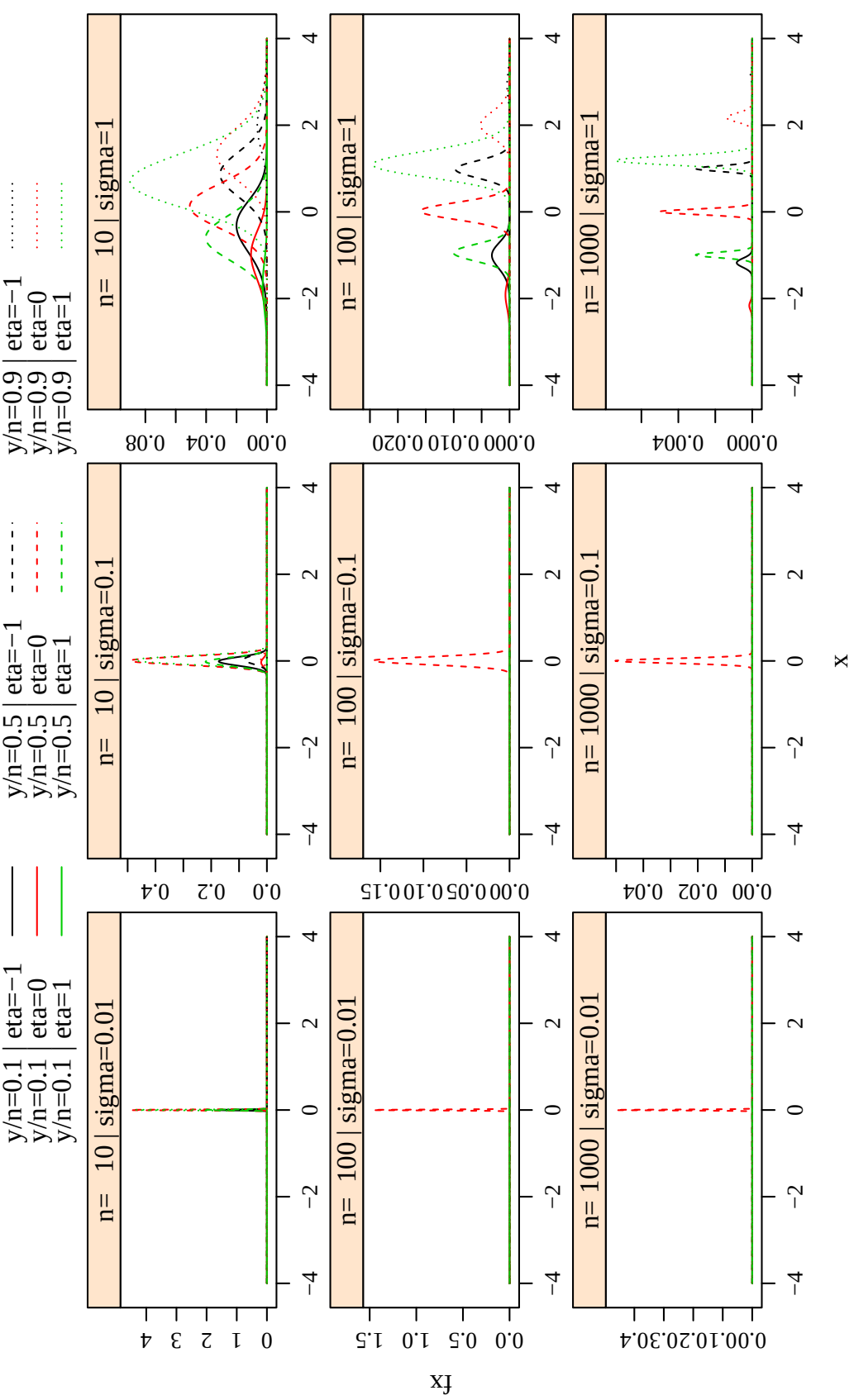

Figure 3.2: Graph of the function $\frac{1}{\sigma_u\sqrt{2\pi}}\binom{n_d}{y_d}\left(\frac{e^{\bar{X}_d\beta+u}}{1+e^{\bar{X}_d\beta+u}}\right)^{y+m}\left(1-\frac{e^{\bar{X}_d\beta+u}}{1+e^{\bar{X}_d\beta+u}}\right)^{n-y} e^{-\frac{u^2}{2\sigma_u^2}}$ for several parameter constellations.

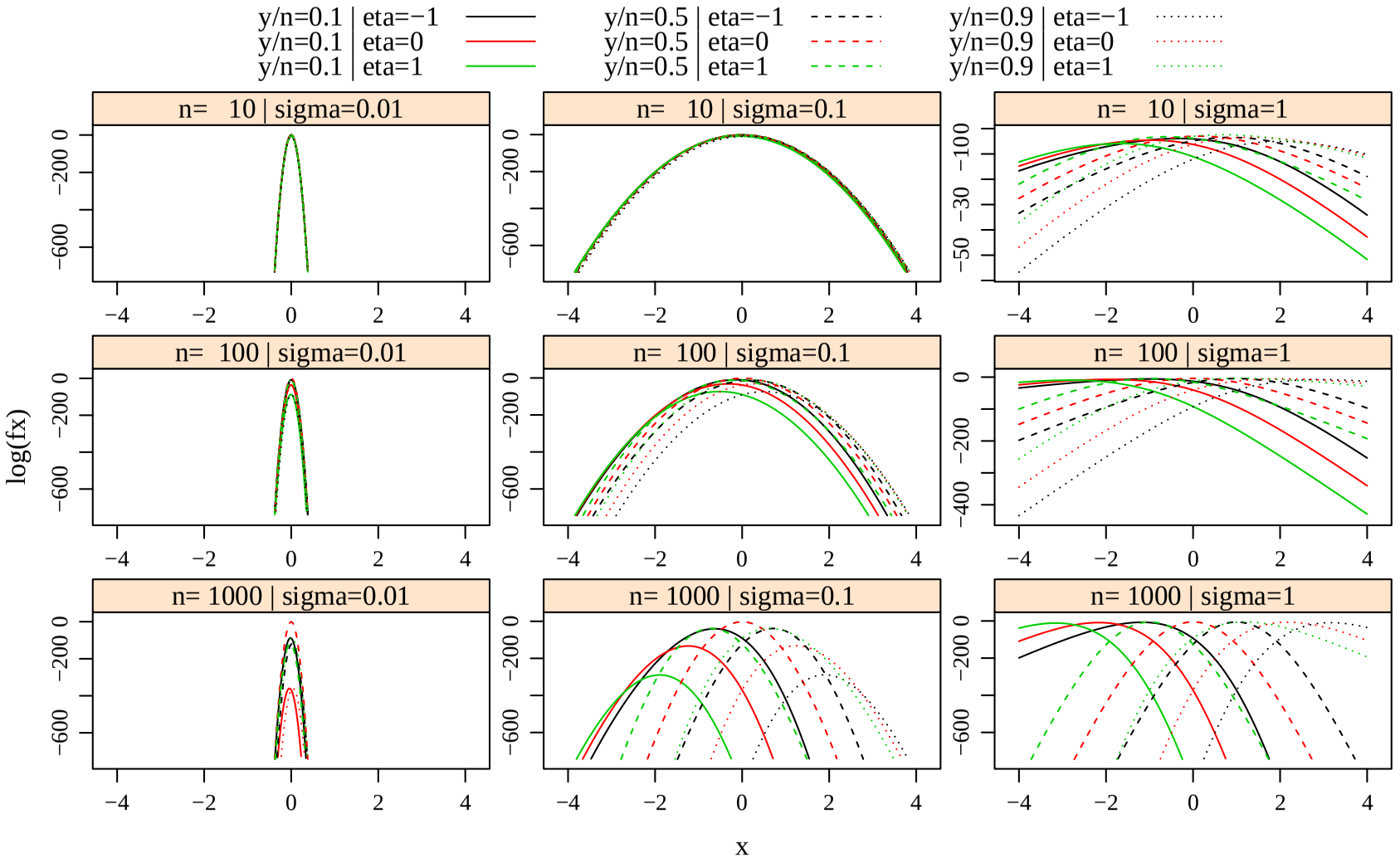

Figure 3.3: Graph of the log of the function $\frac{1}{\sigma_u\sqrt{2\pi}}\binom{n_d}{y_d}\left(\frac{e^{\overline{X}_d\beta+u}}{1+e^{\overline{X}_d\beta+u}}\right)^{y+m}\left(1-\frac{e^{\overline{X}_d\beta+u}}{1+e^{\overline{X}_d\beta+u}}\right)^{n-y} e^{-\frac{u^2}{2\sigma_u^2}}$ for several parameter constellations.

Integration via Gauß-Hermite Quadrature

When using Gauß-Hermite quadrature, the first thing to decide is how many nodes will be used. The more nodes are used, the more precise is the approximation. However, an increase of nodes also increases computation time. The approximation via Gauß-Hermite quadrature for $r = 1..R$ nodes $u^{(r)}$ is given by

$$\widehat{\theta}_{d,\text{AEBP}}^{\text{GHQ,R}} = \frac{\sum_{r=1}^{R} \omega^{(r)} \left[\frac{e^{\overline{X}_i^T\beta+u^{(r)}}}{1+e^{\overline{X}_i^T\beta+u^{(r)}}}\right]^1 \binom{n_d}{y_d} \left[\frac{e^{\overline{X}_i^T\beta+u^{(r)}}}{1+e^{\overline{X}_i^T\beta+u^{(r)}}}\right]^{y_d} \left[1-\frac{e^{\overline{X}_i^T\beta+u^{(r)}}}{1+e^{\overline{X}_i^T\beta+u^{(r)}}}\right]^{n_d-y_d}}{\sum_{r=1}^{R} \omega^{(r)} \binom{n_d}{y_d} \left[\frac{e^{\overline{X}_i^T\beta+u^{(r)}}}{1+e^{\overline{X}_i^T\beta+u^{(r)}}}\right]^{y_d} \left[1-\frac{e^{\overline{X}_i^T\beta+u^{(r)}}}{1+e^{\overline{X}_i^T\beta+u^{(r)}}}\right]^{n_d-y_d}} \tag{3.66}$$

When using Gauß-Hermite quadrature, the choice of the number of nodes is critical for the success of the approximation to the integral. This fact is visualized in Figure 3.4. If only a small number of nodes are chosen, it may well be that none of the nodes will have a function value notably above zero. Even with 99 nodes in the example given in Figure 3.4, only 22 nodes lie in the area of interest and thus 77 of the nodes were computed without much gain. This problem can be overcome by either raising the number of nodes or by transforming the function of interest in such way that it has a more *appropriate* location and shape. That is, one should try to bring the mass of the area under the function near to the point $x = 0$. This can be done by substituting an appropriate term in the function of interest. As seen before, one important determinant for the shape of the function is the σ_u^2. Therefore a shift of the function by an amount ξ and a scaling by the factor σ_u^2 seems reasonable.

It is not clear how to obtain the optimal ξ without running into even more computational problems. However, as the function has definitely only one region of interest (the u^2 is dominant for large absolute values of u), the maximum of the function seems the natural choice for the shift. This maximum is analytically difficult to trace, since for the different constellations of the parameters, it can lie in a wide region. In most cases a Newton-Raphson approximation is successful and fast in finding the maximum of the function.

As the Euler function is a strictly monotonic ascending function, it suffices to find the maximum of a function $g(x)$ in order to obtain the maximum of $e^{g(x)}$. The functions in equation (3.64) can also be written as:

$$\left(\frac{e^{x\beta+u}}{1+e^{x\beta+u}}\right)^{y+m} \left(1-\frac{e^{x\beta+u}}{1+e^{x\beta+u}}\right)^{n-y} e^{-\frac{u^2}{2\sigma_u^2}} \quad , \tag{3.67}$$

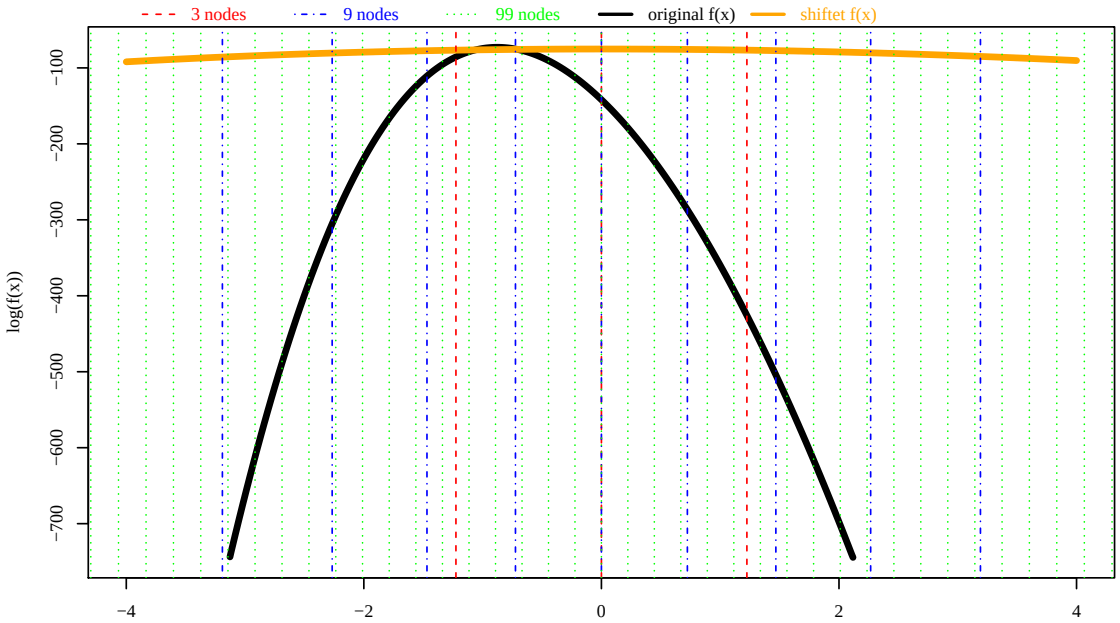

Figure 3.4: The problem of the choice of the number of nodes in the Gauß-Hermite quadrature in the example of: $x\beta = 2.94, y = 800, n = 1000, m = 1, \sigma_u = .1$

where $m = 1$ for the function in the enumerator and $m = 0$ for the function in the denominator. Equation (3.67) can also be written in the form $e^{g(u,\cdot)}$ with:

$$g_d(u,\cdot) = (m+y_d)(x_d\beta+u) - (m+n_d)log(1+e^{x_d\beta+u}) - \frac{u^2}{2\sigma_u^2} \tag{3.68}$$

In order to apply the Newton-Raphson algorithm the first and second derivative of $g(u,\cdot)$ have to be found. These are

$$\begin{aligned} g'_d(u,\cdot) &= (m+y_d) - (m+n_d)\frac{e^{x_d\beta+u}}{1+e^{x_d\beta+u}} - \frac{u}{\sigma_u^2} \\ g''_d(u,\cdot) &= (m+y_d) - \frac{(m+n_d)}{(1+e^{x_d\beta+u})^2} + \frac{(m+n_d)}{1+e^{x_d\beta+u}} - \frac{1}{\sigma_u^2} \end{aligned} \tag{3.69}$$

As the function $e^{g(u,\cdot)}$ is strictly positive, every extreme of the function must be at the same time a local maximum of the function. Therefore the Newton-Raphson can be applied in order to find the root of $g'_d(u,\cdot)$. The starting value may be chosen at $u_0 = 0$.

However, the Newton-Raphson algorithm for some constellations of the parameters fails, which can be identified by running into infinity. In this case, the bisection method or the regula falsi may be applied in order to approximate the root of the function $g'_d(u,\cdot)$.

The maximum ξ found either by Newton-Raphson, the bisection method or Regula-Falsi method can be used to perform a substitution of the function (3.67) of the following form:

Table 3.1: Comparison of the Gauß-Hermite Quadrature for the original and the shifted function with parameters $x\beta = 2.94, y = 800, n = 1000, m = 1, \sigma_u = .1, \xi = -0.8764274$

number of nodes	original function	shifted function
3	$9.35 \cdot 10^{-38}$	$4.05 \cdot 10^{-33}$
9	$1.69 \cdot 10^{-33}$	$4.05 \cdot 10^{-33}$
99	$5.07 \cdot 10^{-33}$	$4.05 \cdot 10^{-33}$
999	$4.05 \cdot 10^{-33}$	$4.05 \cdot 10^{-33}$
9999	$4.05 \cdot 10^{-33}$	$4.05 \cdot 10^{-33}$

$$\int_{-\infty}^{+\infty} \left(\frac{e^{x\beta+u}}{1+e^{x\beta+u}} \right)^{y+m} \left(1 - \frac{e^{x\beta+u}}{1+e^{x\beta+u}} \right)^{n-y} e^{-\frac{u^2}{2\sigma_u^2}} du \tag{3.70}$$

$$= \sigma_u \int_{-\infty}^{+\infty} \left(\frac{e^{x\beta+t\sigma_u+\xi}}{1+e^{x\beta+t\sigma_u+\xi}} \right)^{y+m} \left(1 - \frac{e^{x\beta+t\sigma_u+\xi}}{1+e^{x\beta+t\sigma_u+\xi}} \right)^{n-y} e^{-\frac{1}{2}\left(\frac{t\sigma_u+\xi}{\sigma_u}\right)^2} dt$$

with $t = \dfrac{u-\xi}{\sigma_u}$, $\quad \lambda(t) = t\sigma_u + \xi = u$ and $\lambda'(t) = \sigma_u$. The integration boundaries do not change as $\lambda(\infty) = \infty$ and $\lambda(-\infty) = -\infty$. The effect of this transformation can be seen in Figure 3.4. Whilst the original function drawn in black contains almost no nodes, even with a relatively high number of quadrature points, the transformed function in orange contains a considerable number of the nodes in use, even in the case of only a small number of quadrature points. Logically, the quality of the approximation is also very different when comparing both functions. In Table 3.1 the results of the integration via the Gauß-Hermite quadrature are compared for the original function (3.67) and the shifted function (3.70) when using 3, 9, 99, 999, or 9999 quadrature points. It can be seen that the integral of the original function is approximated in a much more unstable manner for a low number of nodes, than the one of the shifted function. Therefore, the shifted function is preferable to the original function for the integration via Gauß-Hermite quadrature.

Integration via Gauß-Konrod Quadrature

In the case of the Gauß-Konrod quadrature, similar problems as in the case of the Gauß-Hermite quadrature may arise. In particular, the problem may arise that no nodes lie in the region of the function where it is considerably above zero. In this case, the adaptive method will stop very early and approximate the integral with the value zero. Again, the same approach by substituting the function of interest may give better results.

Integration via Laplace Approximation

As stated before in equation (3.68) the function can be written in the form $e^{g(u,\cdot)}$. The second derivative $g''(u,\cdot)$ of $g(u,\cdot)$ is given in equation (3.69). The Laplace approximation can then be computed as:

$$\int_{-\infty}^{+\infty} e^{g(u,\cdot)} \approx e^{g(\xi,\cdot)} \int_{-\infty}^{+\infty} e^{-|g''(\xi,\cdot)|(x-\xi)^2/2} = \sqrt{\frac{2\pi}{|g''(\xi,\cdot)|}} e^{g(\xi,\cdot)} \quad , \tag{3.71}$$

ξ being the maximum of $g(u,\cdot)$. For the ratio of integrals in equation (3.64) one can write

$$\frac{\int_{-\infty}^{+\infty} e^{g(u,m=1,\cdot)}}{\int_{-\infty}^{+\infty} e^{g(u,m=0,\cdot)}} \approx \frac{\sqrt{\frac{2\pi}{|g''(\xi_1,m=1,\cdot)|}} e^{g(\xi_1,m=1,\cdot)}}{\sqrt{\frac{2\pi}{|g''(\xi_0,m=0,\cdot)|}} e^{g(\xi_0,m=0,\cdot)}} = \sqrt{\frac{|g''(\xi_0,m=0,\cdot)|}{|g''(\xi_1,m=1,\cdot)|}} \frac{e^{g(\xi_1,m=1,\cdot)}}{e^{g(\xi_0,m=0,\cdot)}} \quad . \tag{3.72}$$

Comparision of the Different Approaches

In Figure 3.5, the different approaches for the integration approximation of the $\widehat{\theta}_{\text{AEBP}}$ are compared for different constellations. As can be seen, the $\widehat{\theta}_{\text{AEBP}}$ acts according to the common small area estimates. For a low estimated inter area variation $\widehat{\sigma}_u^2$, the effect of the synthetic part $\eta_d = x_d\beta$ dominates the one of the observed mean $\frac{y}{n}$, whereas for a high inter area variation the observed mean is more dominant. Furthermore, for a larger number of observations n_d, more trust is placed on the observed mean and, for a lower number of observations, it relies more on the synthetic part.

In most cases the different approaches yield similar results for the ratio of the two integrals. However, there are differences, especially between the different approximation methods. While the quadrature rules work almost identical, the Monte Carlo approximation with 10,000 trials still has a considerable amount of variation. The Laplace approximation works quite well in most cases, but deviates in the extreme cases by 10^{-5}. Also, for small n, the Laplace approximation fails. This is not surprising, as it is a well known fact that the Binomial distribution only approximates the normal distribution reasonably for a higher number of observations.

All approximations have problems in the situation in which a low inter area variability meets a relatively large number of observations. In this case the denominator is approximated with zero, which yields implausible results for the $\widehat{\theta}_{\text{AEBP}}$. If this happens, then the function can be multiplied by a certain factor (both in the numerator and the denominator).

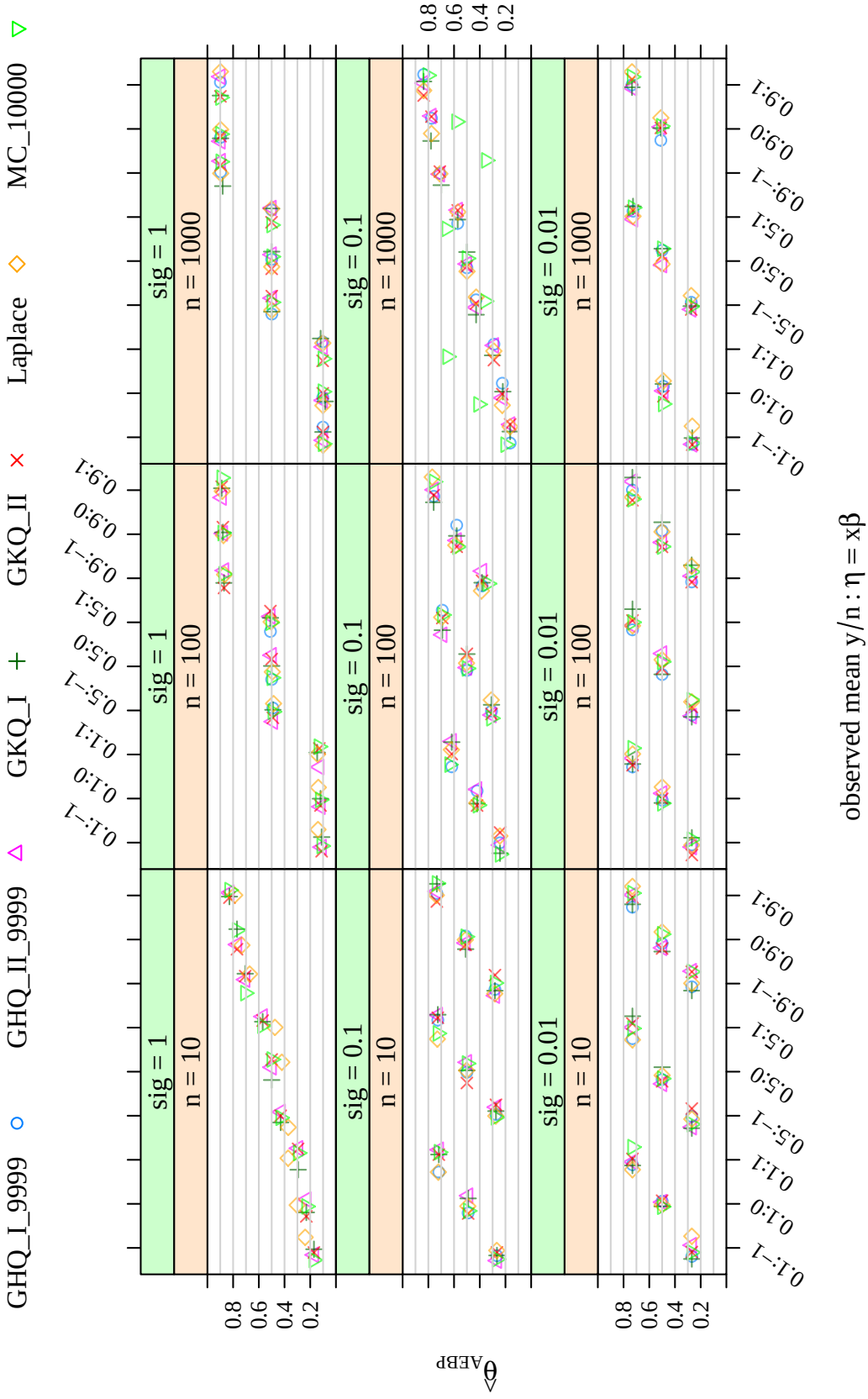

Figure 3.5: Comparison of the different approximation to the $\hat{\theta}_{\text{AEBP}}$. GHQ: Gauß Hermite Quadrature; GKQ: Gauß-Konrod quadrature; I: original function; II: shiftet function; 9999: 9999 quadrature points; MC10000 Monte Carlo approximation with 1000 trials.

3.3.6 MSE Estimation for the AEBP

3.3.6.1 MSE Estimation with the Jackknife Method

Early reasoning on the now so-called Jackknife method goes back to Mahalanobis (1946) who described a method implemented in the Indian Statistical Institute called *interpenetrating subsamples*, where two independent samples are drawn for each domain and then two estimates are computed for quality control (Hall, 2003). Quenouille (1949) used a similar approach to obtain bias reduction, by splitting the sample into two half-samples. Later on, Quenouille (1956) proposed to jackknife the Jackknife and thus obtained a second order correct jackknife estimator. Tukey (1958) proposed to use the Jackknife for the estimation of confidence intervals. According to Miller (1974), Tukey was the first to use the name Jackknife in unpublished work. Miller (1964) showed that the estimator has to have locally linear qualities, such as being a twice-differentiable function of the sample mean. For the estimation of the maximum, as a counterexample, he states that a nonnormal distribution may result in degeneracy at a point and drift to infinity for the Jackknife. For a thorough review of early Jackknife literature see Miller (1974) and for a broader picture see Efron and Tibshirani (1993) and Shao and Tu (1996). Jiang, Lahiri, and Wan (1998) and Chattopadhyay, Lahiri, Larsen, and Reimnitz (1999) propose to use the Jackknife for the estimation of the MSE of an empirical best predictor. Here reduced to an area-level EBP, they show that the MSE of $\widehat{\theta}^{\mathrm{EBP}}$ can be decomposed into:

$$\mathrm{MSE}\left[\widehat{\theta}^{\mathrm{EBP}}\right] = \mathrm{MSE}\left[\widehat{\theta}^{\mathrm{BP}}\right] + \mathrm{E}\left[\widehat{\theta}^{\mathrm{EBP}} - \widehat{\theta}^{\mathrm{BP}}\right]^2 \tag{3.73}$$

Further, they propose to estimate $\mathrm{MSE}\left[\widehat{\theta}^{\mathrm{BP}}\right]$ by

$$\widehat{\mathrm{MSE}}\left[\widehat{\theta}^{\mathrm{BP}}\right]_{Jack} = \widehat{\mathrm{MSE}}\left[\widehat{\theta}^{\mathrm{BP}}\right] - \frac{D-1}{D}\sum_{j=1}^{D}\left(\widehat{\mathrm{MSE}}\left[\widehat{\theta}^{\mathrm{BP}}_{-j}\right] - \widehat{\mathrm{MSE}}\left[\widehat{\theta}^{\mathrm{BP}}\right]\right) \quad , \tag{3.74}$$

and $\mathrm{E}\left[\widehat{\theta}^{\mathrm{EBP}} - \widehat{\theta}^{\mathrm{BP}}\right]^2$ by

$$\widehat{\mathrm{E}}\left[\widehat{\theta}^{\mathrm{EBP}} - \widehat{\theta}^{\mathrm{BP}}\right]^2 = \frac{D-1}{D}\sum_{j=1}^{D}\left(\widehat{\theta}^{\mathrm{EBP}}_{-j} - \widehat{\theta}^{\mathrm{EBP}}\right)^2 \quad . \tag{3.75}$$

The $\widehat{\theta}^{\mathrm{EBP}}_{-j}$ is the $\widehat{\theta}^{\mathrm{EBP}}$ when omitting area j in the estimation of the model parameters of the underlying model (c.f. Chattopadhyay et al., 1999), and $\widehat{\mathrm{MSE}}\left[\widehat{\theta}^{\mathrm{BP}}\right]$ depends on the BP at hand.

Jiang, Lahiri, and Wan (2002) extended this approach to a general model including not only mixed linear models but also generalized linear mixed models. Thus, this Jackknife is applicable to the before mentioned AEBP. Following Jiang et al. (2002) a Jackknife estimator for the MSE of the AEBP is

$$\begin{aligned} \mathrm{MSE}_d^{\mathrm{Jack}} = b_d(\widehat{\phi}) - \frac{D-1}{D}\sum_{j=1}^{D}\left\{b_d(\widehat{\phi}_{-j}) - b_d(\widehat{\phi})\right\} \\ + \frac{D-1}{D}\sum_{j=1}^{D}\left\{\widehat{\theta}_{d,\mathrm{AEBP}}(\widehat{\phi}_{-j}) - \widehat{\theta}_{d,\mathrm{AEBP}}(\widehat{\phi})\right\}^2 \quad , \end{aligned} \tag{3.76}$$

where $\widehat{\phi} = (\widehat{\beta}, \widehat{\sigma_u})$, $\widehat{\phi}_{-j}$ are the estimated model parameters omitting the area j, and

$$b_d(\widehat{\phi}) = \frac{\int_{-\infty}^{\infty}\left[\frac{e^{\overline{X}_d^T\widehat{\beta}+u}}{1+e^{\overline{X}_d^T\widehat{\beta}+u}}\right]^2 l_d(u)du}{\int_{-\infty}^{\infty} l_d(u)du} \quad . \tag{3.77}$$

Lohr and Rao (2009); J. N. K. Rao (2003) propose a modification of the Jackknife by Jiang et al. (2002) which yields a conditional MSE estimator instead of an unconditional one. J. N. K. Rao (2003, § 9.4) states that $\mathrm{MSE}\left[\theta^{\mathrm{BP}}\right] = \mathrm{E}\left[\mathrm{V}\left[\theta^{\mathrm{BP}}\right]\right]$. Then, he applies the Quenouille (1956) bias-reduction method to $\mathrm{V}\left[\theta^{\mathrm{BP}}\right]$ instead of $\mathrm{MSE}\left[\theta^{\mathrm{BP}}\right]$. According to him, this has two advantages. First, it saves the computation of $D+1$ times the term $b_d(\phi)$. Second, the leading term is now specific to the area, thus giving a conditional MSE. Lohr and Rao (2009) improve this approach by reducing the bias of the conditional MSE estimator. This is done by omitting the area d for the conditional MSE estimator for area d from the Jackknife correction term.

Jiang et al. (2002) show that their Jackknife has a bias of order $o(D^{-1})$, the conditional MSE estimator by J. N. K. Rao (2003) has an unconditional bias of order $\mathcal{O}(D^{-1})$ and that of Lohr and Rao (2009) has an unconditional bias of order $o(D^{-1})$. Further, Lohr and Rao (2009) show that their Jackknife also has a conditional bias of order $o_p(D^{-1})$. Within a simulation run on a model based population, Lohr and Rao (2009) compare these three alternatives. They find that for the estimation of the conditional MSE, their Jackknife MSE estimator has the lowest coefficient of variation. In contrast, for the estimation of the unconditional MSE, the one proposed by Jiang et al. (2002) has the lowest coefficient of variation in most cases.

While Lohr and Rao (2009) reduce the amount of computation by simplifying the Jackknife estimate, there is another way to reduce the burden of computation. If the number of areas D is very large, re-estimating D times the term (3.76) is time-intensive. Following Efron (1980, § 2.2) and Kott (2001), one can use a delete-a-group Jackknife instead of a delete-1-Jackknife. In analogy of his work, one can also apply this approach to the area-level case. In this case, the areas are randomly assigned to J equally sized groups. Instead of omitting D times one area in the estimation process, now J times a whole

group is thrown out. This is applied to both Jackknife parts of the $\text{MSE}_d^{\text{Jack}}$. Thus the delete-a-group *of areas* Jackknife is given by

$$\begin{aligned}\text{MSE}_d^{\text{GJack}} = b_d(\widehat{\phi}) - \frac{J-1}{J}\sum_{j=1}^{J}\left\{b_d(\widehat{\phi}_{-j}) - b_d(\widehat{\phi})\right\} \\ + \frac{J-1}{J}\sum_{j=1}^{J}\left\{\widehat{\theta}_{d,\text{AEBP}}(\widehat{\phi}_{-j}) - \widehat{\theta}_{d,\text{AEBP}}(\widehat{\phi})\right\}^2 \quad ,\end{aligned} \tag{3.78}$$

where j is one of the J groups and $\widehat{\phi}_{-j}$ is the vector of the estimated model parameters when all areas in group j are omitted in the estimation process. The choice of the number of groups is arbitrary. In general, the more groups used, the more exact is the Jackknife estimator. If every group consists of only one area, the original Jackknife estimator is obtained. In practice, one would argue with computation cost and time for the number of groups.

3.3.6.2 MSE Estimation With The Parametric Bootstrap Approach

The term *parametric bootstrap* goes back to Efron (1980, § 5.2), who briefly describes the idea behind it and gives an example for a bivariate model. This method is closely related to the so-called *(nonparametric-) bootstrap* by Efron (1979), and is widely used also in non small area applications (see e.g. Lahiri & Li, 2009a). Let

$$\Psi(X,F) \quad , \tag{3.79}$$

be a random variable, e.g. the estimator Ψ, as a function of the random variables $\chi = (\chi_1,...,\chi_K)$ with multivariate distribution F. The interest lies in a certain aspect of the distribution $\mathscr{L}(\Psi)$ of Ψ at the realized sample $\chi_k = \xi_k, \forall k = 1..K$. Then, the following algorithm gives an asymptotically correct approximation (Efron, 1982, § 5.1)

Algorithm 3.3 Nonparametric Monte-Carlo Bootstrap

1. Fit the nonparametric maximum likelihood estimator of F

$$\widehat{F}_n(\xi) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\xi_j \leq \xi} \quad . \tag{3.80}$$

2. Obtain $\Psi^*(\xi^*, \widehat{F}_n)$ by using a *bootstrap sample* from $\widehat{F}_n$:

$$\xi^* = (\xi_1'^*,..,\xi_n'^*)' \overset{\text{iid}}{\sim} \widehat{F}_n \quad . \tag{3.81}$$

3. Repeat step 2 independently R times.

The bootstrap estimator for $\mathscr{L}$ is then obtained by:

$$\mathscr{L}^* = \mathscr{L}(\Psi^*(\xi^*, \widehat{F}_n)) \tag{3.82}$$

In the complex survey design case, the i.i.d. assumption of the observation usually has to be dropped. Ranalli and Mecatti (2012) compare recent methods to account for sampling design and sampling weights. Davison and Hinkley (1997); Davison, Hinkley, and Young (2003) and Shao and Tu (1996) give a broad overview of different bootstrap methods. Lahiri (2003) studies different bootstrap methods for their applicability in small area estimation and survey sampling. He finds, that the parametric bootstrap method is a promising approach for the MSE estimation in small area estimation. An overview on resampling methods in surveys is provided by Gershunskaya, Jiang, and Lahiri (2009).

In contrast to the nonparametric bootstrap, the *parametric bootstrap* takes F_χ instead of $\widehat{F}_n$ in steps 1 and 2. F_χ denotes the distribution of the χ given the model used for the estimation of Ψ. Often some parts of χ are taken to be fixed and only one column is taken to be random. The random column is then usually the vector Y denoting the variable of interest.

Butar and Lahiri (2003); Hall and Maiti (2006) propose a parametric bootstrap approach for small area estimation. It is presented here directly applied to the MSE estimation of the AEBP.

Let $\mathscr{L}$ be the distribution of $(\theta_d - \widehat{\theta}_{d,\text{AEBP}})$. The distribution of $\mathscr{L}$ may be approximated by the *parametric bootstrap* distribution $\mathscr{L}^*$ of $(\theta_d^* - \widehat{\theta}_{d,\text{AEBP}}^*)$. θ_d^* and $\widehat{\theta}_{d,\text{AEBP}}^*$ are computed by the following algorithm starting with $r = 1$.

Algorithm 3.4 Parametric Bootstrap for the AEBP

1. Draw for each area $d = 1..D$ one $u_d^{*(r)}$ from the distribution $\text{N}(0, \widehat{\sigma}_u^2)$.
2. Calculate $\theta_d^{*(r)} = \dfrac{\text{e}^{\overline{X}_d\widehat{\beta}+u_d^{*(r)}}}{1+\text{e}^{\overline{X}_d\widehat{\beta}+u_d^{*(r)}}}$.
3. Draw $y_d^{*(r)}$ from the distribution $\text{Bin}(n_d, \theta_d^{*(r)})$.
4. Estimate the parameters $\widehat{\beta}^{*(r)}$ and $\widehat{\sigma}_u^{*(r)}$ in the same way as $\widehat{\beta}$ and $\widehat{\sigma}_u$, but using the re-sample $y^{*(r)}$ instead of y.
5. Calculate $\widehat{\theta}_{d,\text{AEBP}}^{*(r)}$ using the parameters $\widehat{\beta}^{*(r)}$ and $\widehat{\sigma}_u^{*(r)}$.
6. Calculate the distance $\theta_d^{*(r)} - \widehat{\theta}_{d,\text{AEBP}}^{*(r)}$.
7. Increase r by one.
8. Repeat Steps 1–7 R times.

From the distribution $\mathscr{L}^*$ the MSE of the point estimate $\widehat{\theta}^*_{d,\text{AEBP}}$ is calculable. It can be used as an estimator for the MSE of the point estimate $\widehat{\theta}_{d,\text{AEBP}}$. This MSE is then

$$\text{MSE}^*_{d,\text{AEBP}} = \mathbb{E}^*\left[(\theta^*_d - \widehat{\theta}^*_{d,\text{AEBP}})^2\right] \quad , \tag{3.83}$$

where $\mathbb{E}^*$ is the expectation given the distribution $\mathscr{L}^*$. Then the *parametric bootstrap* estimator of $\text{MSE}_{d,\text{AEBP}}$ is defined as:

$$\widehat{\text{MSE}}^{\text{Boot}}_{d,\text{AEBP}} = \text{MSE}^*_{d,\text{AEBP}} \quad . \tag{3.84}$$

There are several ways to construct confidence intervals from a bootstrap distribution (DiCiccio & Efron, 1996; DiCiccio & Romano, 1988). Among them are Efron's percentile method (Efron, 1979), the bias corrected percentile method (Efron, 1987) and the accelerated bias correction method by (DiCiccio & Efron, 1992). In contrast to Efron's percentile method, Hall (1992, 1997) proposes a percentile method which bootstraps a pivotal instead of a nonpivotal quantity.

Another widely used method is the bootstrap-*t* method by Babu and Singh (1983); Efron (1979). In this work two of these approaches are used. First, $\widehat{\text{MSE}}^{\text{Boot}}_{d,\text{AEBP}}$ is used to obtain bootstrap-*t* confidence intervals. These use

$$\begin{aligned} KI_{\alpha/2,1-\alpha/2}(\widehat{\theta}_{d,\text{AEBP}}) = \Big(& \widehat{\theta}_{d,\text{AEBP}} + \text{Q}_{\alpha/2}(t_{n_d-1})\sqrt{\widehat{\text{MSE}}^{\text{Boot}}_{d,\text{AEBP}}}; \\ & \widehat{\theta}_{d,\text{AEBP}} + \text{Q}_{1-\alpha/2}(t_{n_d-1})\sqrt{\widehat{\text{MSE}}^{\text{Boot}}_{d,\text{AEBP}}}\Big) \quad . \end{aligned} \tag{3.85}$$

This confidence interval is always symmetric around $\widehat{\theta}_{d,\text{AEBP}}$. However, the distribution of $\widehat{\theta}_{d,\text{AEBP}}$ is not necessarily symmetric.

Second, the percentile method by Hall (1992) is applied. It uses the quantiles of the bootstrap distribution of a pivotal quantity to obtain confidence intervals:

$$\begin{aligned} KI_{\alpha/2,1-\alpha/2} = \Big(& \widehat{\theta}_{d,\text{AEBP}} + \text{Q}_{\alpha/2}(\mathscr{L}^*); \\ & \widehat{\theta}_{d,\text{AEBP}} + \text{Q}_{1-\alpha/2}(\mathscr{L}^*)\Big) \quad . \end{aligned} \tag{3.86}$$

where $\text{Q}_\alpha(\mathscr{L}^*)$ denotes the α'th quantile of the distribution $\mathscr{L}^*$. The borders of the confidence interval are not necessarily symmetric around the point estimate. This may happen, e.g., if the bootstrap distribution $\mathscr{L}^*$ is not symmetric itself. For a deeper discussion on the estimation of confidence intervals for predictors see Chatterjee, Lahiri, and Li (2007, 2008).

3.3.6.3 The Parametric Double Bootstrap

The parametric bootstrap MSE estimate discussed in the section before is not second-order unbiased (Chatterjee & Lahiri, 2007). Therefore, further methods are needed in order to obtain a more reliable MSE prediction method. Booth and Hobert (1998) and Hall and Maiti (2006) proposed to use a double bootstrap strategy, where they build a second bootstrap distribution on top of the usual bootstrap distribution. However, as Chatterjee and Lahiri (2007) state, both approaches need analytical work to be done in order to be implementable to a new model. In contrast, Chatterjee and Lahiri (2007) propose a second order correct double bootstrap method that can be easily applied to a wide range of parametric two-level models. The algorithm applied to the AEBP is as follows

Algorithm 3.5 Parametric Double Bootstrap for the AEBP

1. Draw for each area $d = 1..D$ one $u_d^{*(r)}$ from the distribution $\mathrm{N}(0, \widehat{\sigma}_u^2)$.
2. Calculate and store $\theta_d^{*(r)} = \dfrac{\mathrm{e}^{\overline{X}_d\widehat{\beta}+u_d^{*(r)}}}{1+\mathrm{e}^{\overline{X}_d\widehat{\beta}+u_d^{*(r)}}}$.
3. Draw $y_d^{*(r)}$ from the distribution $\mathrm{Bin}(n_d, \theta_d^{*(r)})$.
4. Estimate the parameters $\widehat{\beta}^{*(r)}$ and $\widehat{\sigma}_u^{*(r)}$ in the same way as $\widehat{\beta}$ and $\widehat{\sigma}_u$, but using the re-sample $y^{*(r)}$ instead of y.
5. Begin the loop for the double bootstrap
   (a) Draw for each area $d = 1..D$ one $u_d^{**(rj)}$ from the distribution $\mathrm{N}(0, \widehat{\sigma}_u^{2*(r)})$.
   (b) Calculate and store $\theta_d^{**(rj)} = \dfrac{\mathrm{e}^{\overline{X}_d\widehat{\beta}+u_d^{**(rj)}}}{1+\mathrm{e}^{\overline{X}_d\widehat{\beta}+u_d^{**(rj)}}}$.
   (c) Draw $y_d^{**(rj)}$ from the distribution $\mathrm{Bin}(n_d, \theta_d^{**(rj)})$.
   (d) Estimate the parameters $\widehat{\beta}^{**(rj)}$ and $\widehat{\sigma}_u^{**(rj)}$ in the same way as $\widehat{\beta}$ and $\widehat{\sigma}_u$, but using the re-sample $y^{**(rj)}$ instead of y.
   (e) Calculate and store the estimators
   UMSE $\widehat{\theta}_d\left(y_d, \widehat{\phi}^{*(r)}\right)$ and $\widehat{\theta}_d\left(y_d, \widehat{\phi}^{**(rj)}\right)$
   CMSE $\widehat{\theta}_d\left(y_d^*, \widehat{\phi}^{*(r)}\right)$ and $\widehat{\theta}_d\left(y_d^{**}, \widehat{\phi}^{**(rj)}\right)$
   with $\widehat{\phi}^{*(r)} = \left(\widehat{\beta}^{*(r)}, \widehat{\sigma}_u^{*(r)}\right)$ and $\widehat{\phi}^{**(rj)} = \left(\widehat{\beta}^{**(rj)}, \widehat{\sigma}_u^{**(rj)}\right)$.
   (f) Increase j by one.

(g) Repeat steps 5.a-e J times.

6. Increase r by one.

7. Repeat steps 1–6 R times.

8. Calculate the expectations

UMSE

$$M_{1,d} = \mathbb{E}^* \left[\theta_d^* - \widehat{\theta}_d \left(y_d, \widehat{\phi}^* \right) \right]^2 \tag{3.87}$$
$$M_{2,d} = \mathbb{E}^* \mathbb{E}^{**} \left[\theta_d^{**} - \widehat{\theta}_d \left(y_d, \widehat{\phi}^{**} \right) \right]^2$$

CMSE

$$M_{3,d} = \mathbb{E}^* \left[\theta_d^* - \widehat{\theta}_d \left(y_d^*, \widehat{\phi}^* \right) \right]^2 \tag{3.88}$$
$$M_{4,d} = \mathbb{E}^* \mathbb{E}^{**} \left[\theta_d^{**} - \widehat{\theta}_d \left(y_d^{**}, \widehat{\phi}^{**} \right) \right]^2$$

The appeal of the parametric double bootstrap by Chatterjee and Lahiri (2007) is that one can obtain the usual unconditional MSE as well as the conditional MSE advocated by Fuller (1990) and Booth and Hobert (1998) without having to do any analytical derivations. An unconditional MSE estimator (UMSE^{PDBoot}) is given by

$$\text{CMSE}_d^{PDBoot} = \text{H}(M_{1,d}, M_{2,d} - M_{1,d}) \quad , \tag{3.89}$$

and an unconditional MSE estimator is obtained by,

$$\text{UMSE}_d^{PDBoot} = \text{H}(M_{3,d}, M_{4,d} - M_{3,d}) \quad , \tag{3.90}$$

(Chatterjee & Lahiri, 2007).

For the choice of the function H Chatterjee and Lahiri (2007) propose four different options. The functions H_2 and H_3 are those considered also by Hall and Maiti (2006).

$$\text{H}_1(M, \Delta_M) = (M - \Delta_M)\, \mathbb{I}_{M > \Delta_M} \tag{3.91}$$
$$\text{H}_2(M, \Delta_M) = (M - \Delta_M)\, \mathbb{I}_{\Delta_M \leq 0} + \left(M\, \text{e}^{-\Delta_M/(M-\Delta_M)} \right) \mathbb{I}_{\Delta_M > 0}$$
$$\text{H}_3(M, \Delta_M, n) = \left(M + n^{-1} \tan^{-1}(n\Delta_M) \right) \mathbb{I}_{\Delta_M \leq 0} + M^2 \left(M + n^{-1} \tan^{-1}(n\Delta_M) \right) \mathbb{I}_{\Delta_M > 0}$$
$$\text{H}_4(M, \Delta_M) = 2M \left(1 + \text{e}^{2\Delta_M/M} \right)^{-1}$$

The function H_1 is a kind of natural choice, as it returns the maximum of $(M_{1,d}, M_{2,d})$ for the CMSE and the maximum of $(M_{3,d}, M_{4,d})$ for the UCMSE estimate. For H_4 Chatterjee and Lahiri (2007) state that the order of the error for the CMSE and UMSE is of size $o((p+\nu)^2 D^{-1})$, where $p+\nu$ is the number of parameters of the model. The second order error is of size $\mathcal{O}((p+\nu)^2 D^{-1})$ in both cases.

In a simulation study using the Fay-Herriot model, Chatterjee and Lahiri (2007) find that H_2 seems to overestimate when the sampling variability is low and H_3 seems to underestimate if the sampling variability is large. They also find that H_1 and H_4 perform well for all the choices of sampling variability they made in their simulation study. The double bootstrap is very computing intensive, e.g. for $R = 99$ resamples in the first step and $J = 99$ resamples in the second step one needs 9801 resamples in total.

3.3.7 Unit-Level Logit Mixed Model Predictor

As already stated, linear models may produce unreasonable results for the prediction of proportions. González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2007); Saei and Chambers (2003) propose to use a generalized linear mixed model to predict small area proportions. The predictor is constructed straightforward, as in the Battese-Harter-Fuller estimator, but accounts for the nonlinearity of the link function. Because of the nonlinear link function, it is not possible to do the prediction over the population means of the covariates, but rather the covariates for all units in the population have to be known. This predictor will be called here the Binomial predictor BINP and is defined as follows

$$\widehat{\theta}_{d,\mathrm{BINP}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \frac{1}{1 + e^{-(X_{id}\widehat{\beta} + \widehat{u}_d)}} \quad , \tag{3.92}$$

where $\widehat{u}_d$ and $\widehat{\beta}$ are estimated by a generalized mixed model with binomial assumption on y and the logit function as link function (see Section 2.4). The BINPW is similar to the BINP, with the difference that is uses weighted model parameters. The weights are applied to the likelihood contributions of each unit, which are multiplied by their inverse inclusion probabilities.

From the same model a binomial synthetic estimator may be computed which is given by:

$$\widehat{\theta}_{d,\mathrm{BINSYN}} = \frac{1}{N_d} \sum_{i=1}^{N_d} \frac{1}{1 + e^{-(X_{id}\widehat{\beta})}} \quad . \tag{3.93}$$

If only categorical covariates are used, then it suffices to know the cross table which is defined by the covariates, instead of knowing each unit's covariate. However, in the case of continuous covariates, this relaxation on the information level of the population covariate means is not possible.

The estimation of the MSE of this estimator is very difficult. Due to the nonlinearity of the link function, the MSE estimator of Prasad and Rao (1990) is not directly applicable. González-Manteiga et al. (2007) propose two alternatives for estimating the MSE. The first alternative is to use the MSE estimator proposed by Prasad and Rao (1990) as an approximation. It takes the form

$$\widehat{\mathrm{MSE}}\left[\theta_{d,\mathrm{BINP}}\right] \approx g_{d,1}(\widehat{\phi}) + g_{d,2}(\widehat{\phi}) + 2g_{d,3}(\widehat{\phi}) \quad , \tag{3.94}$$

where in this case

$$\begin{aligned} g_{d,1}(\phi) &= \phi(1-\gamma_d)\sigma_{rd}^2 \quad , \\ g_{d,2}(\phi) &= \sigma_{rd}^2 \left[(\bar{x}_{rd}^{\sigma} - \gamma_d \bar{x}_d^{\sigma})(x' V_s^{-1} x)(\bar{x}_{rd}^{\sigma} - \gamma_d \bar{x}_d^{\sigma})'\right]_{d,d} \quad , \\ g_{d,3}(\phi) &= \frac{2}{\sum_{d=1}^{D} \sigma_{d\cdot}^2 (1-\gamma_d)^2} \left(\frac{\sigma_{d\cdot}}{(1+\phi\sigma_{d\cdot})^3 \sigma_{rd}^2}\right) \quad , \end{aligned}$$

with $\gamma_d = \dfrac{\phi}{\phi + 1/\sigma_d}$, $\overline{\sigma}_{rd} = \dfrac{\sum_{j\in r_d} \sigma_{dj}}{N_d - n_d}$, $\bar{x}_{sd} = \dfrac{\sum_{j\in r_d} \sigma_{dj} x_{dj}}{\sigma_{d\cdot}}$, and $\bar{x}_{rd} = \dfrac{\sum_{j\in r_d} \sigma_{dj} x_{dj}}{\sum_{j\in r_d} \sigma_{dj}}$. However, they state that this MSE estimator does not work well for small sample sizes and, even for the situation of higher sample sizes, they do not seem very enthusiastic about it. Therefore, this approach will not be considered further.

The second alternative is to use a special bootstrap method, which they call *small area wild bootstrap*. This bootstrap is a mixture of the *wild bootstrap* and a *finite population bootstrap*. A thorough overview of a wide range of bootstrap methods can be found in MacKinnon (2006); Shao and Tu (1996). For the case of heteroscedastic and hence nonidentically distributed observations, C.-F. J. Wu (1986) developed the *wild bootstrap* for regression models. It is closely related to bootstrapping residuals (see e.g. C.-F. J. Wu, 1986). In the residual bootstrap, bootstrapped residuals are used to form the new vector $y*$. Then the estimator is recomputed with the $y*$ instead of the y yielding the estimate $\psi*$. The $y*$ are in the simple case of the linear regression

$$y* = x\widehat{\beta} + e* \quad , \tag{3.95}$$

$e*$ being a simple random sample with replacement of the residuals $\widehat{e} = y - x\widehat{\beta}$. Obviously, this can only be done if the observations are identically distributed.

The wild bootstrap instead reuses the residuals from each observation for its own $y*$ by transforming it a bit. The $y*$ in the wild bootstrap is formed by

$$y*_i = x_i\widehat{\beta} + e_i\xi_i \quad , \tag{3.96}$$

where ξ is a random variable with $\mathrm{E}[\xi] = 0$ and $\mathrm{E}\left[\xi^2\right] = \mathrm{E}\left[\xi^3\right] = 1$. For some choices of ξ see R. Y. Liu (1988) and Mammen (1993). For an overview on some bootstrap methods for finite populations see Booth, Butler, and Hall (1994). The general idea behind it is to generate repeatedly finite populations resembling the finite population of interest.

Another possible method would be to use a parametric bootstrap to create the $y*$ by drawing from the model distribution. However, all these bootstrap methods have a common flaw; they need the estimator to be recomputed for each resample. In the case of the generalized linear mixed model at hand, this is quite cumbersome. For example, in the Swiss Structural Survey, a single estimation takes about 40 to 60 minutes. Just by taking 99 bootstrap resamples, which already is a low number, it would take about 3 days for one estimate. Thus, bootstrapping is in practice still not feasible for these kinds of estimators in surveys with a high number of total observations and/or a large population.

3.4 Prediction With An Additional Information Source on an Intermediate Aggregation Level

In small area estimation the auxiliary variables are of utmost importance for a precise prediction of population parameters. Usually, these auxiliary variables are taken from the population registers. In some countries the population registers have a large amount of different variables on the citizens. This is especially the case in northern European countries like Sweden and Finland. In other countries, such as Germany or Switzerland, there is only sparse information about the citizens in the population registers. Furthermore, for many population parameters of interest, the few existing variables have low predictive quality. The question arises as to how to enrich the population registers, so that more predictive quality can be obtained.

Unfortunately, additional registers often are not accessible due to legal restrictions. One legal reason for not obtaining additional register information is the issue of disclosure. Lahiri and Larsen (2005) propose to use record-linkage in order to enhance the use of third party registers where no personal identification number is available. However, the disclosure problem should not be an impediment if the register information is aggregated to a certain amount. The simplest case would be if the register information was made available at area level. Then, the Fay-Herriot estimator could be used directly, and record-linkage would not be mandatory. However, generally, the more detailed the available register information is, the better a model fits and thus better predictions are possible.

Instead of aggregating the additional register to the area-level, another possibility is to aggregate them to some sort of subpopulation of each area. This can be, e.g., a total giving the cross-combination of age classes and gender or proportions in addresses. As this information is between the unit-level and area-level, it will be referred to as intermediate-level information.

The formulas for the estimation and the prediction in the cases of area and unit-level additional information remain the same as before. In the case of the intermediate-level additional registers, the model building changes. As the covariates are now some population parameter at an intermediate aggregation level, the dependent variable also has to be aggregated.

Let A be a partition of the population $\mathscr{U}$ with J cells. Then A_d contains the units which belong in area d partitioned as A ($A_d := A \cap \mathscr{U}_d$). Furthermore, let A_d^j denote the j'th cell of partition A in area d, and τ_{y,A_d^j} be the total of variable y in the cell j of A_d. In this case one has in each of the D areas J observations τ_{y,A_d^j} of the totals of the variable of interest y.

The linear regression can now be applied on this data, using the cells of the partition as covariates. It is not possible to use an intercept in the model matrix as it would become multicollinear. The linear regression model then has the form:

$$\underbrace{\begin{pmatrix} \tau_{y,A_1^1} \\ \tau_{y,A_1^2} \\ \vdots \\ \tau_{y,A_1^J} \\ \tau_{y,A_2^1} \\ \vdots \\ \tau_{y,A_D^J} \end{pmatrix}}_{:=y} \sim \mathrm{N}\left(\underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}}_{:=x} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{pmatrix}}_{:=\beta}, I_{(D\cdot J)}\sigma_e^2 \right) . \tag{3.97}$$

In terms of the general linear regression model 2.1 the τ_{y,A_d^j} may be seen as the y for the cell j in the area d and x is a vector of dimension D of Identity matrices of size J. The observations are then all cells in all areas.

Up to this point, the additional registers are not considered. However, if the additional registers can be made available on the cells A_d^j, then this information can be incorporated into the estimation. Let ξ_k be the k'th vector of the K variables obtained from the additional register. Then the model in (3.97) can be extended to incorporate this information in the following way:

$$y \sim \mathrm{N}\left(\left(x, \xi_1, \ldots, \xi_K\right) \begin{pmatrix} \beta \\ \beta_{J+1} \\ \vdots \\ \beta_{J+K} \end{pmatrix}, I_{(D\cdot J)}\sigma_e^2 \right) \tag{3.98}$$

The assumption that all cross-combinations of cells are iid distributed over all areas is somewhat strong, as usually the areas have at least different levels. But taking the areas

into the fixed effects would lead to a Regression model with too many parameters. Hence, a mixed model is for practicability reasons more appropriate. In terms of a random intercept model as used by the FH or BHF, the areas, the cells or both can be used as random effects. However, in order to use solely the cells as random effect, it would be desirable to have a large number J of them. Another possibility would be to use a so-called two-level model (§ 5.5.4 J. N. K. Rao, 2003 and Moura & Holt, 1999). In this model, random coefficients are also possible. This model can cope for area differences in the level of the fixed effects parameters.

Furthermore, one has to be careful in adding a lot of additional covariates ξ as, due to the building of cells, less *observations* are available in contrast to the unit-level models. If too many parameters have to be estimated, the variability of the β estimate rises. Therefore, the success of this approach relies on few additional covariates ξ which can actually explain much of the variability of the accumulated original y's in the cell A_d^j.

Chapter 4

Variance Reduced Parametric Bootstrap MSE Estimates

4.1 The Parametric Bootstrap from the Monte-Carlo View

For the practitioner and the data producer in national institutes, the information pertaining to the precision of an estimate is very important. However, some of the small area methods presented in chapter 3 do not have an analytical approximation to the variance or MSE of the point estimate. In this case, only resampling methods may aid in finding an appropriate precision estimate, e.g., for the AEBP described in section 3.3.5. One big drawback of resampling methods is that they generally require extremely large computation times. For example, in the case of the Swiss Structural Survey, a single estimation of the AEBP takes approximately 14 seconds. A parametric bootstrap with 99 bootstrap samples would therefore take a little more than 23 Minutes. However, 99 resamples is not a very large number, considering that in the Swiss Structural Survey there are over 2800 areas. In other words, an over 2800 dimensional Problem is approximated by only 99 resamples. Therefore, in practice, much more than 99 resamples would be used, in order to obtain a reliable precision estimate. Even more problematic is the use of double bootstrap (see section 3.3.6.3), as this requires many more evaluations of the estimator. A parametric double bootstrap with 99 resamples in the first and second stage would take more than 38 hours.

Often, there exists a set of possible combinations of the covariates used in the small area estimator of choice. If it is to check which of these combinations is going to be used for the data production, the precision estimates produced for them are also of interest. However, comparing many models would lead to increased problems for the data producer. Depending on the estimator and minimum number of bootstrap samples required, the computations can easily take a couple of days. It is easy to see that the parametric bootstrap methods in many situations will only be feasible if computation time is reduced by a considerable amount.

Recalling the parametric bootstrap for the AEBP in equation (3.83) a more general Version for the MSE estimate may be written as

$$\mathrm{MSE}^*_{d,\mathrm{EST}} = \mathbb{E}^* \left[(\psi^*_d - \widehat{\psi}^*_{d,\mathrm{EST}})^2 \right] \quad .$$

Now the right hand side is written in function of the distribution of $y|X,Z$.

$$\mathrm{MSE}^*_{d,\mathrm{EST}} = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} (\psi_d - \widehat{\psi}_{d,\mathrm{EST}})^2 f_{y|X,Z}(u_1,\dots,u_D,e_1\dots,e_n)\, du_1 \dots du_D de_1 \dots e_n \quad . \tag{4.1}$$

Simplifying the equation (4.1) one can write $h(u) := (\psi_d - \widehat{\psi}_{d,\mathrm{EST}})^2$ and $f_{u,e} := f_{y|X,Z}$. Then the MSE estimate obtains the form

$$\mathrm{MSE}^*_{d,\mathrm{EST}} = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} h(u) f_{u,e}(u_1,\dots,u_D,e_1\dots,e_n)\, du_1 \dots du_D de_1 \dots e_n. \tag{4.2}$$

As the multivariate normal probability distribution function $f_{u,e}$ does not have a closed form integral, the equation (4.2) generally will not be tractable analytically either. The two choices at hand would be to use a numerical integration or a Monte-Carlo integration method for evaluating this multidimensional integral. In the area-level small area context, typically the number of areas D is large. In this case it will not be realistic to use a numerical integration method, as these suffer from the so called *curse of dimensionality* (Donoho, 2000), let alone the case of a unit-level small area estimator. By using a simple Monte-Carlo integration, the classical parametric bootstrap method results. This can be seen if one remembers that the Monte-Carlo estimate for the expectation of a function $g(U)$, where U is distributed according to the probability density function f_U is written as

$$\mathrm{E}[g(U)] = \int_{-\infty}^{+\infty} g(u) f_U(u) du \approx \mathrm{E}[g(U)]^{\mathrm{MC}} = \frac{1}{R} \sum_{r=1}^{R} g(u^{(r)}) \quad , \tag{4.3}$$

where $u^{(r)}$ is drawn independently $r = 1..R$ times from f_U (see e.g. Robert & Casella, 2004, § 3). It follows, so far, that the parametric bootstrap may be written as a special case of a Monte-Carlo integration problem. Thus, methods to improve estimates gained by Monte-Carlo integration may be helpful in estimating the parametric bootstrap MSE estimate as well.

4.2 Variance Reduction Methods for Monte-Carlo Integration

The plain Monte-Carlo integration is not a very efficient method. In the literature, many improvements on the plain Monte-Carlo integration have been proposed, in order to in-

crease efficiency of Monte-Carlo estimates. These methods are subsumed under the so called *variance-reduction methods*, which are broadly discussed by Frey and Rhodes (1999); Hesterberg (1996); Robert and Casella (2004) and Shapiro, Dentcheva, and Ruszczyński (2009), among many others. There are two basic ideas behind the variance reduction methods. The first one is that if the samples used in the Monte-Carlo integration are selected carefully, the rate of convergence of the Monte-Carlo approximation may increase. Besides the classical stratification for high dimensional problems, the Latin Hypercube Sampling was proposed and will be discussed below. The second one is to use some additional information for achieving this goal. Within this class of method, the focus will be laid on the use of control variates.

4.2.1 Latin Hypercube Sampling for Variance Reduction

A well known variance reduction method for survey sampling is the partitioning of the population into more homogeneous subgroups (strata). Instead of drawing simple random samples from the whole population, the sample is drawn randomly from the strata. If the strata are chosen such that the variability between the strata is relatively high to the variation within the strata, then a considerable reduction of the variance of the point estimates may be achieved (Neyman, 1934).

The stratified random sampling is determined by two choices:

1. the number and location of the strata (stratification)
2. the number of sampled units in each stratum (allocation)

Whereas the classical stratification is generally applied on a finite population, in Monte-Carlo integration the finite population can be replaced by a distribution. Therefore, the strata are defined along the support of the probability density function (for this variance reduction method see Ehrenfeld & Ben-Tuvia, 1962; Gaver, 1969). Let f_Y be the probability density function of the variable Y, which is defined on the support $\mathbb{S}$. For example, in the case of the normal distribution $\mathbb{S} = \mathbb{R}$. The strata then are the parts of a partition of $\mathbb{S}$. Usually the strata are taken to be subsequent intervals. So for a number of C strata, the strata may be defined as $I_c = (a_{c-1}, a_c)$, $c = 1 \ldots C$, where $a_0 = \inf(\mathbb{S}), a_C = \sup(\mathbb{S})$ and a_c denotes the supremum of the interval I_c. The choice of the a_c is arbitrary. A common praxis is to choose the a_c such that the resulting intervals are equal probable given the probability density function at hand. Therefore, a straightforward choice would be $I_c = \left(F_Y^{-1}\left((c-1)/C\right), F_Y^{-1}\left(c/C\right)\right)$, $c = 1 \ldots C$, where F_Y^{-1} is the inverse distribution function corresponding to f_Y. For equal probable intervals I_c, the number of sampled units in each stratum in proportional stratified random sampling is $n_c \equiv n/C$, $c = 1..C$. The variance reduction effect for proportional stratified random sampling is maximized

by setting $C = n$, n being the number of observations drawn in one run. The Monte-Carlo approximation uses the same formula as in simple random sampling from the distribution

$$\mathrm{E}\left[g(Y)\right]^{\mathrm{MC}} = \frac{1}{R}\sum_{r=1}^{R} g(y^{(r)}) \quad , \tag{4.4}$$

where, each $y^{(r)}$, $r = 1..R$ consists of a stratified random sample as described before. If these $y^{(r)}$ are used in formula (4.4), then $\mathrm{V}\left[\mathrm{E}\left[g(Y)\right]^{\mathrm{MC}}\right]_{StrRS} \leq \mathrm{V}\left[\mathrm{E}\left[g(Y)\right]^{\mathrm{MC}}\right]_{SRS}$ (see McKay, Beckman, & Conover, 1979 and Fishman, 1996, § 4.3).

In parametric bootstrap applications, especially in small area estimation, the distribution from which it has to be drawn is generally multidimensional. As already stated above, for an area-level estimate on the Swiss Structural Survey, there is easily a 2800-dimensional distribution over which one has to integrate. With the higher dimensionality, the stratification approach runs into troubles, as will be shown now.

Let Y be now a random variable from a K-dimensional distribution F_Y with probability density function f_Y. Again, an arbitrary partition will be the stratification. For simplicity, only the proportional stratified random sampling with equal probable strata is considered here. Let $I_c, c = 1 \ldots C$ be the equal probable strata. Let's assume that every marginal distribution k is univariate stratified with C_k strata. Then, the total number of strata is $C = \prod_{k=1}^{K} C_k$. Using the example of the Swiss structural survey, it is easy to see that this approach will not be feasible. The problem is that for the parametric bootstrap for an area-level estimator, a fixed number n of samples is predefined, which is the number of areas $n = D$. Therefore, the restriction is that a maximum of $C = n = D$ strata may be used. Now, assuming that each marginal distribution of Y is only split into two strata, then $C = \prod_{k=1}^{K} C_k = 2^K$. In the case of $K = 2800$, the minimal C can no longer be calculated with a standard calculator and is definitely larger than the fixed $n = 2800$ observations. The only way to overcome this problem is to reduce the number of strata. However, as a result, the stratification almost vanishes and thus will not be very helpful. In this case the effects tend to be negligible in comparison to simple random sampling.

As alternative, McKay et al. (1979) proposed a method called Latin-Hypercube-Sampling (*LHS*). A Latin-Square is a $n \times n$ matrix filled with n different symbols which only occur once in each row and each column (Euler, 1782). See Bose, Chakravarti, and Knuth (1960) for a computational algorithm to construct Latin-Squares. As Box (1980); Rapanos (2008) state, Fisher (1935) already used the Latin-Squares for his theory on the *design of experiments*. The use of Latin-Squares for sampling is described in Raj (1968, § 4.10).

A Latin-Hypercube, as proposed by McKay et al. (1979), is a generalization of Latin-Squares into the K-dimensional space. The sampling algorithm is described in Algorithm 4.1. Each of the $k = 1 \ldots K$ variables is partitioned into n equally probable intervals $I_i, i = 1 \ldots n$. From each interval i of each variable k one unit $y_{i,k}$ is drawn by simple random sampling. The resulting matrix Y is of dimension $n \times K$ with entries $Y_{i,k} = y_{i,k}$.

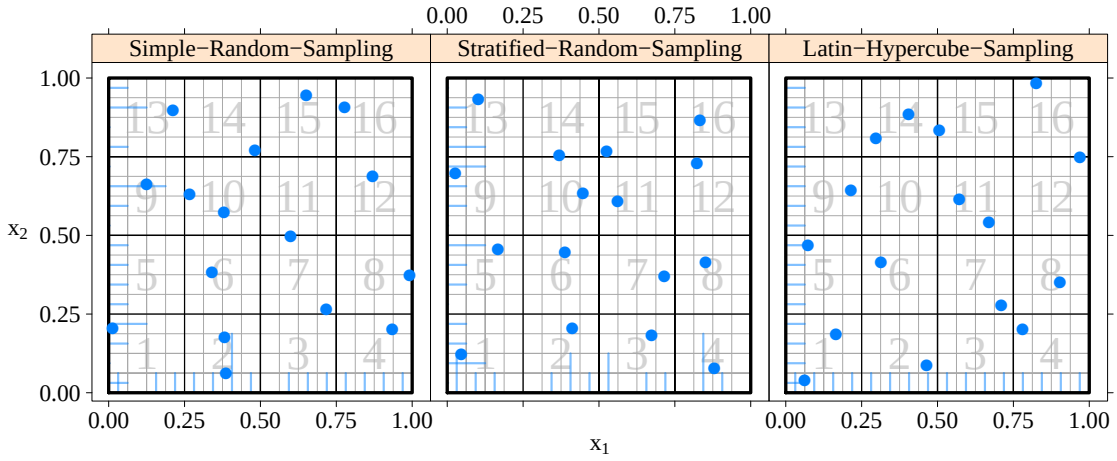

Figure 4.1: Comparison of the sampling schemes: Simple-Random-Sampling, Stratified-Random-Sampling and Latin-Hypercube-Sampling for a sample of $n = 16$ observations from a bivariate uniform distribution on the space $[0,1]^2$

Now the columns of the matrix Y are randomly permuted to form the matrix Y^{P} . The resulting observations are then the $y_{i.}$ entries of the matrix Y^{P}. By doing this, it is ensured that only one point is sampled in each cell of the hypercube. This algorithm draws from a multivariate equal distribution. In order to draw from a normally distributed variable, the *Inverse Transform Sampling* method can be applied (see Section A.1 in appendix or Robert & Casella, 2004, § 2.1.2.

Algorithm 4.1 Latin-Hypercube-Sampling from a K-dimensional Independently Uniformly Distributed Random Variable

Let x be a K-dimensional independently uniformly on $[0,1]^K$ distributed random variable from which one wishes to draw n samples.

1. Set r=1.
2. Draw $i = 1..n$ times $x_{i,k}^{(r)} \sim \text{Uniform}\left[\frac{(i-1)}{n}, \frac{i}{n}\right]$ for each $k \in \{1 \dots K\}$.
3. Permute the columns of X randomly.
4. Repeats steps 2 and 3 for $r = 2..R$.
5. Obtain the $R \times K$ sample-matrix $X = (x_1, x_k)$ of the permuted vectors x_*.

(McKay et al., 1979)

These three methods, SRS, StrRS and LHS are visualized exemplary in Figure 4.1. This

figure shows the example for a $K = 2$-dimensional uniformly in $[0,1]$ distributed random variable with $n = 16$ samples. Each circle shows one sampled point. The bars on the x and y axis show the marginal frequency in the 16 equally probable intervals $[(l-1)/n, l/n]$, $l = 1..16$ of the variables x_1 and x_2 respectively. In the case of SRS from the pdf f_X one can see that the sample is not distributed evenly on the sample space. E.g. no samples resulted in cells three and five, while two samples resulted in cells two and ten. Also the marginal frequency of the sampled x_1 and x_2 is not even.

The StrRS allocates the sample in such a way, that one sample is drawn from each of the cells. As 16 samples are drawn from a two dimensional random variable, one can construct four strata per variable thus resulting in $4^2 = 16$ strata cells. Therefore, by using StrRS the sample is allocated more evenly over the sample space than through SRS. Still, regions may result on the marginals without any samples. E.g. in cells 1..13, and 16 no sample is drawn from the lower quartile of the cells for variable x_2. LHS, in contrast, allocates the samples in such way, that both variables have one sample in each interval $[(l-1)/n, l/n]$ $r = 1..16$. This way, the marginal frequency is evenly distributed over the n intervals. However, in LHS it may happen, like in SRS, that larger regions as seen in cells ten and thirteen remain without samples. With orthogonal sampling, proposed by Owen (1992), one can trace this problem as well.

4.2.2 Control Variables for Variance Reduction

One application of variance reduction in bootstraps is presented by Hesterberg (1996), who uses control variates. This idea will be translated into the purpose of the parametric bootstrap estimation. A control variate is a random variable which is a function of the same input vectors as the function which is bootstrapped. As before, let $h(u,e)$ be the random variable produced within the parametric bootstrap. Then a function $g(u,e)$ is defined with known mean $\overline{g}$. Instead of now calculating the expectation of h via

$$\mathrm{E}\left[h(u,e)\right] = \frac{1}{R}\sum_{r=1}^{R} h(u^{(r)}, e^{(r)}) \quad ,$$

the control variate is introduced as a correction term

$$\mathrm{E}\left[h(u,e)\right]_{\mathrm{CV}} = \frac{1}{R}\sum_{r=1}^{R} h(u^{(r)}, e^{(r)}) + c\left(g(u^{(r)}, e^{(r)}) - \overline{g}\right) \quad . \tag{4.5}$$

As $\mathrm{E}\left[g(u^{(r)}, e^{(r)})\right] = \overline{g}$ and c is a constant it follows that $\mathrm{E}\left[c\left(g(u^{(r)}, e^{(r)}) - \overline{g}\right)\right] = 0$ and therefore $\mathrm{E}\left[h(u,e)\right]_{\mathrm{CV}} = \mathrm{E}\left[h(u,e)\right]$. The optimal constant c, in the sense of minimizing the variance of the estimator in (4.5), is given by

$$c = \frac{\mathrm{COV}\left[h(u,e),\, g(u,e)\right]}{\mathrm{V}\left[g(u,e)\right]} \tag{4.6}$$

(Hesterberg, 1996). This method allows for a reduction of the variance of the Monte Carlo estimate by the rate of $\mathrm{COR}\left[h(u,e), g(u,e)\right]^2$.

In practice, both $\mathrm{COV}\left[h(u,e), g(u,e)\right]$ and $\mathrm{V}\left[h(u,e)\right]$ are not known. Following Hesterberg (1996) these terms may be computed from the bootstrap resamples.

$$\widehat{c} = \frac{\widehat{\mathrm{COV}}\left[h(u,e), g(u,e)\right]}{\widehat{\mathrm{V}}\left[g(u,e)\right]} \tag{4.7}$$

The estimation induces a bias of order $\mathscr{O}(R^{-1})$ which, in most cases, will be very small. The central issue in order to apply this method is to define a function $g(u,e)$, which has a known mean and preferably a strong correlation with $h(u,e)$. This function will depend strongly on the estimator used. For a proof of concept a control variate for the parametric bootstrap MSE estimate of the FH is proposed. As for the FH, an analytical approximation is also available so the performance of the parametric bootstrap MSE estimate may be compared to the analytical MSE estimate.

The function $h(u,e)$ in the case for the estimation of a mean with the FH is given by

$$\begin{aligned} h(u,e)_{d,\mathrm{FH}} &= (\widehat{\mu}^*_{d,\mathrm{FH}}(\overline{X}\widehat{\beta}, u^*, e^*) - \mu^*_d(\overline{X}\widehat{\beta}, u^*, e^*))^2 \\ &= \left[\left(\overline{X}_d\widehat{\beta}^* + \gamma^*_d((\overline{X}\widehat{\beta} + u^*_d + e^*_d) - \overline{X}\widehat{\beta}^*)\right) - \overline{X}_d\widehat{\beta} + u^*_d\right]^2 \end{aligned} \tag{4.8}$$

and assuming that

$$\widehat{\beta} \approx \widehat{\beta}^*$$

this may be approximated by

$$\begin{aligned} h(u,e)_{d,\mathrm{FH}} \approx \dot{h}(u,e)_{d,\mathrm{FH}} &= (\gamma^*_d\,(u^*_d + e^*_d) - u^*_d)^2 \\ &= ((\gamma^*_d - 1)u^*_d + \gamma^*_d e^*_d)^2 \quad , \end{aligned} \tag{4.9}$$

and by further assuming that

$$\begin{aligned} (\widehat{\sigma}^*_u, \widehat{\sigma}^*_{e,d}) &\approx (\widehat{\sigma}_u, \widehat{\sigma}_{e,d}) \\ \ddot{h}(u,e)_{d,\mathrm{FH}} &= ((\gamma_d - 1)u^*_d + \gamma_d e^*_d)^2 \quad , \end{aligned} \tag{4.10}$$

where u^* and e^* for area d are independently normally distributed with mean 0 and variances $\widehat{\sigma}^2_u$ and $\widehat{\sigma}^2_{e,d}$.

Four arbitrary choices for $g(u,e)$ then may be

$$g^{(1)}_d(u,e) = (u+e)^2 \qquad \overline{g}^{(1)}_d = \sigma^2_u + \sigma^2_{e,d} \quad , \tag{4.11}$$

$$g^{(2)}_d(u,e) = ((\gamma_d - 1)u + \gamma_d e)^2 \qquad \overline{g}^{(2)}_d = (\gamma_d - 1)^2\sigma^2_u + \gamma^2_d\sigma^2_{e,d} \quad , \tag{4.12}$$

$$g^{(3)}_d(u,e) = (u)^2 \qquad \overline{g}^{(2)}_d = \sigma^2_u \quad , \tag{4.13}$$

$$g^{(4)}_d(u,e) = (e)^2 \qquad \overline{g}^{(3)}_d = \sigma^2_{e,d} \quad . \tag{4.14}$$

The correlations of these four functions with the approximation $\ddot{h}$ of h are

$$\mathrm{COR}\left[\ddot{h}(u,e)_{d,\mathrm{FH}},\, g_d^{(1)}(u,e)\right] = \frac{2(\gamma_d-1)^2\sigma_u^4 + 2\gamma^2\sigma_{e,d}^4 + 4(\gamma_d-1)\gamma_d\sigma_u^2\sigma_{e,d}^2}{2\left((\gamma_d-1)^2\sigma_u^2+\gamma^2\sigma_{e,d}^2\right)\cdot 2\left(\sigma_u^2+\sigma_{e,d}^2\right)} = 0 \quad , \tag{4.15}$$

$$\mathrm{COR}\left[\ddot{h}_{d,\mathrm{FH}},\, g_d^{(2)}(u,e)\right] = 1 \quad , \tag{4.16}$$

$$\begin{aligned}\mathrm{COR}\left[\ddot{h}_{d,\mathrm{FH}},\, g_d^{(3)}(u,e)\right] &= \frac{2(\gamma_d-1)^2\sigma_u^4}{2\left((\gamma_d-1)^2\sigma_u^2+\gamma^2\sigma_{e,d}^2\right)\cdot 2\sigma_u^2} \qquad (4.17)\\ &= \frac{(\gamma_d-1)^2\sigma_u^2}{2\left((\gamma_d-1)^2\sigma_u^2+\gamma^2\sigma_{e,d}^2\right)}\\ &= \frac{\sigma_{e,d}^2}{2(\sigma_{e,d}^2+\sigma_u^2)} \quad , \qquad (4.18)\end{aligned}$$

and

$$\begin{aligned}\mathrm{COR}\left[\ddot{h}_{d,\mathrm{FH}},\, g_d^{(4)}(u,e)\right] &= \frac{2\gamma_d^2\sigma_{e,d}^4}{2\left((\gamma_d-1)^2\sigma_u^2+\gamma^2\sigma_{e,d}^2\right)\cdot 2\sigma_{e,d}^2} \qquad (4.19)\\ &= \frac{\gamma_d^2\sigma_{e,d}^2}{2\left((\gamma_d-1)^2\sigma_u^2+\gamma^2\sigma_{e,d}^2\right)}\\ &= \frac{\sigma_u^2}{2(\sigma_{e,d}^2+\sigma_u^2)} \quad . \qquad (4.20)\end{aligned}$$

Function $g^{(1)}$ has no correlation to $\ddot{h}_{d,\mathrm{FH}}$ and thus will not reduce the variability of the estimate. As the correlation $\mathrm{COR}\left[\ddot{h}_{d,\mathrm{FH}},\, g_d^{(2)}(u,e)\right] = 1$ this function is supposed to deliver the best results in all settings. Of course, this correlation is higher than the true one, as the assumptions $\widehat{\beta}^* \equiv \widehat{\beta}$ and $\widehat{\sigma}^* \equiv \widehat{\sigma}$ will, in general, be violated in most of the resamples. For the same reason the approximation to the correlations of h with $g^{(3)}$ and $g^{(4)}$ is rather rough, however the picture is promising. The surfaces of the correlation of these functions are visualized in Figure 4.2. For the case, that σ_u^2 and σ_e^2 are of same magnitude, the functions $g^{(3)}$ and $g^{(4)}$ will not bring much improvement. If σ_u^2 is near to zero then it is preferable to use $g^{(3)}$, as even for small σ_u^2 the correlation is quite high. In contrast, if σ_e^2 is near zero, then the function $g^{(4)}$ seems the best choice for analogue reasons. In practice, the function g_3 may cause problems if an estimator for σ_u^2 is used which allows for zero estimates. In this case, the optimal c may not be computed as $g^{(3)}$ is constant and thus $\mathrm{V}\left[g^{(3)}\right] = 0$.

The resulting approximation to the reduction of the variance for the parametric bootstrap MSE estimate for the Fay-Herriot model is depicted in Figure 4.3. The functions $g^{(3)}$ and

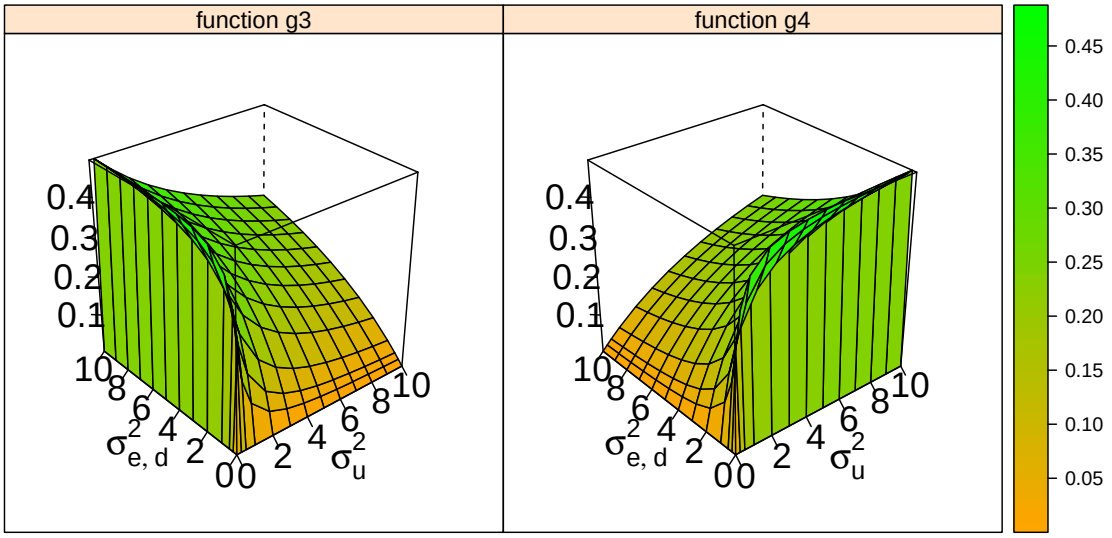

Figure 4.2: Correlation of the function $\ddot{h}$ and the functions $g^{(3)}$ and $g^{(4)}$

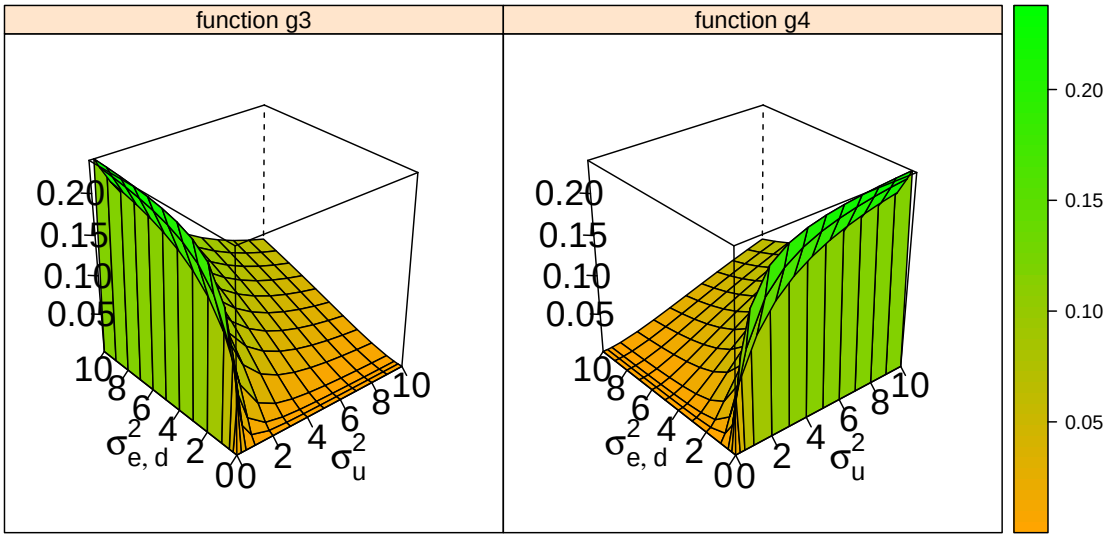

Figure 4.3: Rough approximation of the reduction of variance of the parametric bootstrap MSE estimate for the Fay-Herriot by using the functions $g^{(3)}$ and $g^{(4)}$ as control variates

$g^{(4)}$ seem only to be useful if σ_u^2 and σ_e^2 are small respectively. Again the function $g^{(2)}$ is not plotted, as the correlation $\text{COR}\left[\ddot{h}_{d,\text{FH}}, g_d^{(2)}(u,e)\right]^2 = 1$ and thus would mean that the parametric bootstrap MSE estimate would be estimated with variance 0. As will be seen in the following Monte Carlo Simulation this, of course, is not the case.

4.3 Performance of the Variance Reduced Parametric Bootstrap MSE Estimate

4.3.1 Setting of the Monte Carlo Simulation

In this section, the concurrent approaches to reduce the variance of the parametric bootstrap MSE estimate will be studied in the form of a model based simulation study. For a discussion on the differences between model-based and design-based Monte Carlo simulation see Section 5.1. FH will be used as the estimator. The variance covariance matrix of the random effects and the sampling errors in the case of the FH is a diagonal matrix (see Section 3.3.1). Therefore, the random vectors u, e needed for the parametric bootstrap approximation to the MSE are independent from each other. For this simulation, different populations are constructed according to the model used by the FH. The reason for this is, that by drawing from a population that coincides with the assumptions made in the estimation process, the disturbances available in the simulation are solely of a random nature. There is no systematic error when the superpopulation model is implemented correctly and the estimator is computed in the right way. One assumption of the Fay-Herriot model is somehow critical, namely that the u_i are randomly distributed. This is a helpful assumption for the estimation but, in fact, the deviation of the area mean from the population mean is not random, but rather is a fixed population parameter. Therefore, the population is generated in the following way:

$$
\begin{aligned}
y_d &\sim \text{N}(x_d\beta + u_d, \sigma_{e,d}^2) \\
x_d &\sim \text{MVN}\left((20,10), \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}\right) \\
u_d &\sim \text{N}(0, \sigma_u^2)
\end{aligned}
$$

Hereby it is assumed, that the area deviation from zero remains constant over all populations. That is, the x_d and also the u_d are generated only once, while the $y_d = x_d\beta + u_d + e_d$ are generated for every run randomly by drawing the e_d from a multivariate normal distribution with means zero and variance covariance matrix $(\sigma_{e,1}^2, .., \sigma_{e,D}^2)I_{(D)}$. As the variance covariance matrix is a diagonal matrix, the single e_d are independently distributed. Therefore, it is also possible to draw the D sampling variances $e_d, d = 1..D$ from the univariate normal distribution with mean zero and variance $\sigma_{e,d}^2$.

Table 4.1: Population models for the simulative assessment of the use variance reduction methods for the parametric bootstrap MSE estimate for the Fay-Herriot estimator

population number	number of areas (D)	distribution of $\sigma^2_{e,d}$	value of σ^2_u
1	15	U(3,7)	5
2	40	U(3,7)	5
3	100	U(3,7)	5
4	15	U(0.01,0.1)	15
5	40	U(0.01,0.1)	15
6	100	U(0.01,0.1)	15
7	15	U(3,7)	0.1
8	40	U(3,7)	0.1
9	100	U(3,7)	0.1
10	15	U(.1,7)	5
11	40	U(.1,7)	5
12	100	U(.1,7)	5

By changing the parameters σ^2_u,$\sigma^2_{e,d}$ and the number of areas D some different populations are created. The different parameter constellations are presented in table 4.1.

The evaluation of the results however is not straightforward. The goal of variance reduction methods is to reduce the variability of the parametric bootstrap MSE estimate and decidedly not to reduce the MSE itself. In other words, a certain precision will be reached with less resamples. Therefore, it is of interest how variable the MSE estimate is in function of the number of bootstrap resamples. To visualize this, the following graph is proposed. The number of bootstrap resamples will be plotted on the x-axis against the 95% confidence interval containing the central 95% of the difference of the MSE estimate with r resamples to the *converged* MSE estimate with R resamples $MSE^{(r)} - MSE^{(R)}$. The narrower this band, the less variability in the MSE estimates. This graph will be used to evaluate the results in the following.

4.3.2 Monte Carlo Results for the Variance Reduction Methods for the Parametric Bootstrap MSE Estimate

All the above methods will be compared to the plain parametric bootstrap which, from the Monte Carlo view, uses a simple random sampling from the underlying distribution. Consequently, the plain parametric bootstrap will be denoted as *SRS*, the Latin Hypercube Sampling is named *LHS* and the control variates are called *function* $g^{(2)}$ and $g^{(3)}$. As the Simulations did not find any major differences in the performance of the methods when varying the number of areas, in this evaluation, for the sake of clarity, only the results for the populations with 15 areas are discussed (populations 1, 4, 7, and 10). The first 15 areas of the other populations are plotted in appendix B.

Results for population 1

In Figure 4.4, the graph proposed above is plotted for the Latin Hypercube Sampling approach in population 1. Each panel stands for one Area. As can be seen in this figure, using the Latin Hypercube Sampling does not reduce the variability of the parametric bootstrap MSE estimate, as the confidence bands are exactly overlapping. The same holds for the other populations, as can be deduced from Figures B.3, B.6, and B.9 in the appendix. Therefore, the Latin Hypercube Sampling in this application does not yield the desired variance reduction and is not discussed further.

The function $g^{(2)}$ was found to be the best control variate under the rough approximations made in the previous chapter. Therefore, it is thought to perform best under the concurrent control variates. In Figure 4.5 the performance of the plain parametric bootstrap (SRS) is compared to the control variate $g^{(2)}$. A high reduction of variance of the parametric bootstrap MSE estimate is observed when the control variate is used. Using the grey lines, one can see that the 95 % confidence band of the parametric bootstrap estimate under SRS with 200 resamples is about as wide as the one of function $g^{(2)}$ with 120 resamples. This is a reduction of 40% of resamples needed. As can be seen in Figure 4.6 the function $g^{(4)}$ also leads to a significant variance reduction, even though it is not as high as in the case of function $g^{(2)}$. In contrast, the function $g^{(3)}$ plotted in Figure 4.7 does not have any impact.

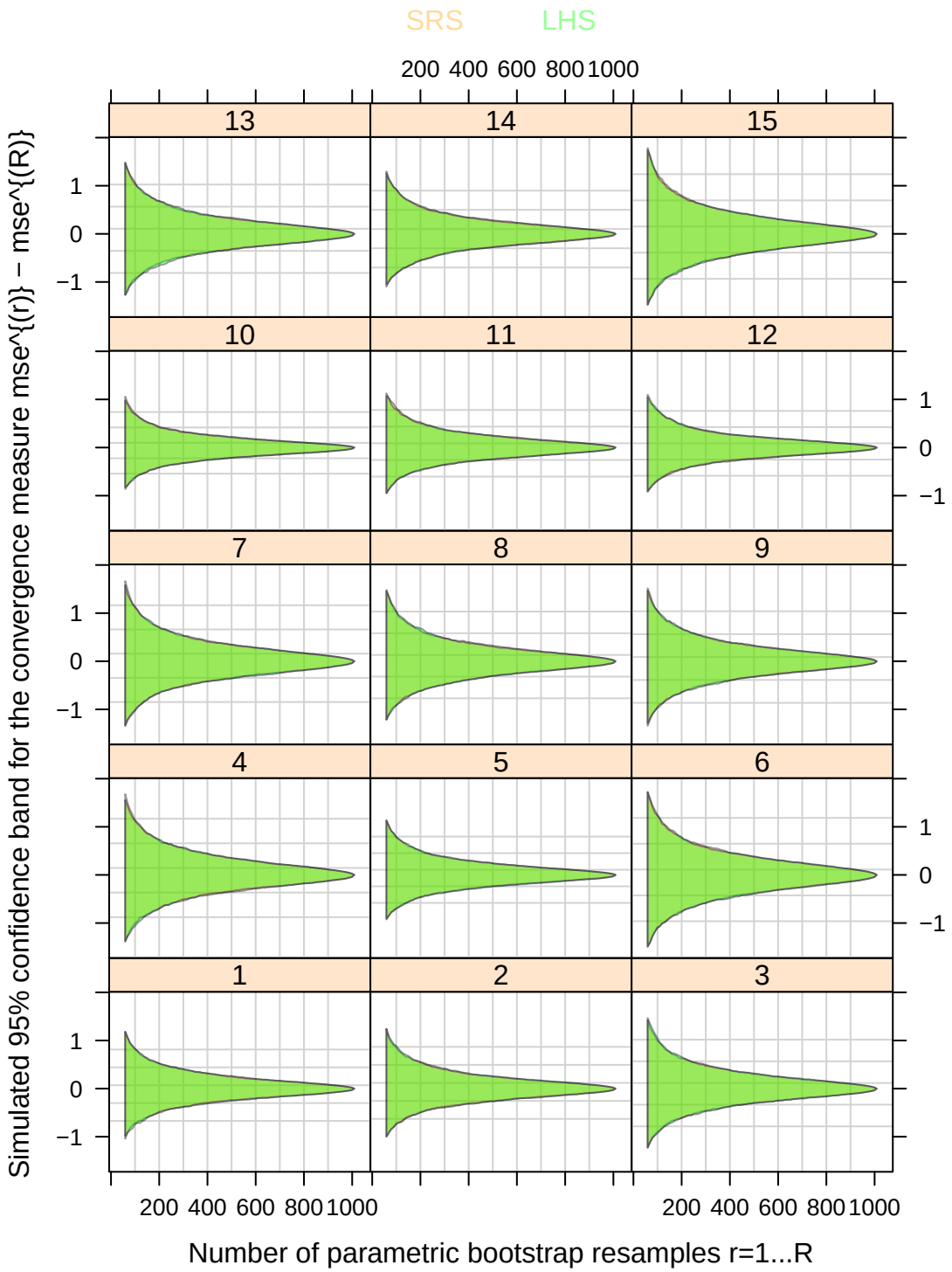

Figure 4.4: Using Latin Hypercube Sampling for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 1

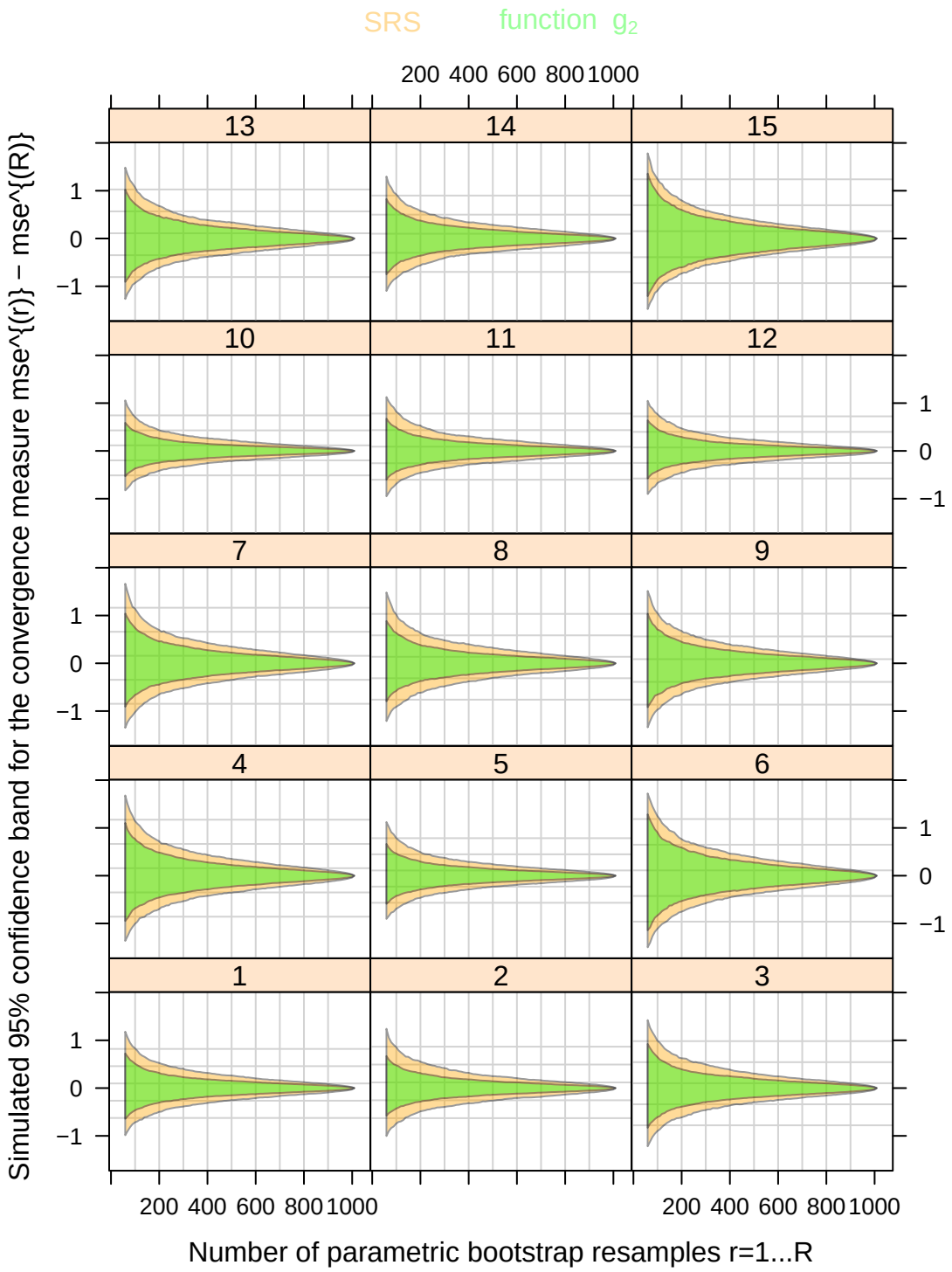

Figure 4.5: Using the control variate $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 1

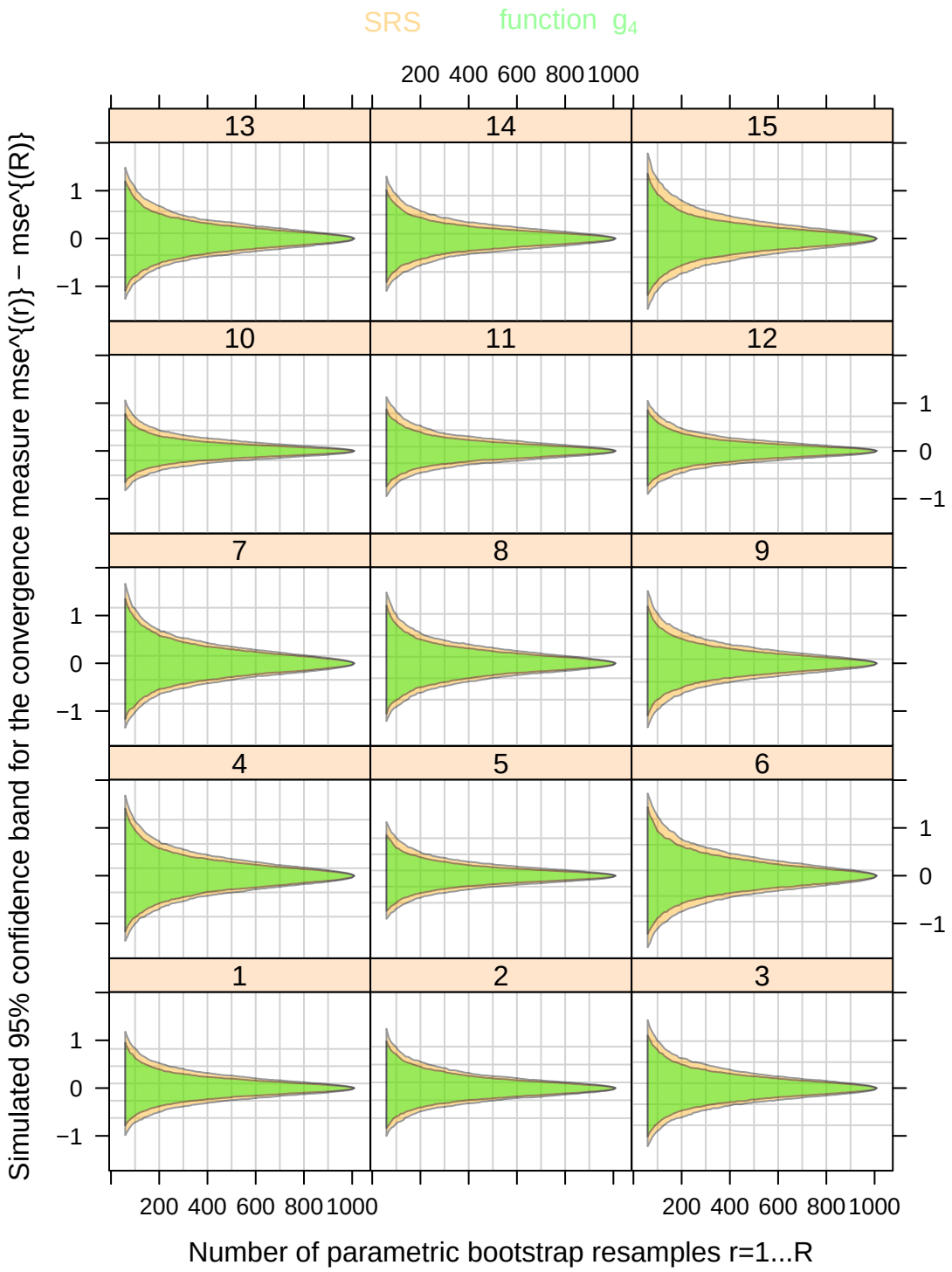

Figure 4.6: Using the control variate $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 1

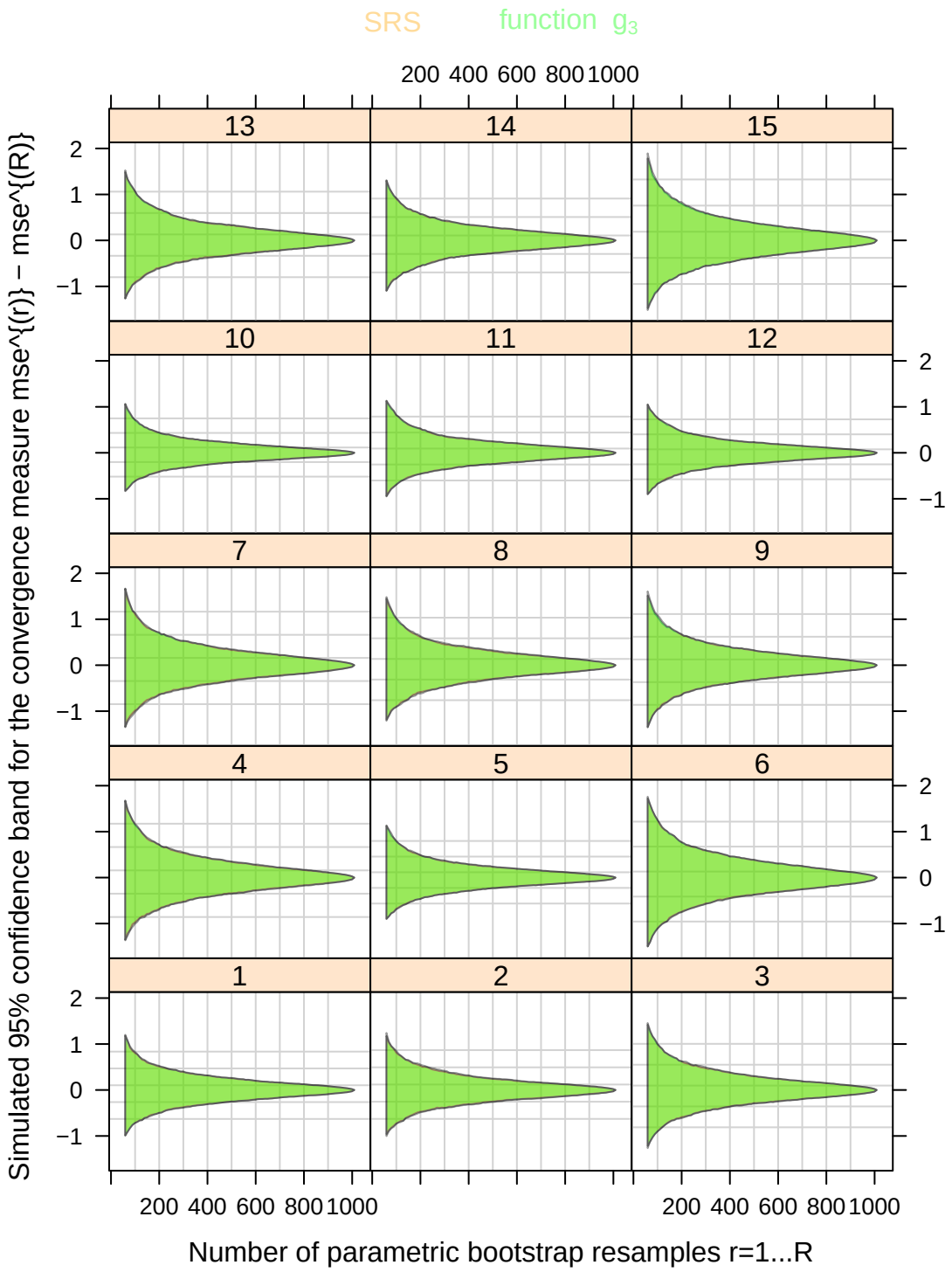

Figure 4.7: Using the control variate $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 1

Results for population 4

In population 4, the σ_u is high and the $\sigma_{e,d}$ are rather small. Thus, the control variates $g^{(2)}$ and $g^{(4)}$ are thought to perform best. In Figures 4.8 and 4.9 the variability of the estimates is plotted. As can be seen in both plots, the amount of variance reduction is massive. In some Areas it reaches over 90% less resamples needed for the same width of confidence bands. The direct comparison of the control variates $g^{(2)}$ and $g^{(4)}$ shows that $g^{(2)}$ performs slightly better then $g^{(4)}$, as was expected from the analytical approximation. The use of control variate $g^{(3)}$ does not have any impact on the variance of the estimate as can be seen in Figure B.23 in the appendix.

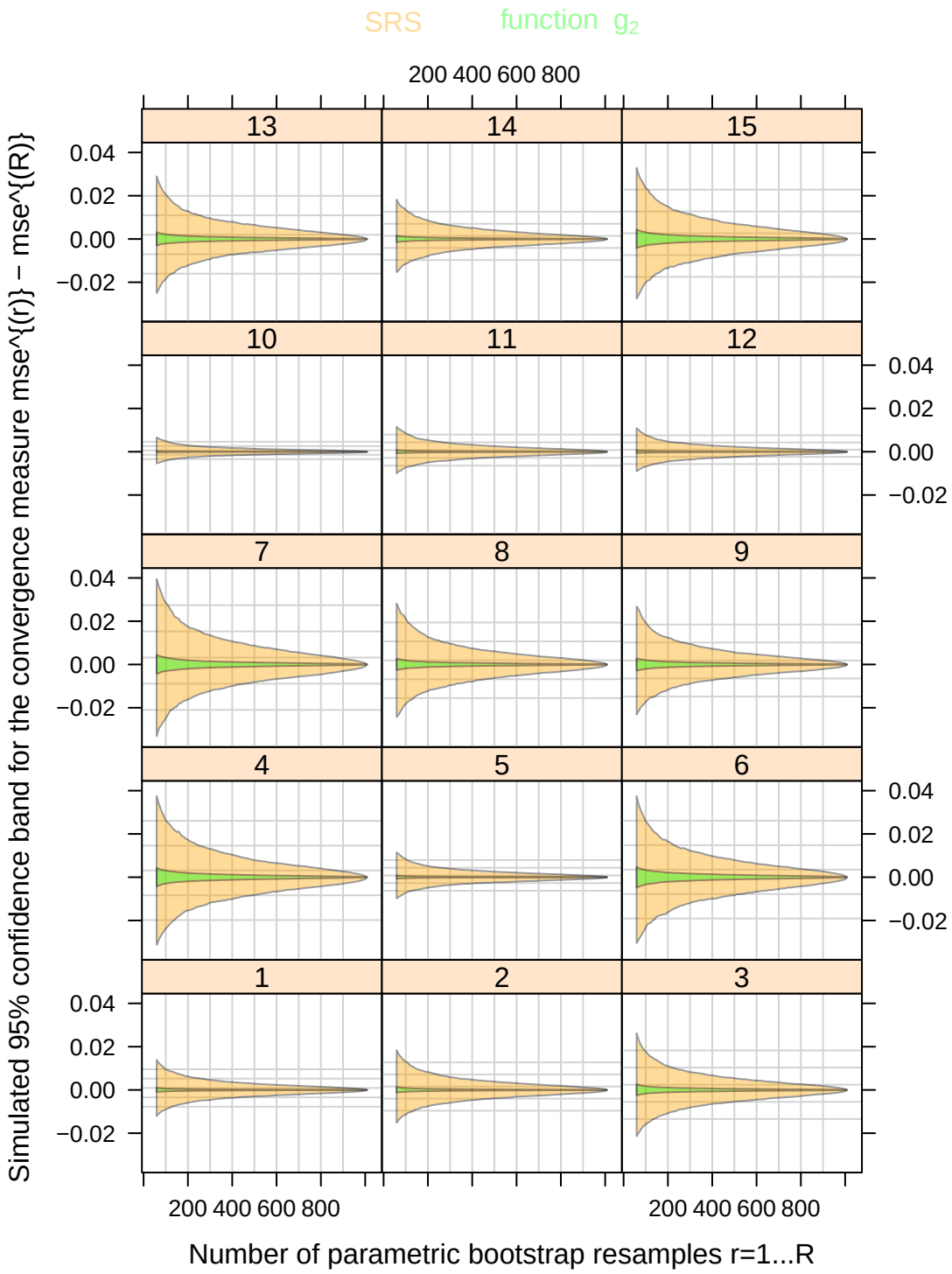

Figure 4.8: Using the control variate $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 4

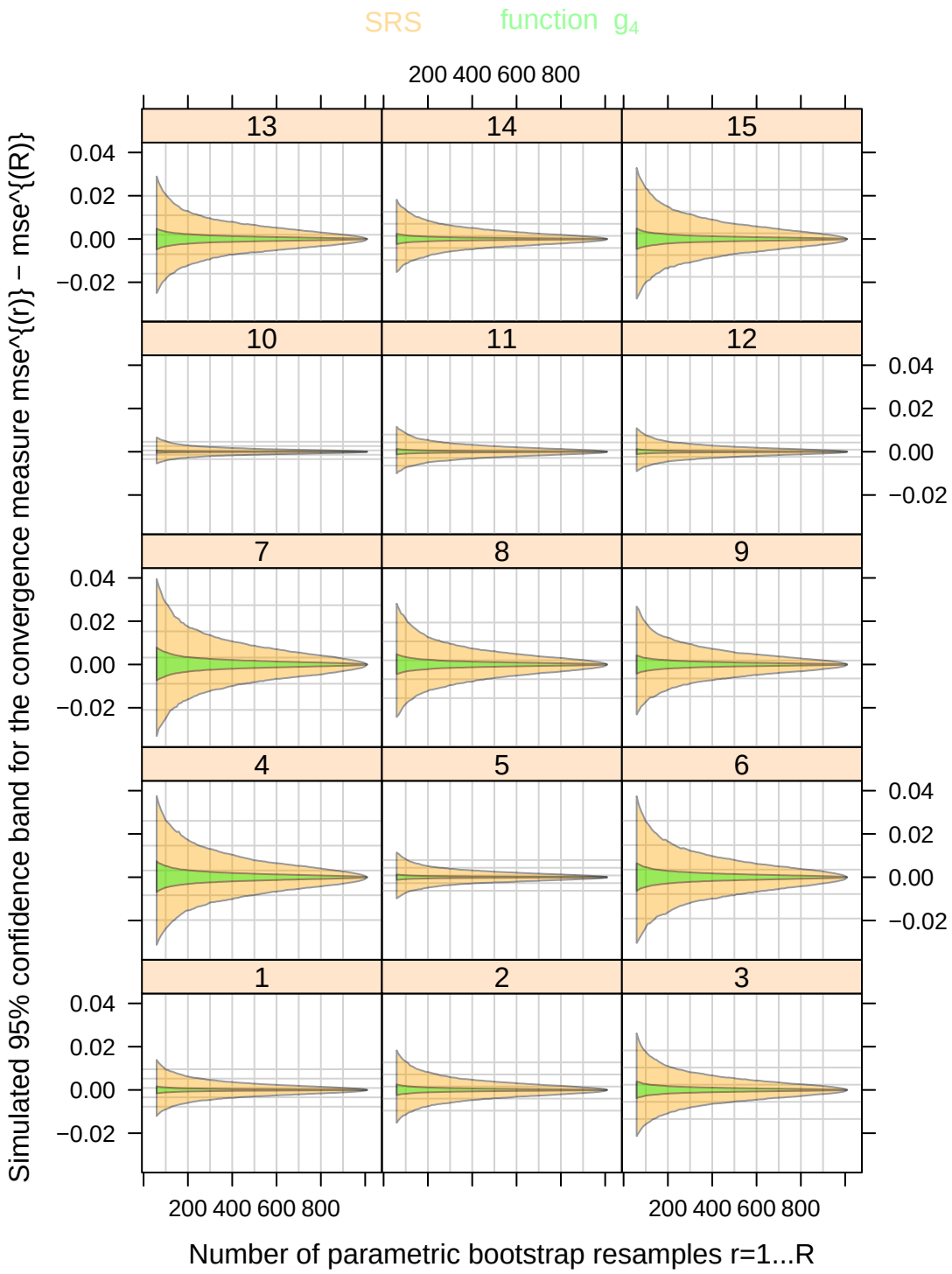

Figure 4.9: Using the control variate $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 4

Results for population 7

Population 7 is characterized by a small $sigma_u$ and larger $\sigma_{e,d}$. From the analytical approximation, in this setting the control variates $g^{(2)}$ and $g^{(3)}$ should be performing best. Therefore, at a first glance, surprisingly $g^{(4)}$ outperforms the other two control variates, even though the reduction of the variability of the parametric bootstrap MSE estimate is moderate in comparison to the other populations (compare Figures 4.10, 4.12, and 4.11). By using the control variate $g^{(3)}$, even an increase in variance may be observed. This, however, may be explained by the fact that in this population the REML variance estimate of σ_u in many samples is zero. By that, in many simulation runs the parametric bootstrap MSE estimate using the control variate $g^{(3)}$ is not computable. When comparing the performance of control variate $g^{(3)}$ in populations 7 and 9, which only differ in the number of areas, 15 and 100 respectively, the effect of the variance estimation becomes apparent. As can be seen in Figure 4.13 there is actually no longer any increase in variance of the estimate. In some areas, even a slight reduction is visible. For the constellation of population 7-9, the amount of possible variance reduction is not very high.

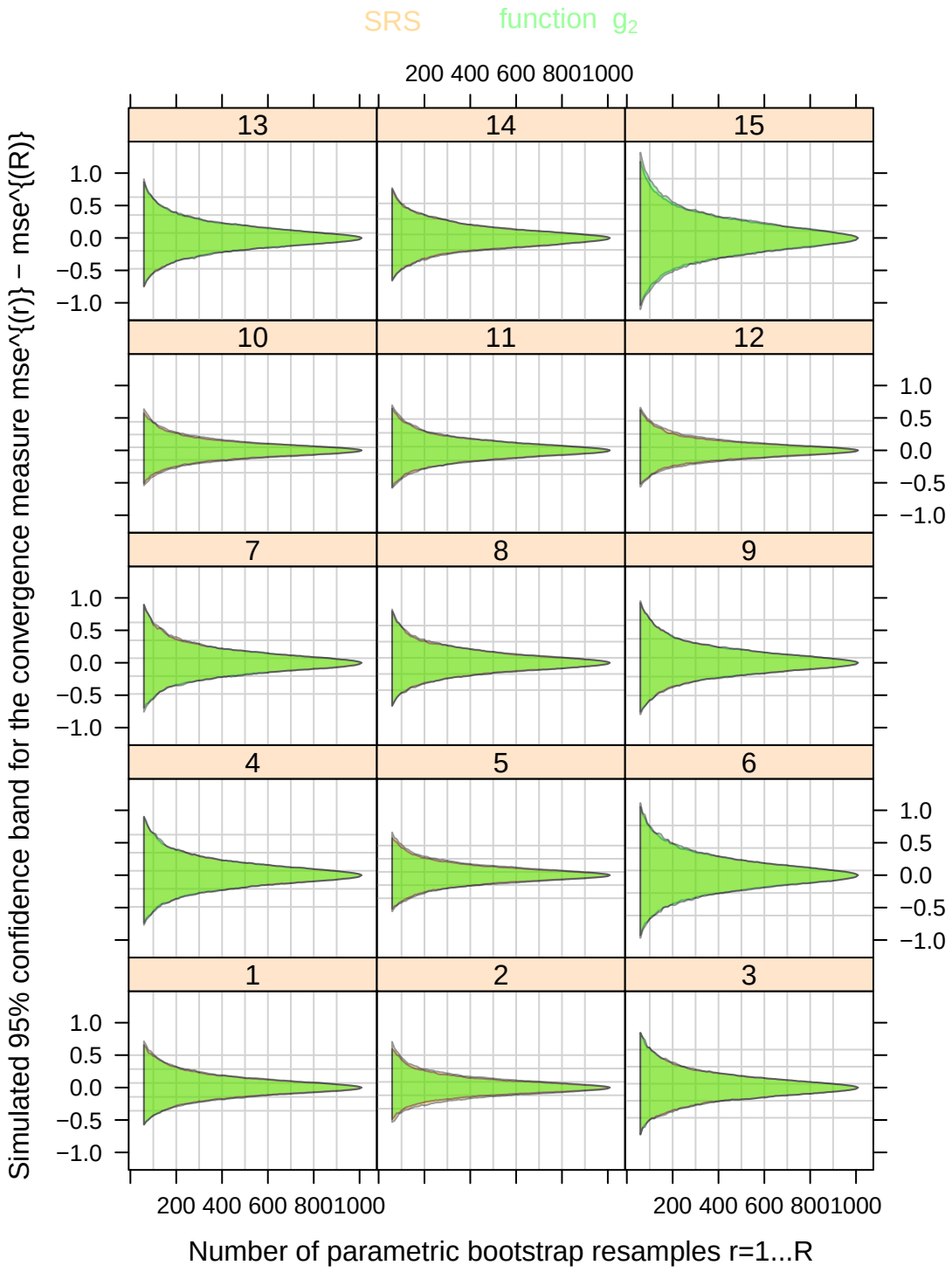

Figure 4.10: Using the control variate $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 7

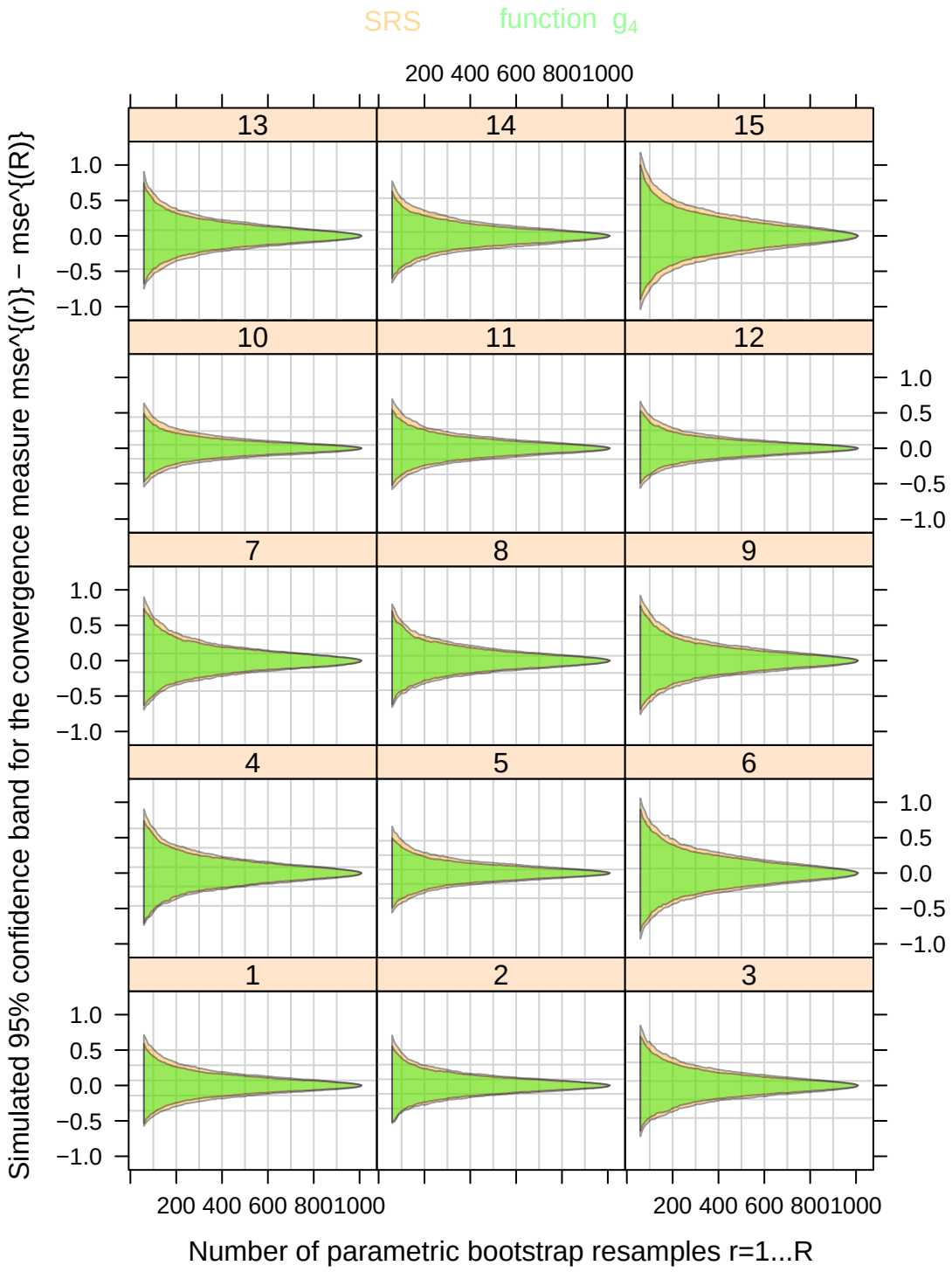

Figure 4.11: Using the control variate $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 7

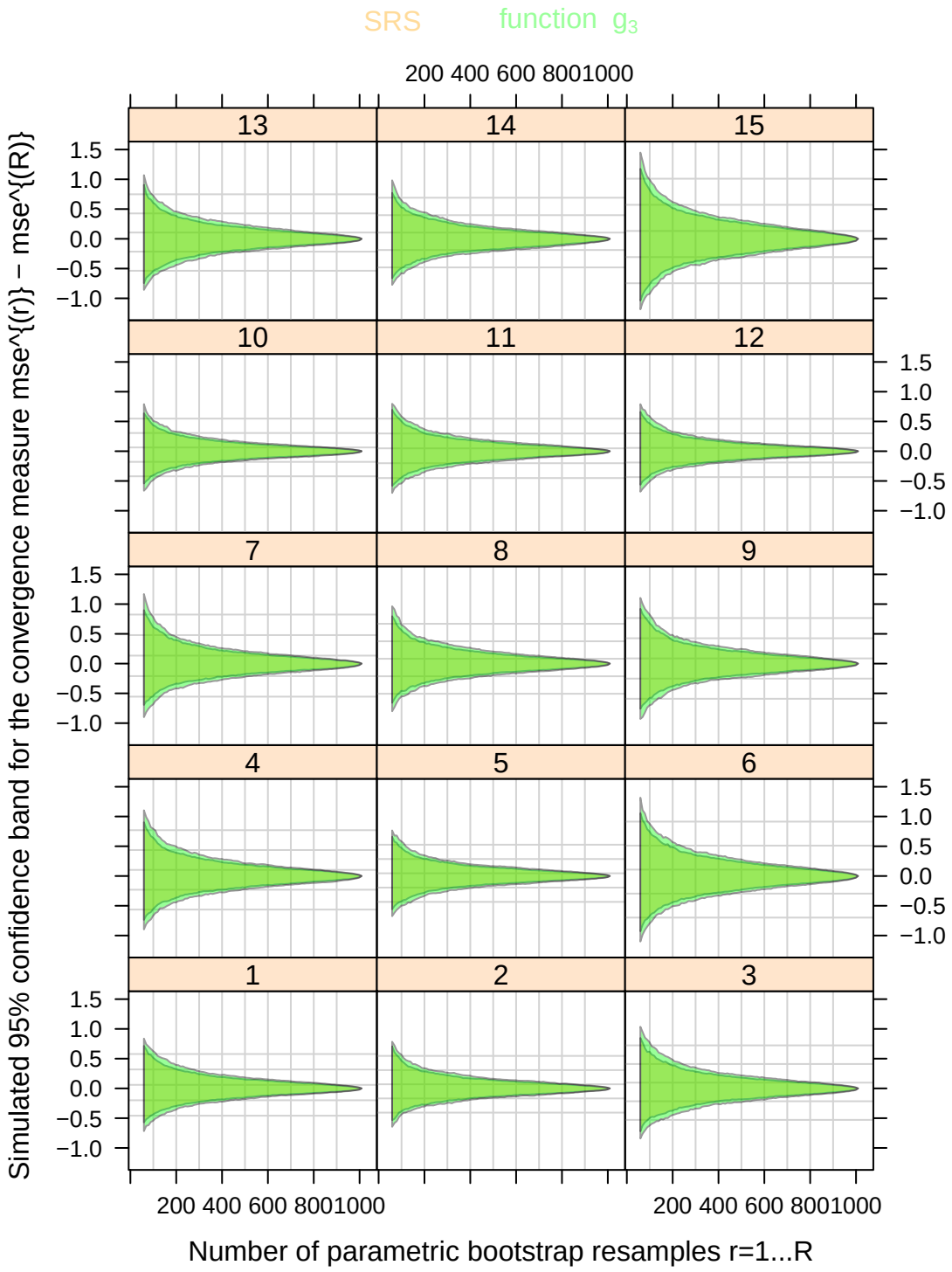

Figure 4.12: Using the control variate $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 7

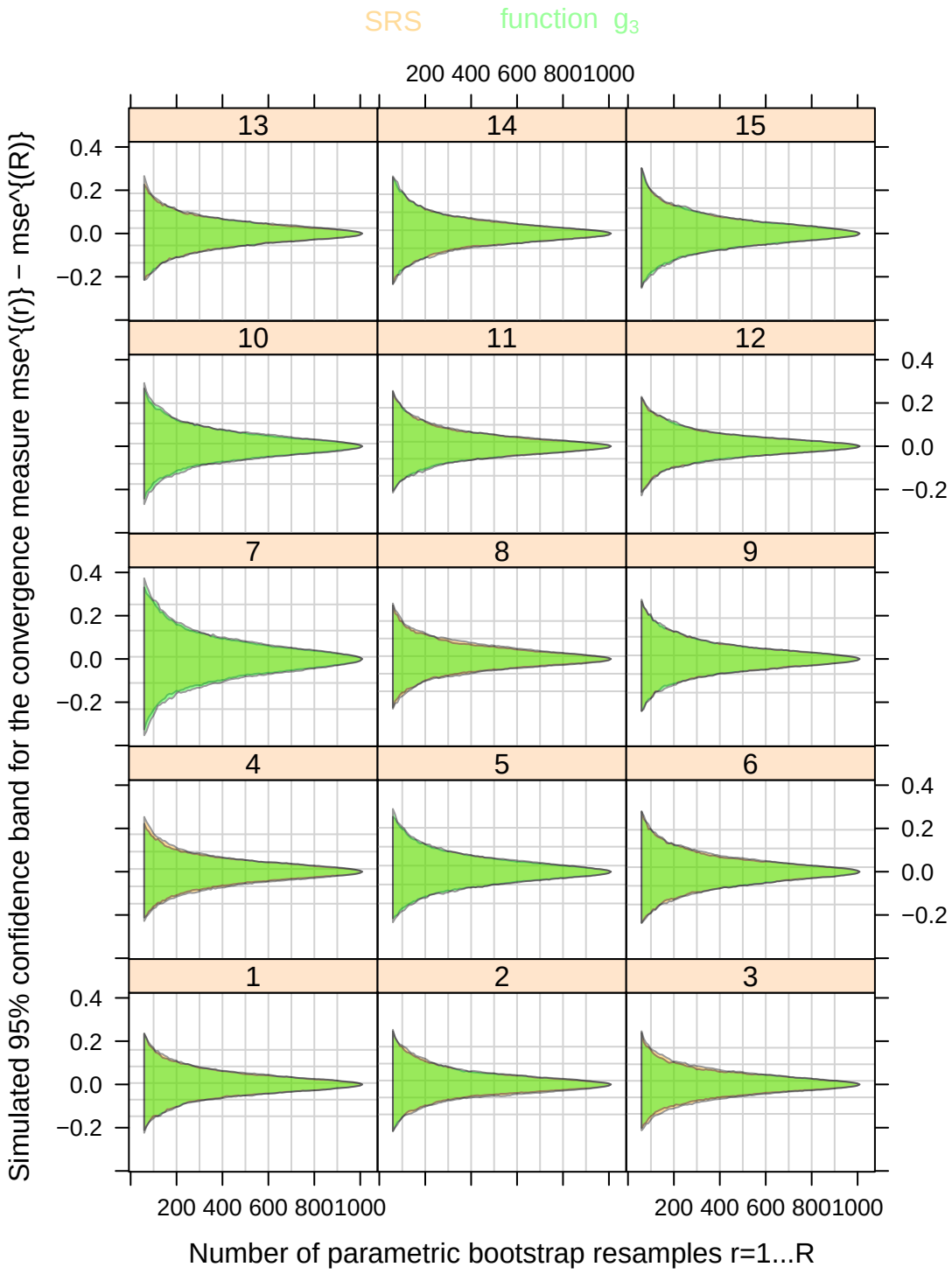

Figure 4.13: Using the control variate $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 9

Results for population 10
The results in population 10 again show a considerable variance reduction when using control variates $g^{(2)}$ and $g^{(4)}$. Also, in this population the control variate $g^{(2)}$ reduces slightly more the variance of the MSE estimate than the control variate $g^{(4)}$ (see Figures B.18, B.27, and B.36 in the appendix).

4.4 Conclusion

The need to reduce computational burden when using parametric bootstrap MSE estimates is apparent. Many small area estimators require a lot of computation time for computing a single estimate. For the parametric bootstrap applications, a large number of resamples are needed, which significantly increases computation time. Especially when practitioners are trying to evaluate certain models at an instant or time near the MSE estimate is an important part of this process. The Latin Hypercube Sampling did not show how to reduce the variability of the parametric bootstrap MSE estimate in any of the populations studied here. However, the use of control variates has been shown to be a computational easy implementable and reliable method. In some populations, the reduction of the needed resamples for a certain variability of the MSE estimate could be reduced by over 90%. This truly enables almost real-time computations of the parametric bootstrap MSE estimate. Only when σ_u^2 is very small, caution must be exercised with the variance estimation method. REML has shown to be problematic due to possible zero estimates. Therefore, in these cases, the use of generalized maximum likelihood methods as proposed by Lahiri and Li (2009b), adjusted density methods proposed by Morris and Tang (2011) or adjusted maximum likelihood methods proposed by H. Li and Lahiri (2010); Y. Li and Lahiri (2007), and Yoshimori and Lahiri (2012) should be used instead.

Chapter 5

Monte Carlo Simulations and Simulation Studies

Recently, Switzerland moved away from a classical census towards a register-based census - the Swiss Structural Survey. This yearly survey was introduced in 2011 and replaces the full census, which was drawn up every ten years. It is planned as a register assisted survey. For this purpose mainly the population register is used. In the long term, it may be possible to incorporate further registers. The population register is taken to be perfect in the sense that it has no register errors (for problems arising with register errors see Burgard, 2009 and Burgard & Münnich, 2012. Thus, counts and means of variables which are in the register are not estimated but counted. These variables are mainly demographic variables, such as *gender* and *age*. All other variables of interest, such as *employment rate*, *distribution of language* or *time spent traveling to the place of work* have to be estimated from the sample. The sample in the Swiss Structural Survey can be seen as a stratified random sample without replacement. It is drawn from the population register which serves as frame. The strata are the communities and the total sample size is allocated proportionally to the population size of each community. In total, 200 000 persons over 15 years are drawn from the roughly 8 million inhabitants of Switzerland (Bundesamt für Statistik Schweiz, 2013a, 2013b).

This change in methodology imposed questions concerning the quality of area estimates. It is of particular interest to have access to how small an area may be to still allow for precise estimates. Precise, in this context, is in the first place a matter of political argument. The question is what precision is needed in order to rely on the estimates. In the small area context, this decision becomes more complex, as many areas are estimated simultaneously. That is, there is a set of precision estimates which have to be evaluated. This leads to a decision theoretic problem which is discussed deeper in (Münnich & Burgard, 2012c). This thesis is partly attached to the project *Simulation der Strukturerhebung und Kleingebiete-Schäzungen*. The Final report of this project is available online (Münnich & Burgard, 2012c).

This chapter is organized as follows. First, a systematization of the different types of Monte Carlo simulation studies is proposed and the single simulation types are explained. Second, Section 5.2 discusses how to measure the precision of estimates within a Monte Carlo simulation study. Also, some notes on the decision theoretic problem in small area estimation are included. As these measures are difficult to revise with tables, due to the multidimensionality of the small area estimation setting, graphical summaries enable a quick and broad overview of the behavior of the different estimators at hand. Hence, third, useful graphs for this purpose are proposed and explained in Section 5.3. Fourth, the central question driving the project is tackled in Section 5.4. Namely, how to handle very small areas in the context of the Swiss Structural Survey. Last but not least, fifth, the possibility of incorporating covariates into the estimators, which are only available on a certain degree of aggregation, is tackled. This is done exemplary by using an additional variable from another register, which may only be provided due to legal reasons on a cross tabulation of some demographic variables for each area.

5.1 Model-Based versus Design-Based Monte Carlo Simulations

Monte Carlo simulations, with the rise of high computer power, became an increasingly feasible way for obtaining diverse measures. In the context of survey sampling, the two main directions of Monte Carlo simulation are the design-based and the model-based Monte Carlo simulations. They are generally conducted in order to obtain knowledge regarding the impact of sampling design on the precision of estimates and to evaluate the appropriateness of estimation methods on certain populations. As there are many different sampling designs, estimators, statistical properties of interest and populations, the Monte Carlo simulations have to be designed separately for each purpose. In Table 5.1 a systematization is proposed, which focuses mainly on the aspects concerning which randomization is chosen to obtain the Monte Carlo approximation and which kind of population is used. The different types of Monte Carlo simulation are illustrated by an example where an approximation to the bias of an estimator is studied. The Monte Carlo simulation would consist in evaluating the integral

$$\text{Bias} = \int_{-\infty}^{+\infty} \left(\widehat{\Psi}(x) - \Psi \right) f(x) dx \quad . \tag{5.1}$$

The randomization is defined by the probability density function $f(x)$, the estimator under consideration produces estimates $\widehat{\Psi}(x)$ and the true value is Ψ. As estimators, the HT (see Section 3.2.1) and the GREG (see Section 3.2.2) will be used in this example. Of course, from the theory it is known that the HT estimator is unbiased and the GREG is at least asymptotically unbiased.

Broadly speaking, the Monte Carlo simulations may be divided into two main groups, the design-based and the model-based simulations. The main difference between the two is that the model-based simulations are characterized by the fact that the randomization is model-based and the randomization of the design-based simulations is design-based. First, the simulation types with model-based flavour are discussed, and then the design-based flavoured.

In the case of a *pure model-based* simulation, this $f(x)$ represents the joint distribution of all variables used in the estimation of $\widehat{\Psi}$. In the case of the HT, only the variable of interest y is to be drawn from the population and a corresponding weight is to be derived. In case of simple random sampling from the model population, all units have equal weights. The model population in this case is simply a distribution like the normal or Bernoulli distribution. In contrast, when the estimator needs more variables, like the GREG, then the covariates x also have to be drawn from the population. In that case, the model population is distributed according to the assumed joint probability density function $f(x)$, e.g., a multivariate normal distribution. Again, if simple random sampling from the model population is performed, then all weights are taken to be equal.

Often, however, the covariates are taken to be fixed and only the variable of interest y is taken to be random. In this case, the $f(x)$ is a probability density function depending on the fixed covariates. This is easier to implement, as it may be extremely difficult to draw samples from the joint distribution of all variables. Generally, this procedure is called *model-based* simulation. At this point, in the context of small area estimation, it is important to note that the handling of the area effect is important. The mixed models underlying most of the small area estimates generally assume a distribution on the area effects. That is, under the mixed model, the area effect is random. For the assessment of the estimation of the fixed effects $\widehat{\beta}$, this randomness has to be considered as such. In small area estimation, however, the area effect is clearly a fixed value. E.g., one area actually has a higher average income than another area. Therefore, it might be sensible to take the area effect as fixed in $f(x)$.

A less common approach is to combine the randomization from the design with the randomization from the model population. That is, $f(x)$ is the joint distribution of the model population and the sampling design. This case would then be called a *model- and design-based* simulation. To clarify this approach an example is constructed where simple random sampling is compared to stratified random sampling and cluster sampling. For ease, the strata, the cluster and the sample sizes are chosen such that the drawn units have the same inclusion probability. This can be changed in order to have varying design weights. The assumed superpopulation is the following normal distribution.

$$(y,x) \sim \mathbf{MVN}\left((0,0), \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right) \quad . \tag{5.2}$$

In analogy to Chapter 4.2.1, the simple random sampling is the design where realisations are drawn without restrictions from this multivariate normal distribution. The sample

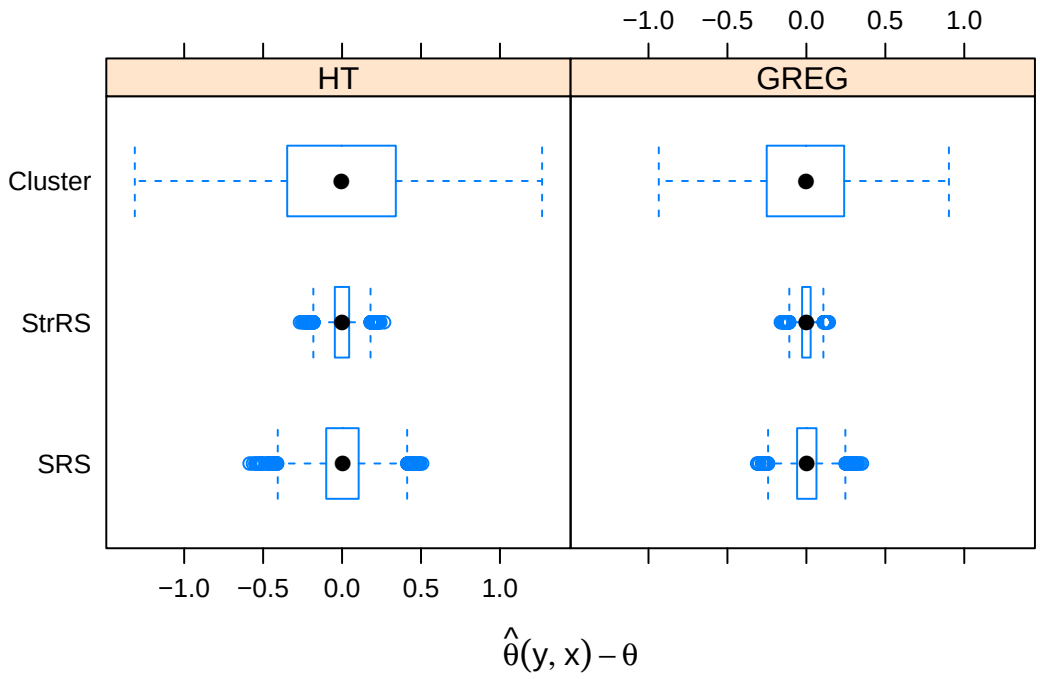

Figure 5.1: Example model-based and design-based Monte Carlo simulation

size is set to 45. For stratified random sampling, in contrast, the sample is drawn within strata. For this example a 3×3 raster with equal probable strata is used, wherein each stratum 5 units are drawn. This is accomplished by using the inverse transform sampling method (see Algorithm A.1) and the Cholesky decomposition (see Algorithm A.2). In the case of cluster sampling, 15 units are drawn by simple random sampling within 3 randomly chosen strata (clusters) of the 9 previously defined strata. Therefore, all three designs have the same sample size of $n = 45$ and all units have equal probability to be included in the sample. The estimators under consideration are the HT and the GREG. The results of this simulation with 10,000 repetitions is visualized in Figure 5.1. One boxplot shows the distribution of the 10,000 iteratively produced point estimates of the mean in the previously defined model population under one sampling design. One can see that, as expected, neither estimator shows a bias in the simulation. As is also expected, the stratified random sampling is more efficient than the simple random sampling. And, as expected, cluster sampling is less efficient than simple random sampling. Overall, the GREG is more efficient than the HT, which is plausible, as the theoretical correlation is $\mathrm{COR}\,[y, x] = 0.8$.

As already introduced in the *design-based* simulations, the randomization is done only via the sampling design. That is, to be exact, the design based simulation needs the real data, in order to be able to solely detect effects due to the sampling design in the given population. However, in most applications Monte Carlo simulations are not needed if the true population is already available. Commonly, only an approximation to the true

Name	Randomization	Population	True Values
pure model-based	model	all variables are drawn from the model population for each replication	theoretical
model-based	model	dependent variables are drawn from the model population for each replication, everything else is fixed	theoretical
model- and design-based	model and design	samples are drawn according to a sampling design from the model population	theoretical
smooth design-based	design	dependent variables or all variables are drawn once from the model population for all replications	empirical
realistic design-based	design	dataset is realistic / contains the relevant characteristics of the real underlying data	empirical
design-based	design	dataset is the real data	empirical

Table 5.1: Comparison of model based and design based Monte Carlo simulation settings

population is available. This could be an older census, or some other data set containing structures and variables which are relevant for the population of interest. As this is not the true real population, but rather a realistic one, this is called *realistic design-based* Monte Carlo simulation.

It is often the case that some covariates are known to a certain degree, e.g. from registers or other surveys, but only a little information about the dependent variable is available. In this case, usually a *smooth design-based* Monte Carlo simulation is performed. As in model-based simulation, the covariates are taken as fixed, and the variable y is a realisation of a random variable given the fixed x. In contrast to model-based simulation, only one realisation of this random variable is considered. With this data set, the design-based simulation is performed.

5.2 Performance Measures

The output of a large Monte Carlo study is usually very huge. Therefore, in order to be able to compare the estimators under different scenarios, the information has to be

reduced to an easy to handle number of figures. Usually, two main features of estimators are considered, on the one hand, the variability of an estimator is of interest and, on the other hand, its bias is taken into consideration. In the following two sections, first the measures for the point estimates and then those for the precision estimates are discussed.

5.2.1 Performance Measures for the Point Estimates

For the calculation of performance measures of the point estimate $\widehat{\Psi}_d^*$ for estimator $*$ in area d, its distribution from the Monte Carlo study is considered. The variability of an estimate can be accessed by the relative dispersion. It is defined as

$$\text{RDISP}_d^* := \text{RDISP}(\widehat{\Psi}_d^*) := \frac{Q(\widehat{\Psi}_d^*, 0.95) - Q(\widehat{\Psi}_d^*, 0.05)}{\Psi} \quad , \tag{5.3}$$

where Ψ is the true value known from the simulation environment, but not in the actual survey. The possible values of the RDISP lie in $(0, \infty)$. The lower they are, the lower the variability of the estimate. The dispersion is a kind of *robust* way of measuring the variability, as the outlying 10% of the estimates are cut off.

The bias can be assessed as the mean difference of the estimates from the true value in relation to the true value. The formula is

$$\text{RBIAS}_d^* := \text{RBIAS}(\widehat{\Psi}_d^*) := \frac{\frac{1}{R}\sum_{r=1}^{R} \widehat{\Psi}_{d\,r}^* - \Psi_d}{\Psi_d} \quad . \tag{5.4}$$

Its values lie in $(-\infty, \infty)$. The nearer the RBIAS is to zero, the lower the bias of the estimator. In survey statistics, there is often a trade-off between the variability of an estimate and its bias (see Chapter 3). That is why it is important to also look at compensatory measures which involve both aspects.

A popular choice for this is the relative root mean squared error. It is computed as the root of the mean squared deviation from the estimates to the true value in relation to the true value. The equation is

$$\text{RRMSE}_d^* := \text{RRMSE}(\widehat{\Psi}_d^*) := \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(\widehat{\Psi}_{d\,r}^* - \Psi_d\right)^2}}{\Psi_d} \quad . \tag{5.5}$$

Often one estimator is better for one area, whilst another estimator is better for a different area. In this case, one has to assess which estimator is best overall. One approach is to

compare the average of the measure over all areas with this average measure result.

$$\text{AVRDISP}^* := \frac{\sum_{d=1}^{D} \text{RDISP}_d^*}{D} \tag{5.6}$$

$$\text{AVRBIAS}^* := \frac{\sum_{d=1}^{D} \text{RBIAS}_d^*}{D} \tag{5.7}$$

$$\text{AVRRMSE}^* := \frac{\sum_{d=1}^{D} \text{RRMSE}_d^*}{D} \tag{5.8}$$

If the interest lies in a minimum requirement in all areas, then one can look at the worst area only. In all three cases, this would be the maximum absolute value:

$$\text{MARDISP}^* := \max_{d=1..D} (\text{RDISP}_d^*) \tag{5.9}$$

$$\text{MARBIAS}^* := \max_{d=1..D} (\text{RBIAS}_d^*) \tag{5.10}$$

$$\text{MARRMSE}^* := \max_{d=1..D} (\text{RRMSE}_d^*) \tag{5.11}$$

A more general Quantile of the distribution of the measures can be considered, where the Median $q = 0.5$ is often used

$$\text{QRDISP}_q^* := Q(\text{RDISP}_d^*, q) \tag{5.12}$$

$$\text{QRBIAS}_q^* := Q(\text{RBIAS}^*, q) \tag{5.13}$$

$$\text{QRRMSE}_q^* := Q(\text{RRMSE}^*, q) \tag{5.14}$$

For the practitioner, in the end, it is of interest whether the point estimate lies within a certain range around the true value. This measure will be called the near-enough rate (NERATE), it ranges from zero to one and is defined as:

$$\text{NERATE}_\Delta^* := \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}_{|\Psi - \widehat{\Psi}_{dr}^*| < \Delta} \quad . \tag{5.15}$$

5.2.2 Performance Measures for the MSE and Variance Estimates of the Point Estimates

In order to have an accuracy measure for the point estimate in the one-sample-case, MSE or variance estimates for the point estimates are computed. These can be evaluated either by their bias or by confidence interval rates.

The relative bias of the MSE estimation is defined in analogy to that of the point estimates as

$$\text{RBIASMSE}^*_d := \frac{\frac{1}{R}\sum_{r=1}^{R} \widehat{\text{MSE}}(\widehat{\Psi}^*_{dr}) - \text{MSE}(\widehat{\Psi}^*_d)}{\text{MSE}(\widehat{\Psi}^*_d)} \tag{5.16}$$

$$\text{with} \qquad \text{MSE}(\widehat{\Psi}^*_d) := \frac{1}{R}\sum_{r=1}^{R}(\widehat{\Psi}^*_{dr} - \Psi_d)^2$$

The values of the *RBIASMSE* lie in $[-1, \infty)$, where values near zero denote an accurate MSE estimator, a value below zero depicts underestimation and a value over zero, overestimation of the MSE. An MSE estimator that generally overestimates the MSE is called conservative.

It is not only important to have a low bias of the MSE estimator, but also it is of interest to have low MSE estimates. Thus, the combination of the magnitude of the MSE estimate and the accuracy of the MSE estimate has to be measured at the same time. A measure that fulfils these requirements is the *confidence interval coverage rate* (CICR). The CICR is defined as follows:

$$\text{CICR}^*_d := \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}_{\Psi_d \in \text{CI}(\widehat{\Psi}^*_{dr})} \quad , \tag{5.17}$$

where $\text{CI}(\widehat{\Psi}^*_{dr})$ is the estimated confidence interval for the estimator $*$ in area d in simulation run r. For the methods used to estimate confidence intervals see Section 3.3.6.

With an arbitrary high MSE estimate, the CIRC will always be 1. Therefore, it is of great interest to compare the CIRC to the magnitude of the MSE. This can be done by incorporating the mean confidence interval length (MCIL) over the simulation runs. This is defined as

$$\text{MCIL}^*_d := \frac{1}{R}\sum_{r=1}^{R} \text{CIL}(\widehat{\Psi}^*_{dr}) \quad . \tag{5.18}$$

5.3 Visualization and Interpretation of Monte-Carlo Simulation Results

Some of these measures are area specific, while others go over all the areas at once. In the setting of the Swiss Structural Survey, almost 3000 areas are available. Tabulating these large amounts of measures obviously will not provide a quick overview about the performance of certain estimators. Therefore, graphical tools are used to summarize this large amount of information. For ease of interpretation, mainly a standard graphical tool is used, the *Box-and-Whisker Plot* (boxplot) first proposed by Tukey (1977, § 2C).

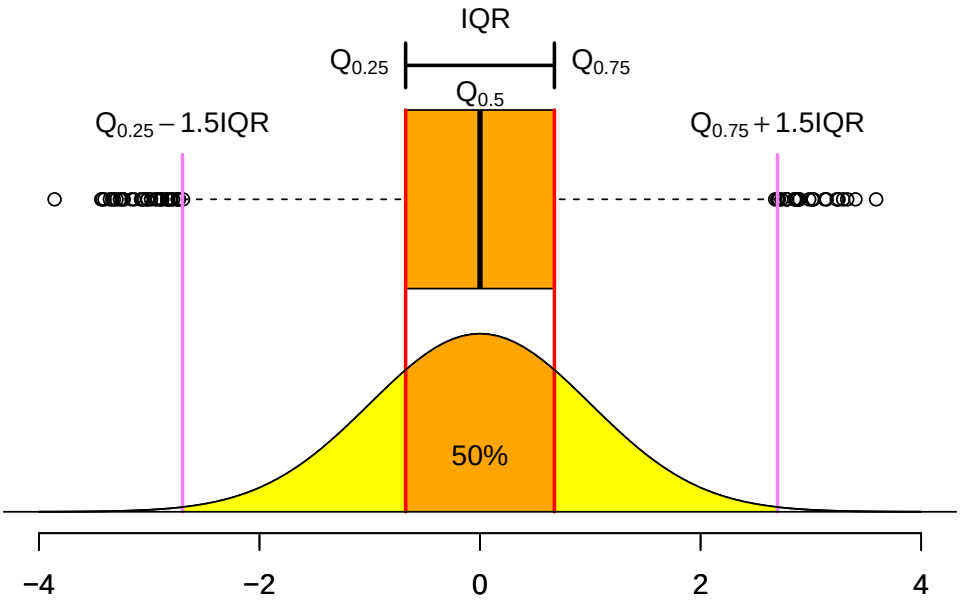

Figure 5.2: Example of a Box-and-Whisker plot for the standard normal distribution

This graph is very useful for getting an idea of the distribution of the data. Its construction is rather simple, as can be explained with Figure 5.2. The box of the boxplot delimits the central 50% of the observed values. Therefore, the left and right borders of this box are the 25% and 75% quantiles respectively. The smaller the inter quantile range (IQR), the more compact most of the observations lie. The Whiskers are positioned following Tukey (1977, § 2C) at the two observations which are farthest away from the median, but not farther than the maximal distance of $1.5 \times \text{IQR}$ from the borders of the box. In addition to the boxplots, the aggregate measures average, maximum absolute value and quantiles found in equations (5.6), (5.9) and (5.12), respectively, are depicted within the plot as well. Besides the median, which is already part of the boxplot, a box showing the central 80% of the values is drawn behind the boxplots. The points lying outside the whiskers can be seen as outliers.

The second graph used is a so-called *waterfalls* graph. This is used to visualize the CICR. The name derives from the fact that in many cases in small area estimation the resulting picture recalls a waterfall. As can be seen in Figure 5.3, on the y-axis the CICR is plotted against the MCIL. The shorter the MCIL, the more precise is the MSE estimate. Therefore, it is desirable if the points are far to the left in the waterfalls graph. On the other hand, the CIRC should have the height of the nominal CICR, that is, for a 95% confidence interval the CICR should be 95%. If the CICR is higher, then the CI is measured somewhat more conservatively. If the CI is lower than this nominal rate, then the MSE underestimates the true uncertainty about the estimate at hand. In this case, a short MCIL purports a precision of the estimate, which cannot be held. To facilitate seeing whether the CICR is at the nominal rate, a horizontal red line is drawn at its value.

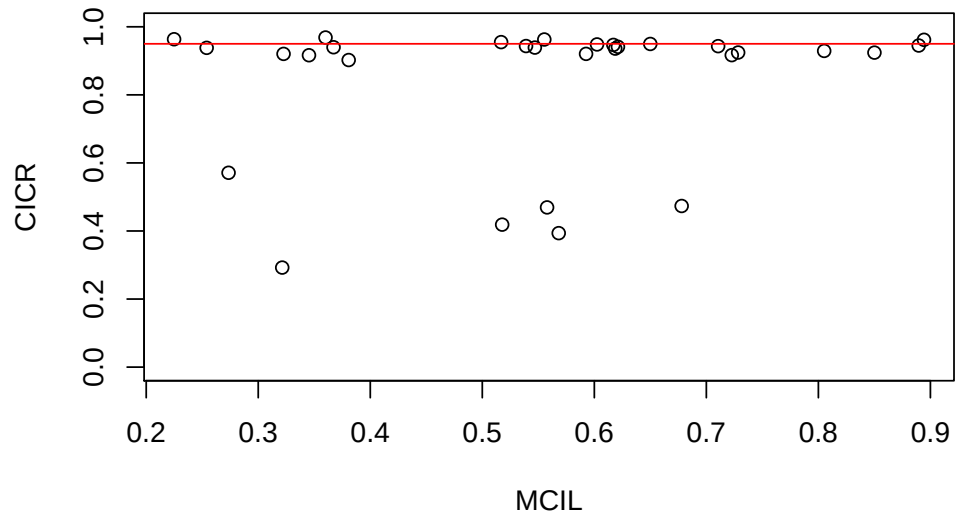

Figure 5.3: Example of a waterfall graph for a nominal confidence interval coverage rate of 95%

5.4 Area Sizes and Omission of *too* Small Areas

A key issue is how to handle very small areas. For many variables of interest, the units in very small areas may behave in a different way than the rest of the population. If many of these very small areas exist, they might have a stronger impact on the model than is appropriate for the share of units within these areas. As traditionally no estimates are published in official statistics for these very small areas, one viable approach would be to omit these areas from the estimation and prediction.

Another possibility is to agglomerate the areas to agglomerations that have a minimum size. In this approach, the estimation and prediction is performed on aggregate level. From the viewpoint of official statistics, these agglomerations must be a sensible political unit. This may lead to the problem that not all agglomerations will actually reach the desired minimal size. However, most agglomerations should meet the required size.

Last but not least, a combination of these two approaches may be viable. That is, the agglomerations are built from areas with a minimal size themselves.

To answer the question of which approach is most feasible, a Monte Carlo simulation is performed. In the next Section, the setting of the Monte Carlo simulation is depicted. Next, the results of the Monte Carlo simulation are summarized. First, the performance for the point estimates is tackled and subsequently the precision estimates are compared.

5.4.1 Setting of the Monte Carlo Simulation

As mentioned in the introduction, one approach is to use agglomerations as areas for the estimation and prediction. The agglomerations are named and built as follows

ZGDEN: Estimation and prediction on all communities.

ZGDEM: Estimation and prediction on all communities with at least 1000 inhabitants. Communities that have less than 1000 inhabitants are assigned to geographical near communities, if politically sensible.

ZGDEMM: Estimation on and prediction of all communities with at least 2000 inhabitants. Communities that have less than 1000 inhabitants are assigned to nearby geographical communities, if politically sensible.

The second possibility presented is to use only the communities of a minimal size. These scenarios are as follows:

CUT0000: All communities are considered.

CUT1000: Only communities with at least 1000 inhabitants are incorporated.

CUT2000: Only communities with at least 2000 inhabitants are incorporated.

The Monte Carlo simulations comprise all combinations of these two scenarios. In Table 5.2 some summary statistics are given for the combinations of the scenarios. Besides the number of areas in each scenario, some quantiles of the area sizes are also presented in the table. As can be seen also in the case of ZGDEMM, there are still areas of a size of less than 2,000 inhabitants. This is due to the fact that some areas could not be allocated in a politically sensible way to other areas. Further, the agglomeration and the cut off were calculated including inhabitants younger than 15 years old. In contrast, in the table, the area sizes are calculated on the basis of inhabitants older than 15 years, which is the relevant frame for this simulation.

As can be seen in this summary, the range of area sizes is very wide. Beginning with an area of size 17 up to an area of a size of 320,324 inhabitants aged over 15 years. This wide range is typical of small area estimation applications in official statistics. However, often the Monte Carlo simulations performed to argue for Small Area estimators rely on less variable area sizes.

The sampling design is almost equivalent to a stratified random sample with proportional allocation to the areas sizes (cf. Section 3.1). Within the areas, a simple random sampling is drawn. Therefore, the sampling weights do not vary much. In every area at least 2 units are drawn and, overall, there are 200,000 units in the sample. For the Monte Carlo simulation, 1,000 samples following this sampling design were drawn. For each of these samples, all the estimators were estimated repeatedly.

Table 5.2: Area Sizes for the scenarios of the Monte Carlo simulation

Agglo-meration	Cut off point	Number of areas	Quantile of the size of the areas min	5%	50%	95%	max
ZGDEN	CUT0000	2,896	17	79	697	7,227	320,324
	CUT1000	1,322	765	890	205	12,065	320,324
	CUT2000	805	1,505	1,696	3,236	14,617	320,324
ZGDEM	CUT0000	1,744	318	793	1,590	10,172	320,324
	CUT1000	1,319	765	890	2,013	12,098	320,324
	CUT2000	805	1,505	1,696	3,236	14,534	320,324
ZGDEMM	CUT0000	1,248	318	1551	1,693	12,556	320,324
	CUT1000	1,131	769	979	2,616	13,314	320,324
	CUT2000	803	1,505	1,695	3,243	14,634	320,324

5.4.2 Results of the Monte Carlo Simulation

In this section, the results of the Monte Carlo simulation are discussed. Hereby, the results for the agglomeration ZGDEM are being suppressed for simplicity, as it turns out that the findings are in between the agglomeration ZGDEN and ZGDEMM. The same applies to the cut off CUT1000, where the results indicate recommendations in between CUT0000 and CUT2000. First the point estimates and then the precision estimates are compared in the following.

5.4.2.1 Evaluation of the Point Estimates

In Figure 5.4 the boxplots of the area-specific RRMSE for the single estimators and scenarios are illustrated. The vertical lines serve as orientation and simplify the comparison between the boxplots of the estimators and the scenarios.

As expected, the GREG outperforms the HT in all scenarios. This is due to the assisting model (see section 3.2.2) used by the GREG. The same covariates are also used by the other estimators but, as is evidenced, with different results. For most areas the model based estimators perform better in terms of RRMSE than the GREG, yet not for all areas. For example, the comparison of the GREG with the YOURAO shows that the right whisker of the boxplot for the YOURAO is near the left margin of the box for the GREG. In other words, with the YOURAO, almost all areas have a lower RRMSE than the first quartile of the GREG. In turn, with the YOURAO, on the average for the simulation, up to five areas lie above the maximal RRMSE of the GREG.

The comparison of the YOURAO with the BHF demonstrates that the YOURAO performs slightly better than the BHF. The BHFW produces RRMSE almost identical to the BHF, as can be seen in Figure 5.5.

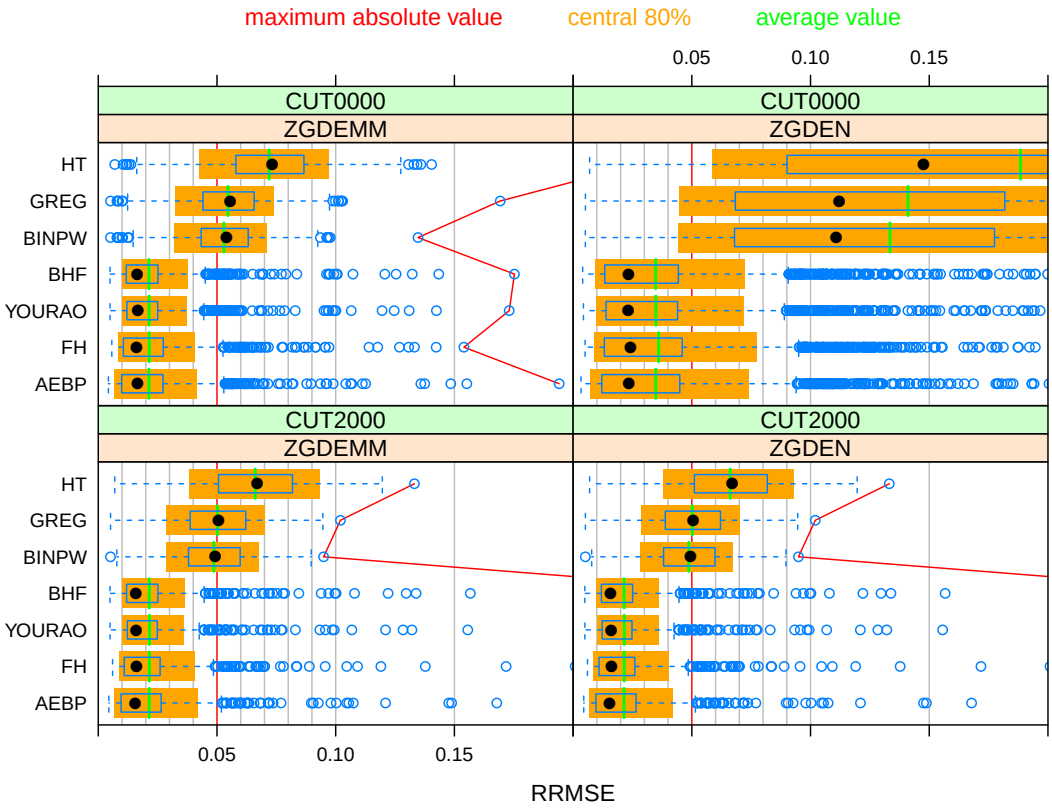

Figure 5.4: RRMSE of the point estimates under the four scenarios.

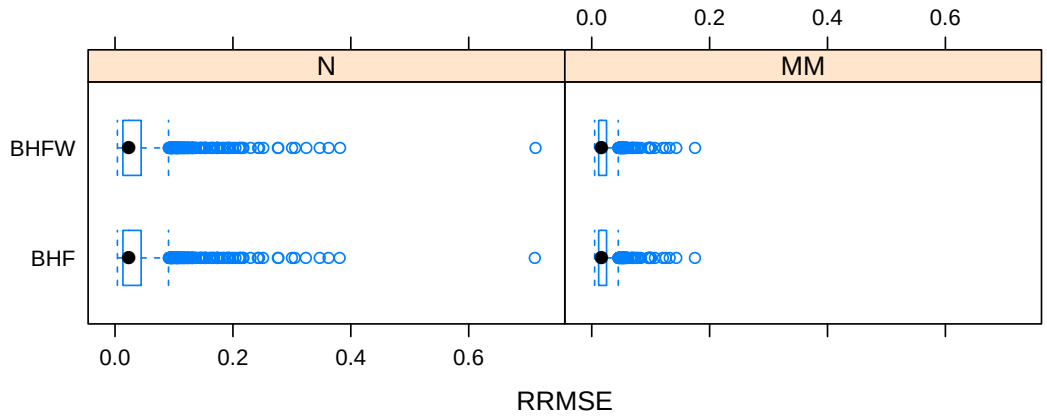

Figure 5.5: RRMSE of BHFW versus BHF

In comparison to the before mentioned estimators, the BINPW is a binomial unit level estimator. In the direct comparison between the GREG and the BINPW, it is striking that the boxes are very similar to each other, yet the BINPW is always slightly better than the GREG. But by far it does not reach the precision of the other small area estimators. The big point for the BINPW is that there are no single areas with an extremely high RRMSE. The BINPW is a composite estimator and its shrinkage-factors in this study mostly accent the design-based component. Hence, it shows similar characteristics as the design-based estimators in this simulation study.

As to the area level estimators, the high reduction of RRMSE with respect to the design based methods is depicted. Even though they only use information on an area level, they outperform the GREG by far. In comparison to the unit level model based estimators, the area level estimators perform only slightly worse. The AEBP, the binomial model for area-level data, again achieves results similar to the FH.

In Table 5.3, summary statistics for the NERATE are presented. The interval into which the estimate should fall is defined by $(\theta - \sqrt{N\theta}, \theta + \sqrt{N\theta})$, where θ is the active population rate to be estimated. The results are controversial. For the design based methods, the NERATE is very low with maximal NERATE over all areas of 0.38 up to 0.82, depending on the scenario. Even worse, the mean NERATE over all areas does not go beyond 33%. That is, two-thirds of the design based estimates in this simulation lie outside the interval around the true value. The BINPW yields only slightly higher rates. In contrast, the FH, YOURAO, BHF and AEPB yield much higher maximum and mean NERATEs. The mean NERATES are almost twice as high as the ones from the design based estimates. The controversy becomes apparent when focusing on the lower tail of the distribution of the NERATEs. There, the design based estimates and the BINPW hold a minimal NERATE, even though it is quite low. The small area estimators fall to a NERATE of zero. Comparing the small area estimates, one can see that the area level estimates have a higher NERATE for the 3rd quartile and a lower one for the 1st Quartile.

With these figures, the decision theoretic problem becomes apparent. Small area estimates provide higher NERATEs for most areas, thus lying within a certain, predefined and targeted range around the true value. On the other hand, independently of which sample is drawn, there is no chance for some areas to obtain an estimate which can be defined as near enough to the true value. Even though in the case of the design based estimators and the BINPW, the NERATE is very low for all areas, the areas have a real chance to obtain an estimate defined as near enough.

In Figure 5.6, the RBIAS is illustrated in the form of boxplots. The vertical red line marks the desirable value zero of the RBIAS. Here, the HT and the GREG show the smallest bias. Single higher differences from the zero in the small areas only arise when the really small areas are also used (ZGDEN CUT0000). The BINPW also performs very well when considering the RBIAS. However, all other small area estimators show high differences from zero for single areas. The more small-scaled the estimation, the higher the RBIAS results. In this context, it should be especially emphasized that the AEBP in the median over the areas shows a smaller RBIAS than the YOURAO and the BHF.

Table 5.3: Table of near enough rates for selected estimates

Aggregation	CUT	Summary	Estimator							
			HT	GREG	FH	BHF	YOURAO	BINPW	AREABINP	AEBP
ZGDEMM	CUT2000	Min	0.1530	0.2370	0.0000	0.0000	0.0000	0.2430	0.0000	0.0000
		1stQu.	0.2220	0.3100	0.3165	0.4010	0.4115	0.3170	0.2630	0.2160
		Median	0.2490	0.3290	0.7160	0.6940	0.6900	0.3360	0.7870	0.7770
		Mean	0.2507	0.3266	0.6178	0.6321	0.6324	0.3344	0.6274	0.6148
		3rdQu.	0.2760	0.3450	0.9520	0.9230	0.9105	0.3525	0.9840	0.9870
		Max	0.3800	0.4240	1.0000	1.0000	1.0000	0.4430	1.0000	1.0000
	CUT0000	Min	0.0650	0.2440	0.0000	0.0000	0.0000	0.2490	0.0000	0.0000
		1stQu.	0.2240	0.3090	0.3342	0.4188	0.4300	0.3170	0.3430	0.2728
		Median	0.2470	0.3280	0.8025	0.7670	0.7615	0.3350	0.8360	0.8320
		Mean	0.2503	0.3258	0.6498	0.6617	0.6620	0.3338	0.6613	0.6464
		3rdQu.	0.2740	0.3450	0.9790	0.9530	0.9460	0.3530	0.9870	0.9922
		Max	0.3820	0.4020	1.0000	1.0000	1.0000	0.4240	1.0000	1.0000
ZGDEN	CUT2000	Min	0.1530	0.2370	0.0000	0.0000	0.0000	0.2430	0.0000	0.0000
		1stQu.	0.2220	0.3100	0.3170	0.4010	0.4110	0.3170	0.2650	0.2220
		Median	0.2490	0.3290	0.7160	0.6980	0.6920	0.3360	0.7880	0.7860
		Mean	0.2509	0.3266	0.6189	0.6332	0.6334	0.3344	0.6284	0.6158
		3rdQu.	0.2760	0.3450	0.9520	0.9230	0.9110	0.3530	0.9840	0.9880
		Max	0.3800	0.4240	1.0000	1.0000	1.0000	0.4430	1.0000	1.0000
	CUT0000	Min	0.0000	0.0530	0.0000	0.0000	0.0000	0.1370	0.0000	0.0000
		1stQu.	0.2040	0.3000	0.6655	0.6718	0.6655	0.3010	0.7208	0.6988
		Median	0.2510	0.3250	0.9950	0.9970	0.9920	0.3260	1.0000	1.0000
		Mean	0.2456	0.3232	0.7782	0.7899	0.7900	0.3285	0.7908	0.7840
		3rdQu.	0.2990	0.3450	1.0000	1.0000	1.0000	0.3470	1.0000	1.0000
		Max	0.5530	0.8220	1.0000	1.0000	1.0000	0.9160	1.0000	1.0000

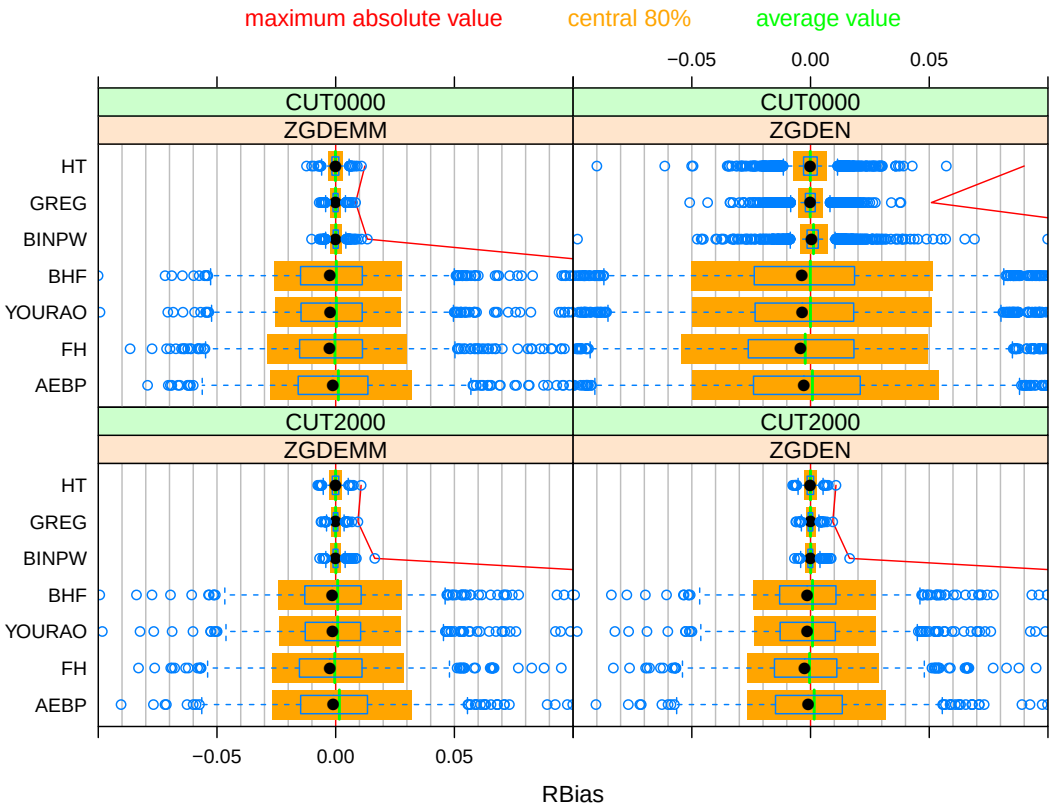

Figure 5.6: RBIAS of the point estimate for the four scenarios

The design-based estimators, such as the GREG and the HT, are (asymptotically) unbiased. Yet, the estimate between different samples often alternates very strongly. This can also be observed in figure 5.7. Thereby the HT shows greater variations of the estimates than the GREG, which achieves more stability in the point estimation because of the assisting model. As can be seen as well, the point estimates of the GREG and HT vary very strongly, particularly for smaller areas (ZGDEN/CUT0000). Furthermore, the RDISP of the BINPW is extremely high for a small area estimator. However, it still shows smaller RDISPs than the GREG.

The comparison of the area-models indicates that the FH has a slightly but yet obviously higher RDISP than the AEBP. The YOURAO and the BHF even have a slightly higher RDISP than the FH. Overall, the YOURAO, the BHF, the AREABINEBP and the AREAEBP do not show an elevated RDISP, even when using the very small areas instead of excluding them.

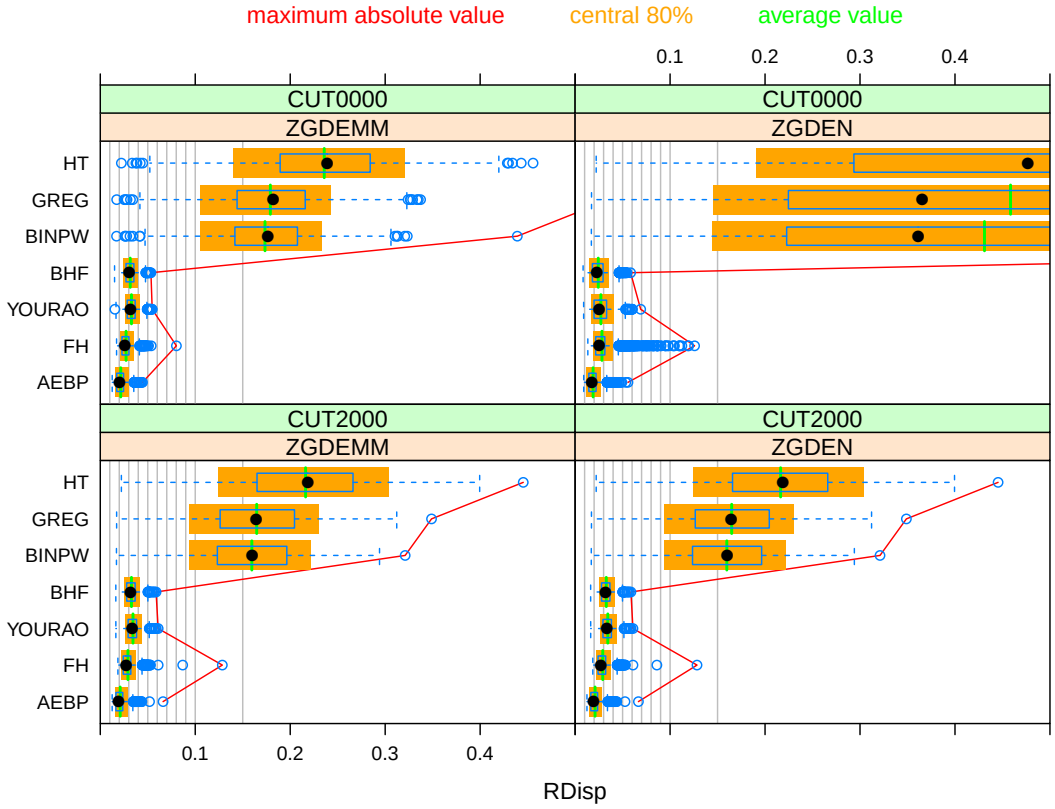

Figure 5.7: RDISP of the point estimate for the four scenarios

It is of great interest to know how the RRMSE behaves depending on the community size. Compliance with certain quality requirements is often linked to the community size. Thus, a method that does provide considerably more precise point estimators in small areas, but loses too much precision in huge areas, often, at least for huge areas, is not suitable. In Figure 5.8 this question is examined. For this purpose, the RRMSE per area

is plotted against the logarithmic area sizes for the scenario ZGDEN/CUT0000. For the HT and the GREG a clear relationship between area size and RRMSE is visible. The greater an area is, the lower the RRMSE becomes. A similar finding is provided by the BINPW. Here too, the same relationship between RRMSE and area size is observed. It is again depicted here that, due to the proportional allocation of the sample size, larger areas have a higher absolute sample size.

The YOURAO and the BHF show a different result. For both of them, even relatively small areas already achieve a quite low RRMSE, whereas some slightly greater areas also show a higher value than the GREG or the BINPW. This pattern can also be observed in the area level estimators. Altogether, it can be asserted that the small area estimators perform very well for all kinds of area sizes, with a few outliers.

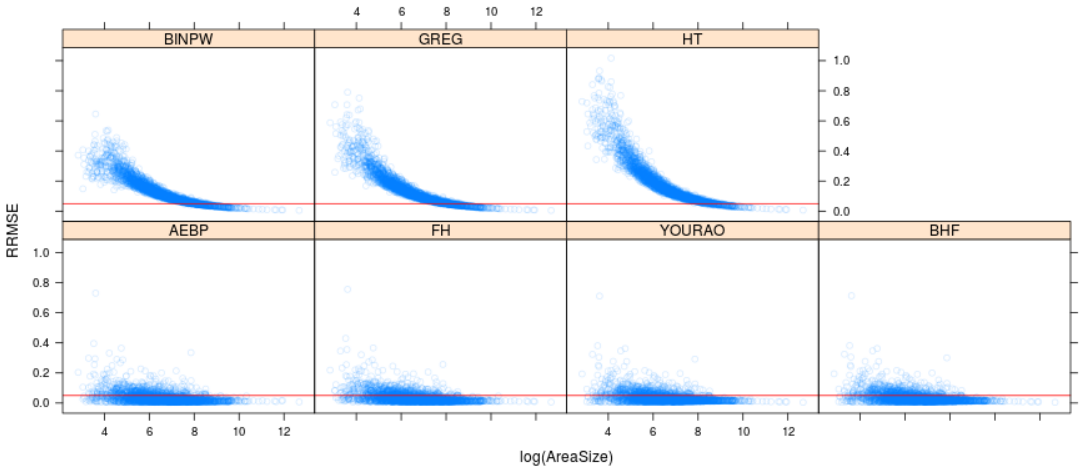

Figure 5.8: RRMSE of the point estimates versus the logarithmic area sizes in scenario ZGDEN/CUT0000

Next, in regard to the question of the relationship that exists between the area size and the quality of the point estimator, it is also important to consider whether the exclusion of too small areas from the models can lead to better estimation results for the more important huge areas.

This question can be analysed with the help of Figure 5.8. This illustrates the difference of the RRMSE when using all areas (ZGDEN/CUT0000), compared to the situation where the areas below 2,000 inhabitants are omitted (ZGDEN/CUT2000). Negative values mean that the RRMSE of a community is lower for CUT2000 than for CUT0000, i.e., that the exclusion of areas with less than 2,000 inhabitants leads to a reduction of the RRMSE in the other areas. The figure shows that there is no difference, either for the GREG or the HT, between the two scenarios. For the HT, it has to be this way, as no information from other areas has been added and thus the information used does not change between the two scenarios. For the GREG, differences between the scenarios are theoretically possible. Usually they are very small, as the GREG contains the correction term for the error of the model.

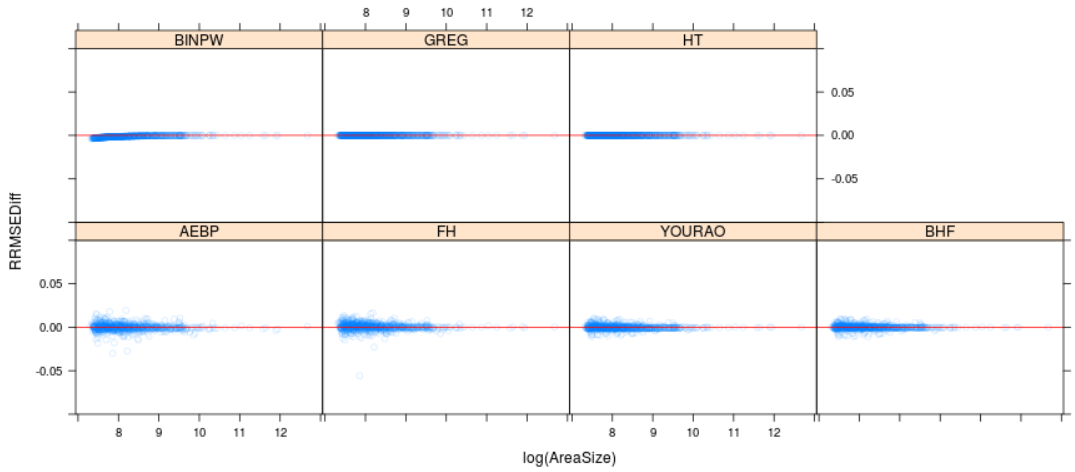

Figure 5.9: Change in RRMSE of the point estimates versus the logarithmic area sizes when dropping the very small areas (scenario ZGDEN/CUT2000)

For the BINPW, through the exclusion of the areas with less than 2,000 inhabitants, a slight improvement is to be seen. However, this only benefits areas that are not much greater than the excluded areas, with less than 2,000 inhabitants. The very huge areas are not affected by this. For the other estimators in the figure, there is no clear pattern visible, so it cannot be concluded whether the exclusion CUT2000 represents an improvement or a decline for the estimation results for the YOURAO, the BHF, the FH or the AEBP.

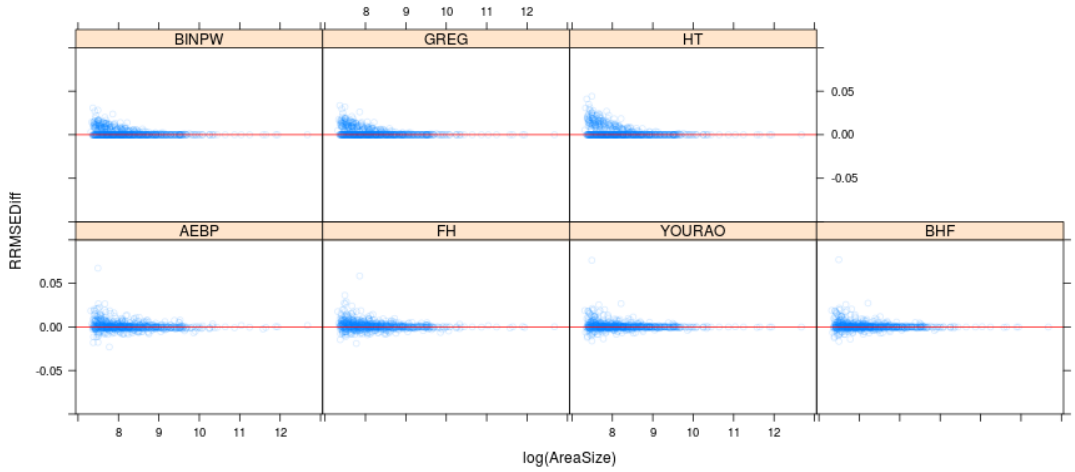

Figure 5.10: Change in RRMSE of the point estimates versus the logarithmic area sizes when dropping the very small areas (scenario ZGDEMM/CUT2000)

The same analysis for the scenarios with ZGDEMM agglomerations of 2000 shows another picture. As is seen in Figure 5.10, all estimates worsen somewhat when the areas with less than 2,000 inhabitants are excluded from the estimation. As some parts drop out because of the exclusion of the small areas from some agglomerations, the differences

of the RRMSE are no longer identical to zero for the HT and the GREG as well. Furthermore, the BINPW worsens considerably when excluding the communities with less than 2000 inhabitants. This leads to the conclusion that a combination of both scenarios means that the use of an agglomeration and the additional exclusion of small areas can even lead to a worsening of the estimation results. For this reason, in the following only the variance and MSE estimation for both of the CUT0000 scenarios are examined.

5.4.2.2 Evaluation of the MSE Estimates

In general, the precision of a point estimator is indicated by an appropriate variance or MSE estimator, which has to be calculated from the sample as well. In the following, as a simplification, variance estimators, as well as MSE estimators are denoted as MSE estimators. Variance estimators, in a strict sense, are only the variance estimators of the GREG and the HT. However, as these estimators are unbiased (HT) or asymptotically unbiased (GREG), their variance estimators are in a wide sense also MSE estimators. In the simulation, the quality of the precision measures can be tested. Of particular interest as precision measures are the confidence interval coverage-rates in combination with the average confidence interval lengths.

For confidence interval coverage-rates in a simulation, the share of the confidence intervals that covers the true value is calculated. Ideally, this corresponds to the nominal coverage-rate of 95 % (see equation (5.17)). In the context of the small-area-problems, often small samples result in each area. In this case, often a confidence interval build with the student distribution instead of the normal distribution is preferred. The same applies to the confidence interval for the small area estimates with an analytic approximation to the MSE or a jackknife MSE estimate. For the estimators which rely on bootstrap methods, such as the AEBP, two approaches are possible as mentioned in Sections 3.3.6.2 and 3.3.6.3. On the one hand, the MSE of the bootstrap distribution is taken as the MSE estimator for the MSE of the point estimate and, on the other hand, the confidence interval is obtained by taking the corresponding interval from the bootstrap distribution.

Therefore, if a 95 % confidence interval is analysed, the confidence interval should cover the true value in 95 % of the samples. Furthermore, it is desirable that the confidence interval lengths be as short as possible. A short confidence interval indicates a high precision of the estimation. However, if the confidence interval lengths are short and the coverage-rates are low, the MSE estimator pretends to have a higher precision of the point estimation then it effectively has.

For the AEBP, four different CI estimators are used. These are built as follows

AEBPJKALL Jackknife, each area is dropped once (see Section 3.3.6.1).

AEBPJKSRS Grouped jackknife with 100 groups of an almost equal number of areas. Each group of areas is dropped once (see Section 3.3.6.1). The areas are assigned to the groups via simple random sampling without replacement.

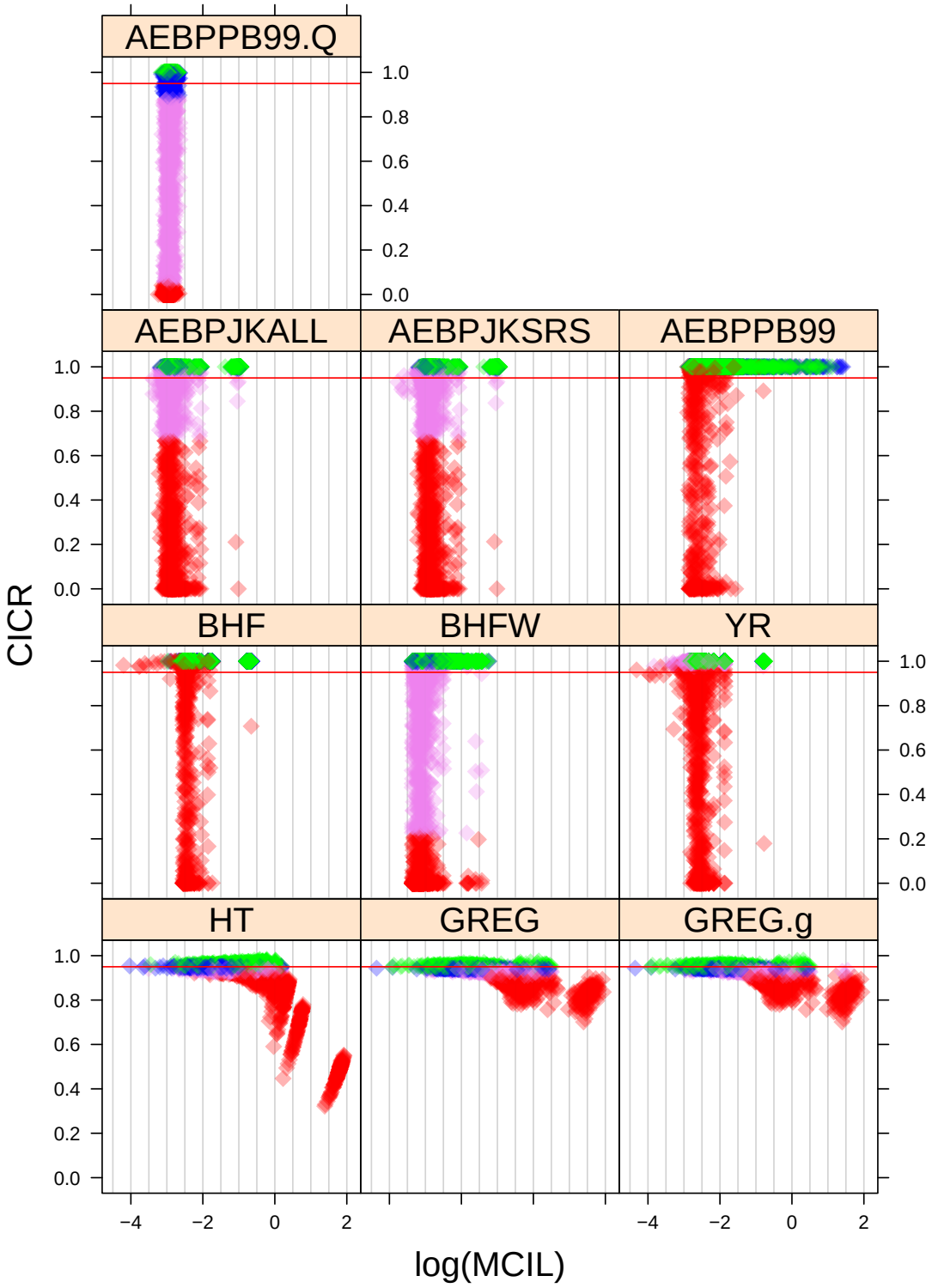

Figure 5.11: Confidence interval coverage rates versus the logarithmic mean confidence interval length for scenario ZGDEN/CUT0000

AEBPPB99 Parametric bootstrap with 99 bootstrap resamples. CI are built using the MSE of the bootstrap distribution.

AEBPPB99.Q Parametric bootstrap with 99 bootstrap resamples. CI are built using the 2,5% and 97,5% quantiles of the bootstrap distribution.

As mentioned in Section 3.3.7 the MSE estimator of the BINPW is still quite cumbersome, especially in terms of computation times. The use of a resampling procedure takes a prohibitively long time. Further, the point estimates are not as promising as the ones from the AEBP or the unit level small area estimators. The FH does not perform better in terms of RRMSE than the AEBP. The AEBP has the nice property in that it always gives estimates for proportions that lie in the interval $[0,1]$, which is not necessarily the case for the FH. Therefore, the MSE estimates of the FH and the BINPW will not be discussed further.

In Figure 5.11 for the scenario ZGDEN/CUT0000, the logarithmic MCIL are depicted on the x-axis and the CICR on the y-axis. Ideally, the points would lie entirely at the left on a red 95% line. The coloration should be interpreted as follows: the areas in the fourth quartile of the confidence interval rates for each estimator are shown in green, in blue those in the third quartile, in purple those in the second quartile and in red those in first quartile. For example, for the BHF in Figure 5.11, more than 75% of the confidence interval CICR for the estimator lies above the desired 95% CICR. On the contrary, for the AEBPPB99.Q more than 25% shows CICRs close to 0%. That the CICR for the BHFW perform worse than for the BHF is due to the fact that for the BHFW, the Prasad-Rao MSE estimator (see Section 3.3.3.1) without adjustment was used for the weighted estimation. It is to be expected that an approach similar to the one of Torabi and Rao (2010) would lead to considerable improvements.

For the HT, more than half of the confidence interval rates barely lie under the 95% mark (all red, purple and a few of the blue points). In particular, it can be observed that the confidence interval lengths are extremely long. Values above zero on this scale mean that the confidence interval length is above one, which is no longer helpful for the interpretation of a ratio. For example, a point estimator of 0.4 with a symmetric CI of length 1 would be $\big(0.4-0.5;0.4+0.5\big) = \big(-0.1;0.9\big)$. Even if this confidence interval had a coverage rate of 95%, the information is strongly limited. However, even with this extremely long confidence interval, the confidence interval coverage rates for the HT drop off up to 40%. A better picture is provided by the GREG. Here too, in part, the confidence interval lengths become extremely long and the confidence interval coverage rates do not reach the desired 95% coverage rates. However, these do not drop off as dramatically as occurs with the HT. The use of the g-weights for the variance estimation does not lead to a considerable improvement compared to the residual variance estimator, despite the theoretically better asymptotic characteristics. Thus, the extensive increase of computing time does not pay off.

Contrary thereto, the confidence interval coverage rates, as well as the confidence interval lengths, behave completely different for the small area estimators. Here, for all of the

listed small area estimators, the confidence interval coverage rates for many areas are 95% or more. At the same time, the confidence interval lengths are considerably lower for all areas. However, in return, the confidence interval coverage rates in some areas break down even more extremely than for the HT and for the GREG. The best performance in this comparison is provided by the unweighted unit-level estimator (BHF). As already mentioned, for more than 75% of the areas, the confidence interval coverage rates lie above 95%.

Furthermore, the logarithmic MCIL are still relatively moderate, with mostly under -2, thus MCIL of under 0,135. In relation to the confidence interval lengths, the AEPJKALL and the AEBPJKSRS certainly perform better, but for these many more areas do not reach the desired 95% CICR. For the parametric bootstrap confidence interval estimations AEBPPB99.Q, very short confidence interval lengths are achieved, but in turn considerable losses in the coverage rates are involved. For the confidence interval estimations on the basis of AEBPPB99, indeed more areas achieve a confidence interval coverage rate of 95%. In return, the MCIL is considerably closer to the CICR compared to the one of the AEBPPB99.Q.

The examination of the CICR in the case of ZGDEMM (see Figure 5.12) shows that the CICR is much more stable for the HT, as well as for the GREG. Although they do not reach the desired 95% in most cases, they are generally very close to this value. Also, the MCIL are considerably shorter, but with up to 0.36 are still very long. Furthermore, the small area estimators show that the use of the agglomeration ZGDEMM leads to a stabilization of the CICR. Hereby, most of the logarithmic MCILs lie under 0.1. In addition, many areas exceed the 95% line with their CICR.

An improvement of the confidence interval estimation for the parametric bootstrap is to be expected for an increase in the replications. At the same time, the parametric bootstrap has the lowest logarithmic MCIL. On the average, these lie at about 0.05. The use of parametric bootstraps with an insufficient number of replications can lead to pretended and, in practice, misleading accuracy, because of the many poor CICR paired with the very short MCIL. Contrary thereto, the Jackknife seems to function very well in the form of the AEBPJKALL, as well as in the form of the AEBPJKSRS (grouped). Most of the areas have a CICR near the 95% line and the logarithmic MCIL are relatively short, with about 0.05, even shorter than for the BHF. For this reason, in the case of the ZGDEMM it is a question of consideration whether the AEBPJKSRS with short MCIL and a few areas that do not achieve the 95% coverage rates or the BHF, that offers better coverage rates and in turn higher MCIL, should be used instead.

In the context of the examination of CICR related to the MCIL, possible biases of the variance or MSE estimator play a central role. In the simulation, this bias is quantified with the RBIAS of the variance and the MSE estimation (see (5.16)). It is very important to take note that the MSE of the point estimation has a lower Monte Carlo precision than the mean of the variance- and, respectively, MSE estimations, as the convergence according to the weak law of large numbers is slower . To be able to specify more precise RBIAS of

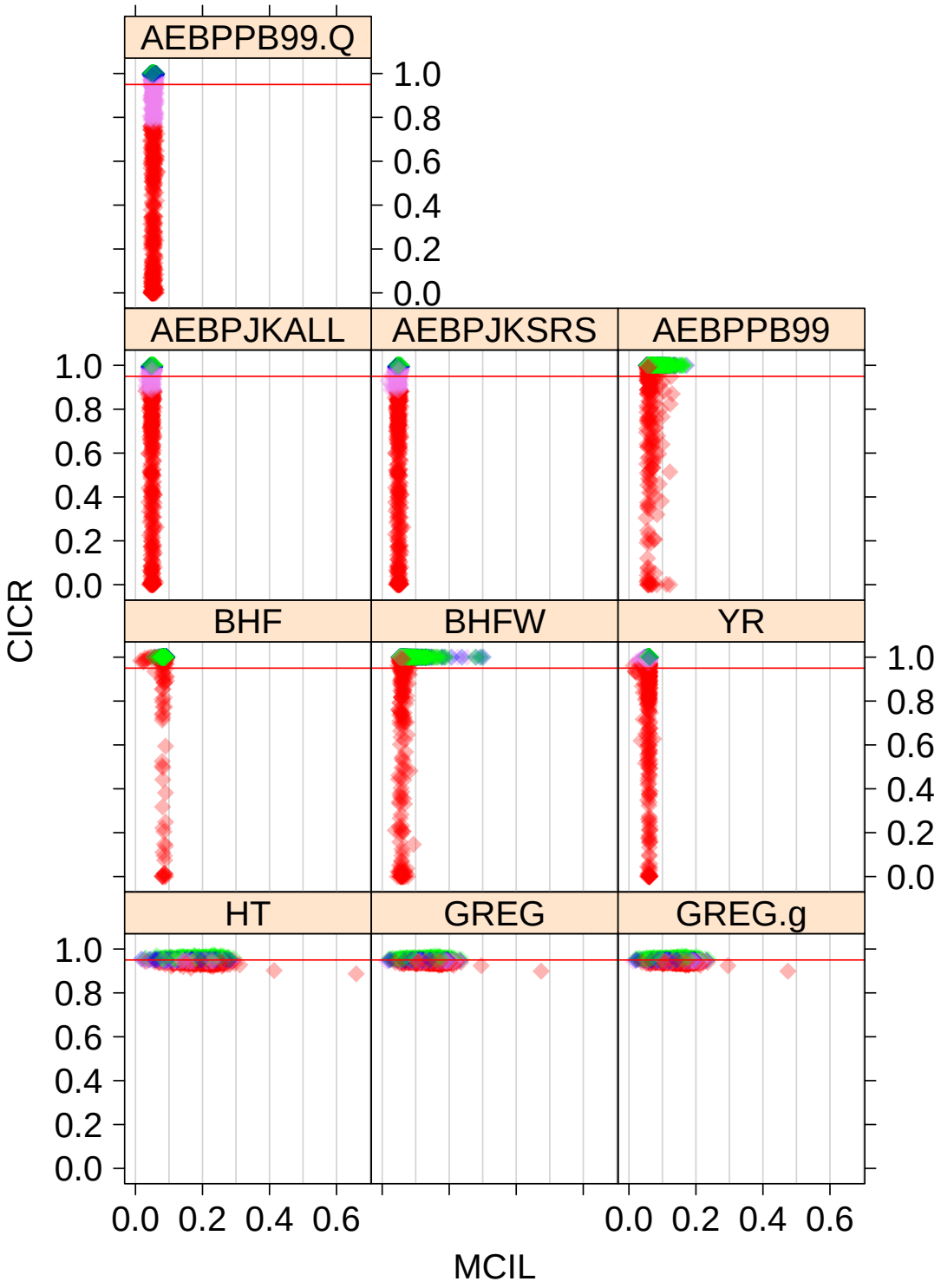

Figure 5.12: Confidence interval coverage rates versus the mean confidence interval lengths for scenario ZGDEMM/CUT0000

the variance- and MSE-estimations at this point, the computation of point estimators for more than 1,000 samples would be necessary. The effort to undertake such estimations, for example with 100,000 samples, exceeds acceptable time margins by far.

In Figure 5.13 the RBIAS of the MSE estimators is depicted for the scenario ZGDEN/-CUT0000. In order to better see the differences, the plot is repeated three times, only varying the range of the y-axis. The variance estimators of the HT and the GREG are almost unbiased in all examined regions. In the median of the areas, the BHF, YR and the AEBPPB99 show a positive bias of the MSE estimates. The BHFW, AEBPJKALL and AEBPJKSRS, in contrast, show a negative bias in the median of the area. As the MSE estimate for the BHFW is a plugin of the YOURAO estimator, it is not surprising that it fails to some extent in estimating the MSE correctly. A bit unexpected is the fact that the grouped jackknife for the AEBP (AEBPJKSRS) performs visibly better than the AEBJKALL.

The different zoom-levels again demonstrate clearly how strong the difference of MSE-estimations in single areas can be. Even if overestimations of MSEs in principle are more acceptable than underestimations, as this involves a more careful evaluation of the estimations' quality, such an overestimation should be avoided.

Figure 5.14 illustrates the distribution of the variance and MSE-estimation from the simulation for the small area 16. This figure demonstrates the problems of the variance estimation for the HT and the GREG. A peculiarity of estimators not using distributional assumptions is that for small sample sizes there are often only a limited number of possible estimates. This directly influences the variance estimation of the HT where, for this reason, there are only a small number of possible variance estimates available as well. Due to the use of the model, the GREG can take more different point estimates, and its variance estimator also attains more different variance estimates. In this constellation, using the variance of the point estimate as a benchmark for the variance estimator is obviously not very helpful. The model based estimators can deal better with the small sample sizes in the MSE estimation. This is due to the fact that these estimators have parametric assumptions about the distribution of the point estimator and estimate on this basis. Thus, even for small sample sizes, the point estimator can take many different values.

5.4.3 Conclusion

The model based estimation methods perform substantially better in relation to the quality of the point estimation of the active population ratio than the design based methods. As explained before, the BINPW does not show the expected performance and thus lies between the design based and the other model based estimates in terms of RRMSE. The findings indicate that among the model based estimates, the BHF and the YR provide the best results. As the design weights in the Swiss Structural Survey do not vary (considerably) because of the proportional design, the YR yields no improvement. Furthermore, it has to be noted that the exclusion of small areas in the estimation brings no considerable

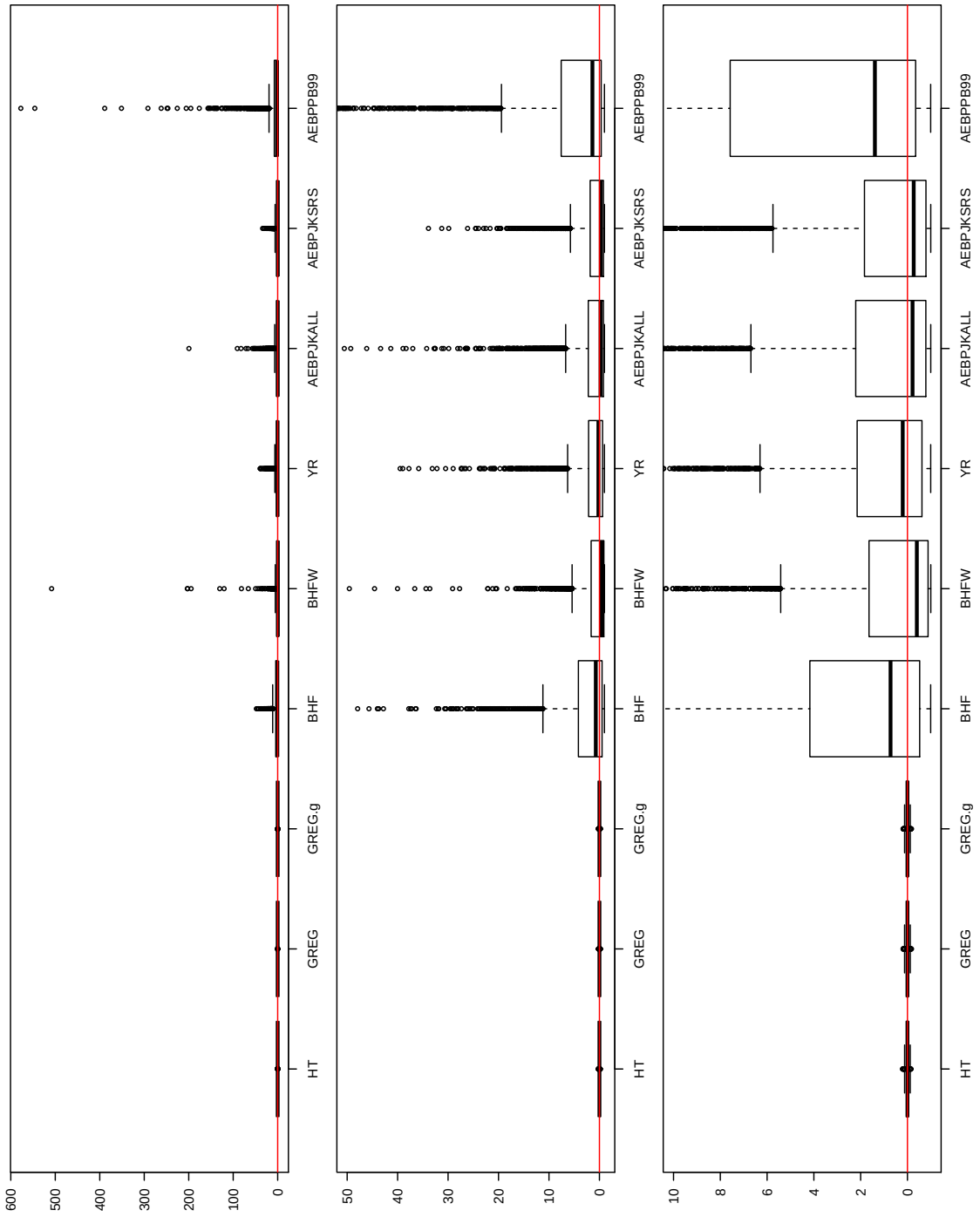

Figure 5.13: RBIAS of the variance and MSE estimates in the scenario ZGDEN/-CUT0000

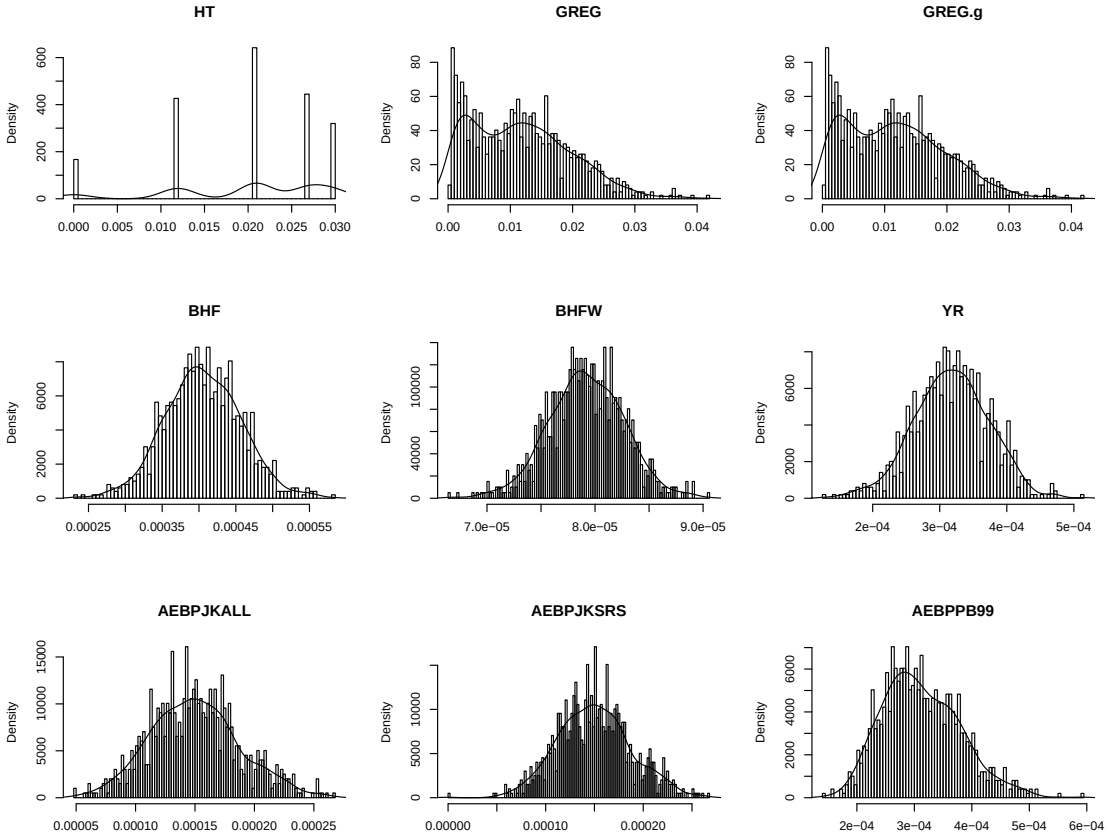

Figure 5.14: Distribution of the variance and MSE estimators in area 16 (< 300 inhabitants).

improvement of the estimation quality for the remaining areas. The drawback of the linear estimation methods, such as the BHF, is that they may provide estimates of proportions of over one or under zero. Taking this into account, along with the fact that the AEBP yields almost as good estimates, the AEBP is also to be considered.

As can be seen in the part on the MSE estimation, the jackknife MSE estimates work quite well for most of the areas. Also, the MSE estimate for the BHF provides satisfactory results. Although the design based CICR are more reliable than the ones from the BHF and AEBP, they have a very long MCIL, leading to uninformative estimation results. Therefore, also from the view of the MSE estimation, it is reasonable to opt for the use of the BHF or the AEBP as the estimator for the proportion of the active population. Of course, this is only one decision in the decision theoretical problem. If some areas are given more weight in the decision of which estimator to use than other areas, the decision of which estimator to use may vary greatly.

5.5 Small Area Estimation with Additional Information Sources on Different Aggregation Levels

For the estimation of the proportion of unemployed, an interesting additional information source would be the job centre register. For disclosure reasons, these registers are not usually available on unit level. Most figures reported are either on area level or on demographic domains on a national level. However, it might be possible to obtain the figures on demographic domains on area level, in the event the area is large enough to prevent disclosure. This situation is studied in this section using the modelling proposed in Section 3.4.

5.5.1 Setting of the Monte Carlo Simulation

The demographic domain considered is a cross combination of the variables *AGE* and *SEX*, both of which are usually available from the population registers. The variable *AGE* is partitioned into four classes and *SEX* into two classes. Hence, the total number of cells for the intermediate model is eight. Therefore, for the intermediate model, in every area there are eight observations of the cell totals. However, it can well be, especially for really small areas, that some of the cells have no units even in the universe. These cells will logically always have zero observations in the sample.

In this Monte Carlo simulation, the quality of the small area estimates on the three different levels, namely *unit-level, intermediate-level*, and *area-level*, are compared for underlying linear normal mixed models (FH and BHF) and binomial logit mixed models (AEBP and LogitMM). In order to see the improvement by using the additional registers, the estimation is performed once *with* the additional registers and once *without* the

Table 5.4: Coding of the variable AGC

AGC 1		$<$	*AGE*	$\leq$	20
AGC 2	20	$<$	*AGE*	$\leq$	60
AGC 3	60	$<$	*AGE*	$\leq$	65
AGC 4	65	$<$	*AGE*		

Table 5.5: Estimators and its information sources when estimating with additional registers with disclosure issues.

	Linear Normal Model			Logit Model		
Aggregation Level	Estimator	Register		Estimator	Register	
		Population	Additional		Population	Additional
Unit	BHF	(✓)	X	LogitMM	(✓)	X
Intermediate	BHF	✓	(✓)	LogitMM	✓	(✓)
Area	FH	✓	✓	AEBP	✓	✓

legend: ✓ := usually available; (✓) := sometimes available; X:= usually not available.

additional registers. Furthermore, two area sizes were studied. On the one hand, all municipalities were used as areas (denoted with *N*) where the sizes go down into the tens. On the other hand, accumulated municipalities were used as areas (denoted with *MM*). For this, the municipalities were accumulated to areas of over 2,000 inhabitants where it was geographically and politically feasible. An overview of the estimators and the information sources that are generally available is given in Table 5.5. While for the common researcher, at least the counts of certain cells in the population registers are are made available by the national authorities, the unit-level population register information might not always be available.

5.5.2 Results of the Monte Carlo Simulation

In this simulation the focus lies on the quality of the point estimates and the viability to use an intermediate level model as proposed in Section 3.4. Hence, the RDISP, RBIAS and RRMSE measures are considered in the following.

Relative Dispersion of the Estimates

From Figure 5.15, the first thing to see is that when using the additional register information, the linear and the binomial estimators act almost oppositely in terms of variability of the estimates. In the linear case, with additional register information, the lower the aggregation level of the estimation, the less variable the estimates for both agglomerations *N* and *MM*. In contrast, for the binomial estimators, the higher the aggregation level of the

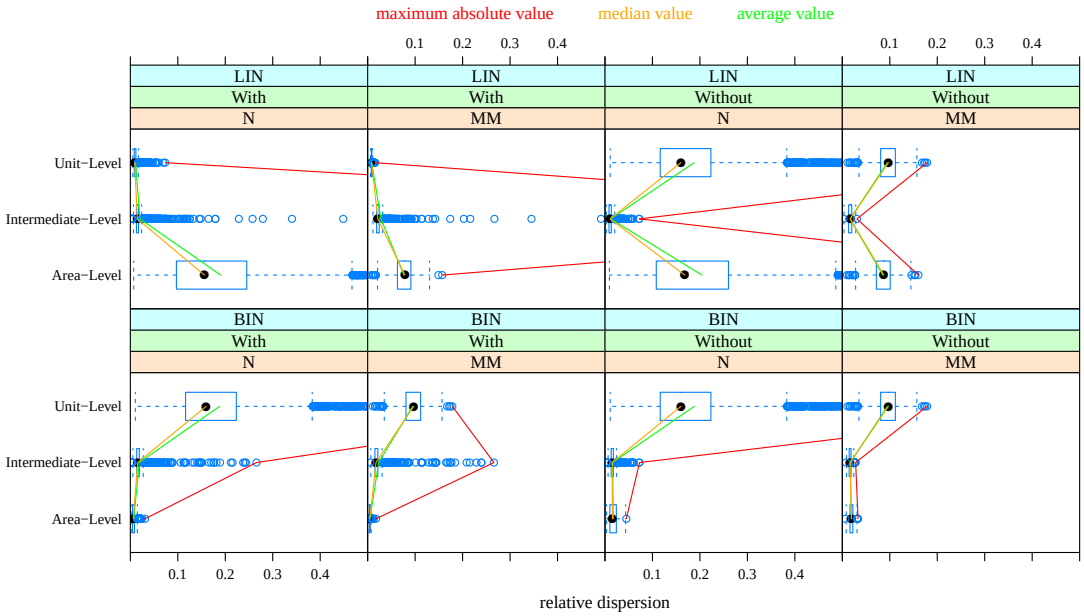

Figure 5.15: Relative dispersion of the estimation on different aggregation levels.

estimator, the lower its variability. The lowest variability is yielded in this comparison by the binomial Area-Level estimator (AEBP).

When no additional register information is available, then the picture is a bit different. Surprisingly, in the linear case, the intermediate estimator overrides both the unit-level and the area-level estimators in terms of variability. The same applies to the binomial estimators, where the area-level estimator has almost the same low level of variability as the binomial intermediate level estimator.

In all cases, the use of *MM* instead of *N* diminishes the variability. For the intermediate level estimators this effect is rather small, in contrast to the unit level and area level estimators, where the reduction is considerable. However, the use of *MM* instead of *N* comes with the cost of having less detailed estimates.

Another surprising point is that the variability of the estimates produced by the intermediate level estimators does not benefit much when adding the additional register information. Even the point with the maximal relative variation, denoted by the red line, increases enormously when using the additional information. Comparing the orange and the green line, which do not differ very much, one can see that the outliers in the intermediate level models are not numerous enough to increase the average variability over the median variability of the estimates.

Relative Bias of the Estimates

Even though small area estimators in general do not produce (design) unbiased estimates,

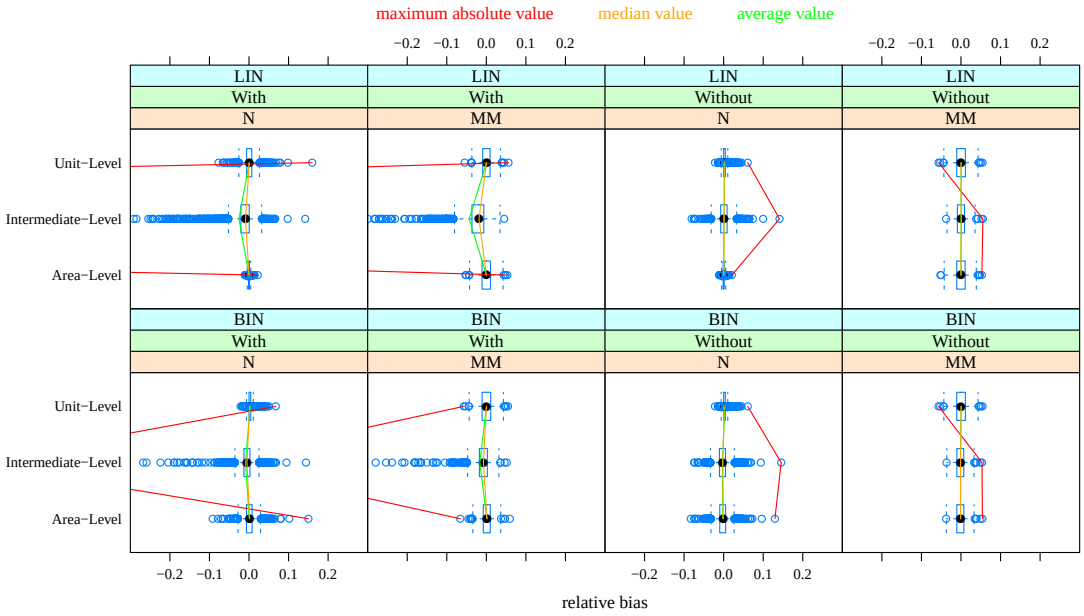

Figure 5.16: Relative bias of the estimation on different aggregation levels.

in this setting the RBIAS of the estimators is relatively low overall. Furthermore, the few areas with a bit of RBIAS are almost symmetrically distributed around 0. This can be seen either by observing the border of the boxes, by comparing the positively and negatively biased estimates or by confirming that the orange and green lines are almost identical. In order to better assess whether these estimators are empirically unbiased in this setting, one would have to increase the amount of simulation runs by far over 10,000 runs, as the convergence for improving the results in some decimal points is quite slow.

Only the intermediate level estimate seems to be slightly biased when using the additional register information. As in this case the AVRBIAS is visibly different from the median RBIAS, the outlying area estimates are considerably away from zero. This is the case for the linear and for the binomial intermediate level estimators with additional register information.

In general, one can see that by using the additional register information, the relative bias of many areas rises, sometimes enormously. This effect is much stronger in the case of the intermediate level estimators then in the case of unit level or area level estimators. From these results it seems that the model does not hold equally well for all areas.

Relative Root Mean Squared Error of the Estimates

In this setting, the relative root mean square error is driven mainly by the variability of the estimates and less by their bias.

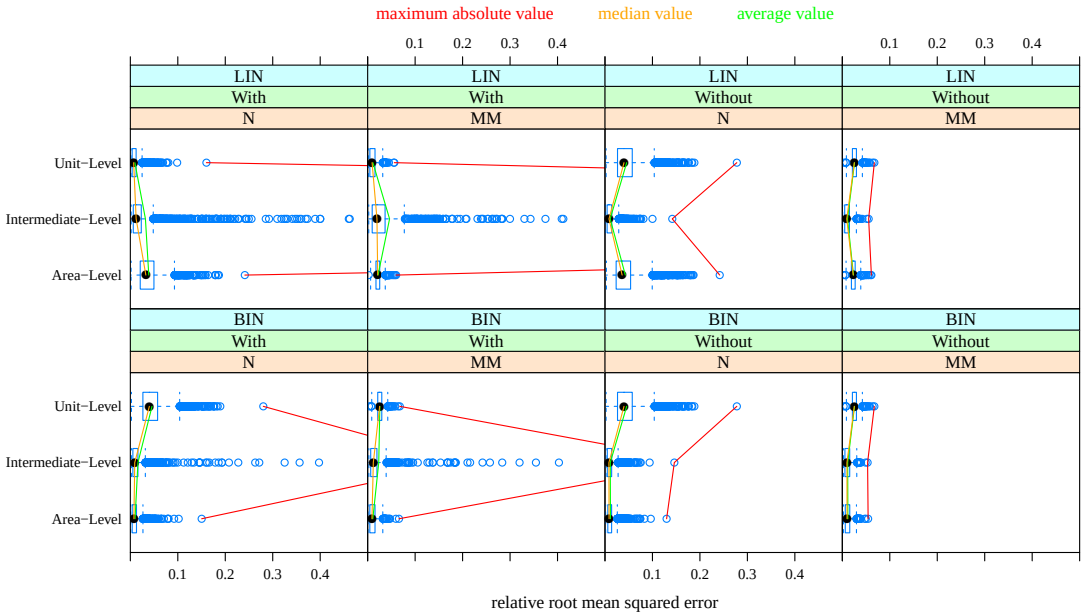

Figure 5.17: Relative root mean squared error of the estimation on different aggregation levels.

The linear unit level and area level estimators gain by using the additional register information. While this effect in the case of the area level estimator is only marginal, in the case of the unit level model it is important. In the intermediate level, the situation is the inverse. Here, the use of the additional information source deteriorates the estimation considerably. When using the agglomeration *MM* over the agglomeration *N*, the overall picture of the area RRMSEs is less variant. However, this effect is mainly due to the fact that in the case of the agglomeration *MM* for the really small areas, no estimates are produced.

In contrast to the linear estimators, under the binomial estimators the area level estimator performs better in terms of RRMSE than the unit level estimator. As in the case of the linear intermediate level estimator, the binomial intermediate level estimator has a higher RRMSE in some areas. This higher RRMSE is induced by both the variability and the bias of the estimates. Again, by using the agglomeration *MM*, the most extreme RRMSE are smaller, but no estimates are available for areas under 2,000 inhabitants.

5.5.3 Conclusion

If the partition used to construct intermediate level estimators partitions the dependent variable in such a way that a large amount of its variability is explained, then it may even yield better estimates compared to the unit level models.

The use of additional register information has to be conducted carefully. Especially when the additional register information is gained by combining similarly defined variables from decentralised registers, caution has to be taken that the different definition of this variable is not counterproductive.

When the information is only available on area level, then for the estimation of a proportion or a total it seems advisable to use the binomial model instead of the linear one. This is not only theoretically more appealing, as the proportion may not exceed the interval $[0, 1]$, but it also enhances the quality of the estimates, as they give very suitable results over all the combinations considered here.

Furthermore, even if the information is available on unit-level, it might be interesting to use a binomial area level estimator, as this yields good results in comparison to the unit level linear estimator, better results than the binomial unit level estimator and similar or better results than the binomial intermediate level estimator.

Chapter 6

Summary and Outlook

The aim of this research has been to evaluate small area techniques for applications in official statistics. The estimation in the small area context is generally based on regression models which have been presented in Chapter 2, *Regression Models for Small Area Estimation*, The regression models tackled include the general linear model, the general linear mixed model, the generalized linear model and the generalized linear mixed model. Besides the theory, algorithms for the estimation of the regression models have also been shown.

In Chapter 3, *Small Area Estimation*, the estimation and prediction of small area estimators have been depicted. First, a brief review is given of sampling designs developed for surveys that set out to use small area estimation techniques. Second, on the basis of this, classical design based estimators are described in a form suitable for small area estimation problems. Subsequently, third, the model based small area estimators are explicated. Beginning with the standard models in small area estimation, a presentation is given of the Fay-Herriot (FH) and the Battese-Harter-Fuller (BHF) models, the pseudo small area models incorporating design weights. In addition to point estimation, MSE estimation is also discussed. As many of the variables of interest in small area applications are binary, the models apt for binary dependent variables are also described. In particular, a binomial area-level predictor (AEBP) is developed, which relies on the work of Jiang and Lahiri (2001). For this estimator, a Jackknife, a grouped Jackknife, a parametric bootstrap and a parametric double bootstrap MSE estimator are also proposed. Last but not least in this chapter, fifth, a way to make use of register data which might only be available on certain aggregation levels due to disclosure reasons is proposed.

One major problem for the MSE estimation via the parametric bootstrap is the huge computation time, especially for complex estimators. Therefore, in order to reach practicability, the computational burden of this method has to be reduced. In Chapter 4, *Variance Reduced Parametric Bootstrap MSE Estimates*, variance reduction methods for easing the computational burden of the parametric bootstrap are proposed and evaluated exemplary for the FH within a model based simulation study. The two variance reduction methods

proposed are the *Latin Hypercube Sampling* (LHS) and the method of using *control variates* (CV). The LHS has not shown to be helpful in the studied model based scenarios. In contrast, the CV has shown to allow for a great reduction in the resampling required to obtain a certain stability of the MSE estimate. The reduction attained in the number of resamples in the model based simulation reached a massive 90%. Caution has to be exercised with the estimation of the σ_u when σ_u is very small. In this case special estimates for the variance component as proposed by H. Li and Lahiri (2010); Y. Li and Lahiri (2007) may improve the results.

In Chapter 5, *Monte Carlo Simulations and Simulation Studies*, the small area techniques are studied by design based Monte Carlo simulations on the basis of the Swiss Census of 2001. In order to clarify the differences between the various types of Monte Carlo simulations in the Survey Statistics context, a systematization of different Monte Carlo simulation types is proposed. Also, an example is given of what a combination of model and design based Monte Carlo simulation may look like. Subsequently, adequate measures are presented in order to assess the output of the simulation studies.

The questions tackled in a first design based Monte Carlo simulation are how small an area can be for small area estimation and whether the exclusion of very small areas can improve the estimation of the larger ones. For this purpose, four scenarios have been introduced where the estimation is done either on all areas, only on areas larger than a certain size, on different aggregation levels, or on all three. The exclusion of the very small areas has not shown to have a considerable impact on the estimation of the larger areas. Moreover, aggregating did not improve the estimation. The estimator which performed the best was the BHF estimator. The AEBP works nearly as well, with the advantage that the estimated proportion in the AEBP always ranges between zero and one. This is not necessarily the case for the BHF. In terms of the precision estimation, it should be noted that all estimators have some troubles reaching the nominal confidence interval coverage rate (CICR) in all areas. This is also true for the design based estimates, as the distribution of point and variance estimates is discrete, with a low number of possible values for small sample sizes. Hence, in really small areas the asymptotics are seen to have problems. Also, the linearised MSE estimate for the BHF and the grouped jackknife for the AEBP have problems meeting the nominal CICR. But, in combination with the much lower relative root mean square error (RRMSE) of the point estimates, they seem to be preferable to the design based methods. This is confirmed by a look at the near enough rate (cf. Section 5.2), which shows the strong advantages of the BHF and AEBP.

The second simulation study evaluates the approach to include intermediate level information (cf. Section 3.4). The result is that this approach is a viable way to improve the estimation for both linear and binomial small area estimators. Nevertheless, this approach has to be handled carefully. The additional information will not always bring the desired effect of a reduction in the RRMSE. A method to detect whether it would be useful in a certain situation would be an interesting research topic. In order to be applicable, a working MSE estimate has to be found. This is a further research topic.

The proposed variance reduction method CV has shown to be very promising (cf. Chapter 4). Currently, this method is being studied in depth also for other small area estimates. An interesting issue here is to find adequate functions which can be used as control variates. In addition, the extension of the CV method to the parametric double bootstrap is being developed. This is a very important development, as it would make the parametric double bootstrap computational more feasible.

Also, an R-package is in preparation, in order to allow for the estimation of the AEBP without having to deal with numerical issues. The MSE estimation techniques will also be included.

In this work, many small area techniques have been evaluated under a design based Monte Carlo study. Modern small area methods have shown to provide an interesting alternative to the classical design based estimators. In many situations, the small area estimates outperform the design based estimators. For small areas, in the case of binary variables even the confidence intervals for the small area estimates obtained higher coverage rates, along with smaller confidence interval lengths than the design based methods. Especially, the proposed variance reduction method of using control variates for the parametric bootstrap MSE estimation will computationally enable the use of complex small area estimators.

Appendix A

Statistical and Mathematical Background

A.1 Random Number Generation

For drawing random samples a crucial prerequisite is to have some kind of random numbers. Physically produced random numbers may be obtained e.g. by throwing a coin or rolling a dice. The physical generation of random numbers is obviously time consuming, particularly if millions of random numbers are needed. An early solution for this problem was found by Tippett (1927) who published a table of $40,000$ digits, which he obtained by taking digits *at random* from census reports. However, for many situations like sampling from a large population, $40,000$ digits are not enough. Therefore Kendall and Babington-Smith (1938) proposed a apparatus which they call *The Randomizing Machine*. This machine works similar to a wheel of fortune. A disk is divided into ten equally sized sections enumerated with the digits 0..9. In a dark room this disk is brought to rotate. Then a flashlight illuminates a small part of the disk in *at random intervals* for a short duration, in such way, that each time only one number can be observed. With this machine they obtained in their setting in average one random digit every three to four seconds (Kendall & Babington-Smith, 1938).

Although unlimited amount of random numbers may be produced with this machine, it still takes a lot of time to obtain them. Further, these random numbers cannot be reproduced at a later stage, which in some cases is desirable. In the not uncommon situation, where one would like to reproduce an experiment conducted in the past, and for some reaseon, e.g. storage costs, only some of the random numbers are still available. By drawing new random numbers the outcome of the experiment will change. Therefore, one cannot reproduce the results and check whether the mathematical routines were implemented correctly. For a more detailed overview of the different approaches applied to obtained random numbers see Knuth (1981, § 3.1). Lehmer (1951) proposed the so called

Linear Congruential Method (LCG). According to Knuth (1981, § 3.2.1) this method is one of the most popular random number generators, and is easy to implement. However, this class of generators is has some drawbacks which are discussed in Entacher (1998).

One big advantage of the LCG is that, given a set of parameters, all random numbers can be reproduced easily by just starting the LCG again. On the other side it might seem counterintuitive to call number produced this way random, as they can be reproduced deterministically. In fact, for encryption problems the LCG is not suitable. But as Knuth (1981, § 3.1) states *'[...] the sequence* isn't *random, but it* appears *to be.[...] Being* apparently random *is perhaps all that can be said about any random sequence anyway*. There exist many different test to see whether a sequence of random number are *apparently random*.

A widely applied random number generator is the Mersenne-Twister by Matsumoto and Nishimura (1998). His great popularity is due to the fact, that the length of its random sequence is $2^{19937}-1$ and thus, long enough for most applications. Furthermore, it reaches a 623-dimensional equidistribution (Matsumoto & Nishimura, 1998). The higher the dimensions are in which the random numbers seem to be equally distributed the more they appear to be at random (c.f. L'Ecuyer, 1994). One drawback of the Mersenne-Twister is that in case of bad initial values, it needs about $700,000$ random numbers before converging to a good random sequence (c.f. Panneton, L'Ecuyer, & Matsumoto, 2006). A random number generator that needs much less time (about 700 random numbers) to converge to a good random sequence is given by Panneton et al. (2006) and is called WELL.

Algorithm A.1 Inverse Transform Sampling

Let U be a K-dimensional independently distributed random variable with pdf f_U and cdf F_U. One wishes to obtain a random vector $u=(u_1,..,u_K)$ with the u_k lying in the interval $[a_k,b_k] \subseteq$ image of f_{U_k}.

1. Set k=1.
2. Draw $x_k \sim \text{Uniform}(F_{U_k}(a_k), F_{U_k}(b_k))$.
3. Set $u_k = F_{U_k}^{-1}(x_k)$, $d = 1..D$.
4. Repeat steps 2 and 3 $k = 2..K$ times.
5. Combine the u_k to $u=(u_1,..u_K)$.

(Robert & Casella, 2004, § 2.1.2)

If the random variable U is to be normally distributed one has the problem that there is no analytical representation of F_U. In this case step 3 can be approximated by using the algorithm of Wichura (1988), which gives correct values for up to 16 digits. Further, if

U is multivariate normal with $u \sim \text{MVN}(\mu, \Sigma)$, one can use the Cholesky-decomposition $A^T A = \Sigma$ in order to obtain a sample from this distribution as described in Algorithm A.2 on page 130 in the appendix.

Algorithm A.2 Generation of Multivariate-Normal Random Vectors

Let U be a K-dimensional multivariate-normal distributed random variable with the mean vector μ and the variance-covariance matrix Σ. One wishes to obtain a random vector $u = (u_1, .., u_K)$ from K standard-normal random numbers $x_k, k = 1..K, X = (x_1, .., x_K)$.

1. Compute the Cholesky-Decomposition $A^T A = \Sigma$.

   (see. e.g. Rizzo, 2008, § 3.2)

2. Compute $U = J\mu^T + AX$, with J being a column vector of ones.

Step 2 in algorithm A.2 is a linear transformation. Thus, the standardized marginal distributions of the x_k in algorithm A.2 are the same as before, or exactly reverse in order (in the case that there was a negative covariance). Both, StrRS as LHS build symmetric strata, therefore it wouldn't matter if the standardized marginal distributions were in reverse order.

Concatenating algorithms the stratified random sampling and 4.1 with the algorithms A.1 and A.2, samples can be drawn from a multivariate-normal distribution with arbitrary mean μ and variance-covariance-matrix Σ.

A.2 Computational Integration Methods

A.2.1 Numerical Integration

A Common problem in statistics is to find the integral of an arbitrary function, which is not analytically tractable. A prominent case is the cumulative distribution function *cdf* of a normally distributed random variable X with mean μ and variance σ^2. The cdf can be written as

$$F_X(x) = \int_{-\infty}^{x} \frac{e^{-\frac{(j-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} dj \quad . \tag{A.1}$$

This integral has no closed form. Therefore, in many cases the expectation of a function ψ of a random variable X cannot be derived analytically. The expectation is defined as:

$$\mathrm{E}[\psi(X)] = \int_{-\infty}^{+\infty} \psi(x) f_X(x) dx \quad , \tag{A.2}$$

f_X being the *probability density function* (pdf) of the variable X. E.g. for a normally distributed random variable X with mean μ and variance σ^2 the pdf is

$$f_X(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \quad . \tag{A.3}$$

In order to obtain an approximation of this expectation two main concepts can be used: Numerical Integration and Monte Carlo Integration. For simplicity consider the case that the integral of the function $\psi(x) = e^{-x^2}$ is to be computed over the interval $(-\infty, \infty)$. This function has a bell shape (see figure A.1) and is closely related to the pdf of the standard normal distribution.

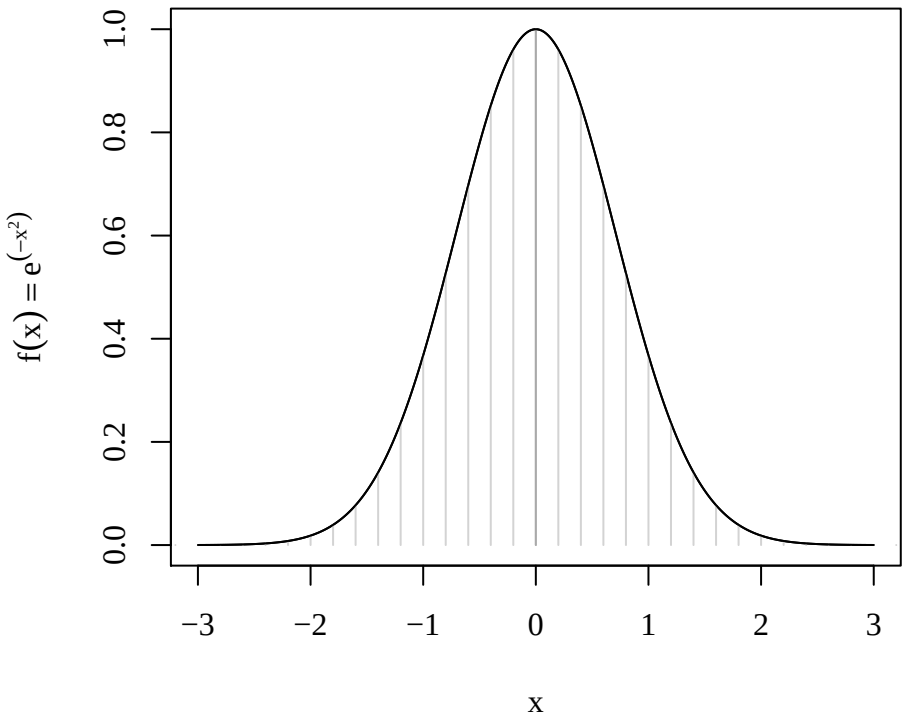

Figure A.1: Graph of the function $\mathrm{f}(x) = e^{-x^2}$

The integral of this function can be expressed as $\int e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \operatorname{erf}(x) + c$, where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function and c a constant.

In numerical integration the integral is approximated generally by a second function which is easily integrable such as polynomials. In case of a polynom of degree J all its coefficents may be computed by knowing $J+1$ pairs $\{(x_0, \mathrm{f}(x_0)), \ldots, (x_J, \mathrm{f}(x_J))\}$.

A general form of the formula used for numerical integration is

$$\int_a^b \mathrm{f}(x)dx \approx \sum_{j=0}^{J} \omega_j \mathrm{f}(x_j) \tag{A.4}$$

where the integral is approximated by the sum of J function evaluations $\mathrm{f}(x_j), a = x_0 < x_1 < x_k < x_J = b, j = 0..J$ weighted by appropriate weights $\omega_j, j = 0..J$. The differences between most of the numerical integration techniques lie in the choice of the $x_j \in [a,b]$ and ω_j.

The Midpoint Rule

The midpoint rule is the most simplistic of the numerical integration rules. It merely takes one function evaluation $\mathrm{f}(x), x = \frac{a+b}{2}$ and one weight $\omega = b-a$. Thus, the approximation to the integral is $(b-a)\,\mathrm{f}(\frac{a+b}{2})$. This method is visualized in figure A.2 on the left side. An improvement can be easily achieved by applying the midpoint rule on J equidistant intervals within the intervals of interest. The interval borders for the j-th interval are then $a_j = a + \frac{j(b-a)}{J}, j = 0..(J-1)$ and $b_j = a_{j+1}$. The points that have to be evaluated are then, following the midpoint rule, $x_j = \frac{a_j+b_j}{2}$, and their weights are $\omega_j = (b_j - a_j)$. The approximation formula for the integral via the cumulative midpoint rule is then $\sum_{j=0}^{J}(b_j - a_j)\,\mathrm{f}(\frac{a_j+b_j}{2})$. The cumulative midpoint approximation is visualized in figure A.2 on the right. As can be seen by comparing both sides in A.2, the cumulative midpoint method approximates the value of the integral much better than the pure midpoint rule. The narrower the intervals are chosen the better the approximation will be, however, the computation time also rises, as more evaluations of the function are needed. This approach is similar to the Riemann Sums.

The Trapezoidal Rule

In contrast to the midpoint rule, where only the function value at the middle of the interval is taken into consideration, the trapezoidal rule uses both endpoints of the interval. Therefore the two function evaluations $\mathrm{f}(a)$ and $\mathrm{f}(b)$ are needed. The applied weights are $\omega_0 = \omega_1 = \frac{b-a}{2}$. The approximation of the integral via the trapezoidal rule is then $\frac{b-a}{2}\big(\mathrm{f}(a) + \mathrm{f}(b)\big)$. This approach is visualized in figure A.3 on the left. The trapezoidal rule can also be used within a cumulative setting. Similar to the case of the cumulative midpoint rule the interval of interest can be split into J equally wide intervals $[a_j, b_j]$ with a_j and b_j as before. Then the trapezoidal rule is applied to each of

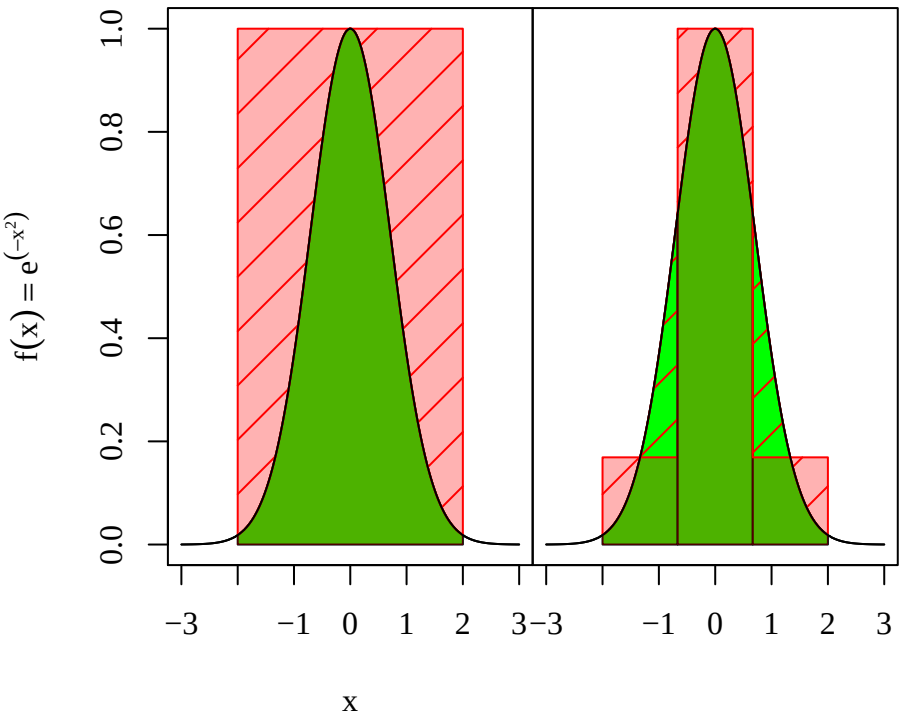

Figure A.2: The Midpoint Rule and the Cumulative Midpoint Rule Applied to the Function $\mathrm{f}(x) = e^{-x^2}$

these intervals and accumulated to obtain the approximation to the integral of interest $\sum_{j=0}^{J} \frac{b_j - a_j}{2} \big(\mathrm{f}(a_j) + \mathrm{f}(b_j)\big)$. As in this case $\mathrm{f}(b_j) = \mathrm{f}(a_{j+1})$ only $J+1$ function evaluations have to be performed and the formula can be simplified to $\frac{b-a}{2J} \big(\mathrm{f}(a) + \mathrm{f}(b) + 2\sum_{j=1}^{J-1} \mathrm{f}(a_j)\big)$. By comparing the cumulative trapezoidal rule on the right (in figure A.3) with the pure trapezoidal rule, the improvement becomes obvious. The trapezoidal rule is especially interesting for periodic functions as described in .

The Newton–Cotes rule

The Newton-Cotes rule can be split into two main branches. On the one hand there are the *open* Newton-Cotes rules. Open means, that the endpoints of the interval upon which the integral is to be approximated, are not used as a point of support for the approximation formula, that is $\sum_{j=1}^{J-1} \omega_j \mathrm{f}_j$. The midpoint rule is an open Newton-Cotes rule of degree one. On the other hand there are the *closed* Newton-Cotes rules, where the endpoints of the interval are used as points of support, that is $\sum_{j=0}^{J} \omega_j \mathrm{f}_j$. The trapezoidal rule is a closed Newton-Cotes rule of degree one. Generally, the equally spaced points x_j (sometimes also $\mathrm{f}(x_j)$) are given for the Newton Cotes rules. The degree of the Newton-Cotes rule denotes the degree of the polynomial used to approximate the function f. For a given degree it gives the optimal weights ω_j.

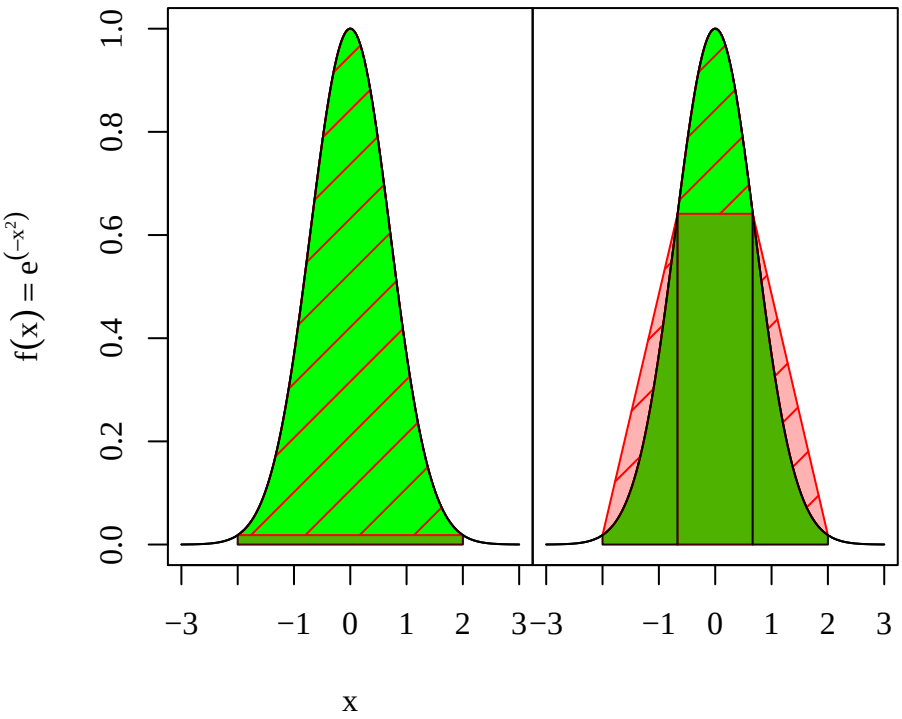

Figure A.3: The Trapezoidal Rule and the Cumulative Trapezoidal Rule Applied to the Function $f(x) = e^{-x^2}$

The names, formulae and error terms for the most popular Newton-Cotes quadrature rules are provided in table A.1.

One important advantage of the cumulative closed Newton-Cotes rules, is that it is easy to improve the precision of the approximation. In case that the error estimate seems too large to the researcher, the support can be enlarged by adding the points in the middle of the former points of support to the former support. The already performed function evaluations can be used again, and only the new set of points of supports have to evaluated. This property is called nesting.

Gaussian Quadrature Rules

The Newton-Cotes rules use ex-ante fixed points of support for the evaluation of the function and gives the optimal weights ω for these points. In contrast, the Gaussian quadrature rule determines the optimal pairs (x_j, ω_j) for the points of support and the weights for a given number of function evaluations. Similar to the Newton-Cotes rules, the Gaussian quadratur rule uses polynomials to approximate the shape of the function. Depending on the number of points that are to be evaluated, the used polynomial's degree changes.

Table A.1: Newton-Cotes rules

Open Newton–Cotes Formulae Name (Degree)	Formula	Error term
Closed Newton–Cotes Formulae		
Name (Degree)	Formula	Error term
Trapezoid rule (1)	$\frac{b-a}{2}(f_0+f_1)$	$-\frac{(b-a)^3}{12}f^{(2)}(\xi)$
Simpson's rule (2)	$\frac{b-a}{6}(f_0+4f_1+f_2)$	$-\frac{(b-a)^5}{2880}f^{(4)}(\xi)$
Simpson's 3/8 rule (3)	$\frac{b-a}{8}(f_0+3f_1+3f_2+f_3)$	$-\frac{(b-a)^5}{6480}f^{(4)}(\xi)$
Boole's rule (4)	$\frac{b-a}{90}(7f_0+32f_1+12f_2+32f_3+7f_4)$	$-\frac{(b-a)^7}{1935360}f^{(6)}(\xi)$

Table A.2: Two Gaussian Quadrature Rules

# nodes	x_j	ω_j	Error term
Gauss–Legendre Quadrature	$[-1,1]$		
1	(0)	(2)	
2	$(-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})$	$(1,1)$	
3	$(-\frac{3}{\sqrt{5}},0,\frac{3}{\sqrt{5}})$	$(\frac{5}{9},\frac{8}{9},\frac{5}{9})$	
Gauss–Hermite Quadrature	$(-\infty,\infty)$		

Therefore, the optimal pairs (x_j,ω_j) given a polynomial of certain degree p have to be derived. The general form of the approximation to the integral again is $\sum_{j=1}^{J}\omega_j f(x_j)$.

Since the Gaussian quadrature rule in general does not nest, once the approximation via the Gaussian quadrature is done, an improvement of the approximation needs a completly new evaluation.

Gauss-Konrod Quadratur Rule

The Gauss-Konrod quadratur extende the Gauss quadrature with J points of support, by $J+1$ points of support. As such in total $2J+1$ function evaluations are needed, where J function evaluations are already available from the before performed Gauss quadrature. The difference of the Gauss and the Gauss-Konrod approximation can be used as an error estimate. Kahaner, Moler, Nash, and Forsythe (1989, § 5.5), give an example of Gauss-Konrod quadrature depicted in table A.3.

Table A.3: Example of a 15-points Gauss-Konrod quadrature rule on the interval $[-1, 1]$ rounded to 5 digits

Gauss nodes		Weights
±0,949 11	G1	0,129 48
±0,741 53	G2	0,279 71
±0,405 85	G3	0,381 83
0,000 00	G4	0,417 96

Kronrod nodes		Weights
±0,991 46		0,022 94
±0,949 11	G1	0,063 09
±0,864 86		0,104 79
±0,741 53	G2	0,140 65
±0,586 09		0,169 00
±0,405 85	G3	0,190 35
±0,207 78		0,204 43
0,000 00	G4	0,209 48

Multidimensional Quadrature

The extension of numerical quadrature rules to the multidimensional space is quite cumbersome. Smolyak (1963) provided the mathematical foundation for an approach called sparse grid, which aims to set the nodes in the multidimensional space in an optimal way in order to obtain precise approximation to the integrand at least for smooth functions. Griebel, Zenger, and Zimmer (1992) developed a set of algorithms for the implementation of this method which was named *sparse grids integration*. A compact overview and some extensions to the sparse grid method can be found in Gerstner and Griebel (1998). Never-the-less, the multidimensional numerical integration techniques suffer from the so called *curse of dimensionality* (Bungartz & Dirnstorfer, 2003). That is, the number of needed funtion evaluations grows exponentially with the number of dimensions. In contrast, Monte-Carlo techniques do not suffer from this problem. Depending on the problem at hand, Monte Carlo integration perform better than the multidimensional numerical integration rules by Bungartz and Dirnstorfer (2003). This is especially the case the higher the dimension of the problem, and the less smooth the function is.

A.2.2 Monte Carlo Integration

As stated before, the definintion of the expectation of a function $\mathrm{f}(X)$ where X is a random number with probability distribution function f_X is

$$\mathrm{E}[\mathrm{f}(X)] = \int_{-\infty}^{+\infty} \mathrm{f}(x) f_X(x) dx \quad . \tag{A.5}$$

If X is random variable on the interval (a,b) then equation A.5 reduces to

$$\mathrm{E}[\mathrm{f}(X)] = \int_{a}^{b} \mathrm{f}(x) f_X(x) dx \quad . \tag{A.6}$$

If the random variable X is uniformly distributed on (a,b) then A.6 is even simpler

$$\mathrm{E}[\mathrm{f}(X)] = \frac{1}{b-a} \int_{a}^{b} \mathrm{f}(x) dx \quad . \tag{A.7}$$

A Monte Carlo integration can then be performed by drawing R random numbers $x_r, r = 1..R$ from the uniform distribution on (a,b) and computing the the following approximation:

$$\mathrm{E}[\mathrm{f}(X)] = \overline{\mathrm{f}(X)} \approx \overline{\mathrm{f}(X)}^{\mathrm{MC}} = \frac{1}{R} \sum_{r=1}^{R} \mathrm{f}(x_r) \quad . \tag{A.8}$$

Using the strong law of large numbers one can show, that with probability of 1 $\overline{\mathrm{f}(X)}^{\mathrm{MC}}$ converges to $\overline{\mathrm{f}(X)}$. More generally, for an arbitrary probability distribution function f_X the Monte Carlo approximation to the expectation is similar to equation A.8, but with x_r drawn from the distribution f_X.

A.3 Finding the Root of Real Valued Functions

One basic problem in statistics is to find the root of a continous function $\mathrm{f}(x)$ that maps from $\mathbb{R}^n \mapsto \mathbb{R}$. That is, the $x = x_1, ..x_n$ for which a function $\mathrm{f}(x) = 0$ has to be found. There are several methods to find this x.

Bisection Method

The bisection method searches for a root of a continous function f within a pre-defined interval $[a,b]$. It starts from the assumptuion that if $\exists a_0, b_0 \in [a,b], \mathrm{f}(a_0)\,\mathrm{f}(b_0) < 0$ then there must exist a $x \in [a_0, b_0]$ with $\mathrm{f}(x) = 0$. It is easy to see, that this method will not find roots which are at the same time extreme values of the function f.

Algorithm A.3 Bisection Method

Let f be a continous real valued function with a root in the interval $[a,b]$. Set $r = 0$ and $\varepsilon > 0$ arbitrary small depending on the precision needed.

1. Find $a_0, b_0 \in [a,b]$ such that $\mathrm{f}(a_0)\,\mathrm{f}(b_0) < 0$. E.g. by plotting the function and deducing them from the graph.
2. Set $x_r = \dfrac{a_r + b_r}{2}$. If $\mathrm{f}(x_r) = 0$ or $|\dfrac{x_r - x_{r-1}}{x_r}| < \varepsilon$ then stop and take x_r.
3. If

   $\mathrm{f}(a_r)\,\mathrm{f}(x_r) < 0$: set $a_{r+1} = a_r$ and $b_{r+1} = x_r$.

   $\mathrm{f}(b_r)\,\mathrm{f}(x_r) < 0$: set $b_{r+1} = b_r$ and $a_{r+1} = x_r$.
4. Increase r by one and proceed with step 2.

The Bisection method can be extended to solve multidimensional as shown by Morozova (2008).

Regula Falsi Method

The Regula Falsi method is very similar to the bisection Methods. It differs mainly in the update formula for the x_r. Instead of using the trivial guess of $x_r = \frac{1}{2}(a_r + b_r)$ it approximates x_r by equating the secant which crosses the points $(a_r, \mathrm{f}(a_r))$ and $(b_r, \mathrm{f}(b_r))$ to zero. As the x_r from the Regula Falsi method in most cases is much nearer to the searched value than in the Bisection method, a faster convergence can be expected.

Algorithm A.4 Regula Falsi Method

Let f be a continous real valued function with a root in the interval $[a,b]$. Set $r = 0$ and $\varepsilon > 0$ arbitrary small depending on the precision needed.

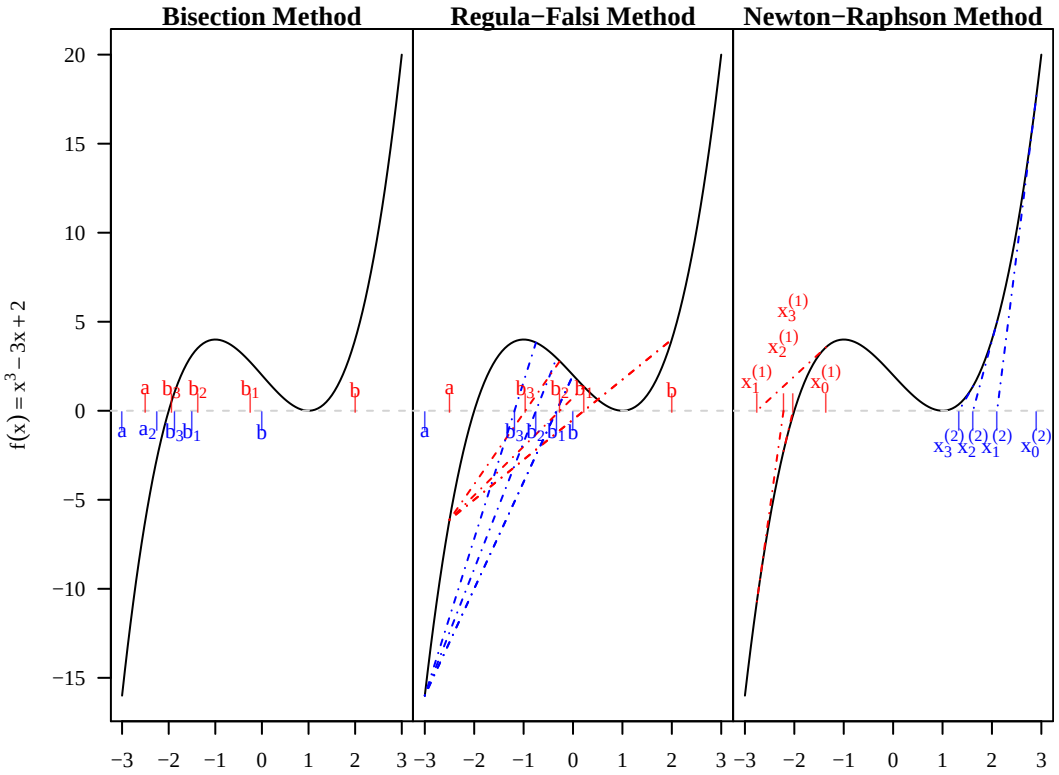

Figure A.4: Examplary comparision of the Bisection, Regula Falsi and Newton Raphson Methods

1. Find $a_0, b_0 \in [a,b]$ such that $f(a_0)\,f(b_0) < 0$. E.g. by plotting the function and deducing them from the graph.
2. Set $x_r = \dfrac{a_r f(b_r) - b_r f(a_r)}{f(b_r) - f(a_r)}$. If $f(x_r) = 0$ or $|\dfrac{x_r - x_{r-1}}{x_r}| < \varepsilon$ then stop and take x_r.
3. If

   $f(a_r)\,f(x_r) < 0$: set $a_{r+1} = a_r$ and $b_{r+1} = x_r$.
   $f(b_r)\,f(x_r) < 0$: set $b_{r+1} = b_r$ and $a_{r+1} = x_r$.
4. Increase r by one and proceed with step 2.

Newton-Raphson Method

A Brief history of the Newton-Raphson method is given by Cajori (1911) who states, that the first printed version of the ideas of Newton for finding the root of a polynomial is in a book by Wallis (1685, § 94). Following Cajori (1911) theese first thoughts of Newton

were then extended by Raphson (1690) who gave the *Newton-Raphson* method the present form.

Instead of having to define an interval in which the root has to be found, the Newton-Raphson method only needs to have a single value x_0 near enough to the root. The term *near enough* is dependent on the function.

The Newton-Raphson method can be derived using a taylor expanion around $\mathrm{f}(x_r)$

$$\mathrm{f}(x) = \mathrm{f}(x_r) + (x - x_r)\,\mathrm{f}'(x_r) + E_1 \tag{A.9}$$

as x is a root of f, obviously $\mathrm{f}(x) = 0$ and thus by solving the equation (A.9) for x one obtains

$$x + \frac{E_1}{\mathrm{f}'(x_r)} = x_r - \frac{\mathrm{f}(x_r)}{\mathrm{f}'(x_r)} \quad . \tag{A.10}$$

The smaller the approximation error E_1 of the taylor series expansion is in relation to the tangent of f at point x_r, the nearear will the term $x + \frac{E_1}{\mathrm{f}'(x_r)}$ be to x. Thus it is sensical to set $x_{r+1} := x + \frac{E_1}{\mathrm{f}'(x_r)}$ as next approximation step. This yields the following algorithm.

Algorithm A.5 Newton-Raphson Method

Let f be a continous real valued function with a root. Set $r = 1$ and $\varepsilon > 0$ arbitrarily small depending on the precision needed. Obtain the derivative f' of f.

1. Find x_0 near the root of f. E.g. by the Bisection or Regula Falsi methods.
2. Set $x_r = x_{r-1} - \dfrac{\mathrm{f}(x_{r-1})}{\mathrm{f}'(x_{r-1})}$. If $\mathrm{f}(x_r) = 0$ or $\left|\dfrac{x_r - x_{r-1}}{x_r}\right| < \varepsilon$ then stop and take x_r.
3. Increase r by one and proceed with step 2.

Maximizing a Likelihood

One of the most used classes of estimation methods are the maximum-likelihood estimators. The main idea behind this method is rather simple. Assuming a distributional relationship between a set of variables one derives a function called likelihood function. This function ℓ maps from $\mathscr{R}^p \mapsto \mathscr{R}$ where p is the number of parameters needed to describe the before mentioned relationship. The function is constructed in such way, that it

provides the likelihood of the parameter set $\beta_1, .., \beta_p$ given the distributional assumptions and the data.

In order to find the parameter set β for which the function ℓ is maximal (i.e., the parameters with the *maximal likelihood*) one can also search for the root of the first derivative ℓ' of ℓ. For numerical reasons this is generally accomplished by searching for the root of $-\ell'$. This approach may be problematic if the likelihood function at hand is multimodal, as several roots of ℓ' exist in that case. As described before, the Newton-Raphson method provides one possibility in order to find the root of this function.

Newton-Raphson Method for Maximum-Likelihood Estimation

In order to apply the Newton-Raphson method to the maximization of a likelihood ℓ, it is necessary to attain its first and second derivative (ℓ' and ℓ''). Starting with an initial guess x_0 and setting in algorithm A.5 f $:= \ell'$ and f$' := \ell''$ one obtains the Newton-Raphson algorithm for maximum likelihood estimation. However, in general the function ℓ maps from $\mathbb{R}^p \mapsto \mathbb{R}$. Therefore, ℓ' is the Jacobian of ℓ and ℓ'' is the Hesse matrix of ℓ.

If the first and second derivatives of ℓ have no analytical solution, then one can also use numerical derivatives.

Fisher-Scoring Method for Maximum-Likelihood Estimation

The Fisher-Scoring method is a modification of the Newton-Raphson, where instead of using the second derivative of ℓ the *Fisher information* is used (Bailey, 1961; Fisher, 1925; Kale, 1961, 1962). The Fisher information can be interpreted on the one hand as the variance of the likelihood function, and on the other hand as the expectation of the second derivative of the likelihood function. When using a canonical link with a exponential family distribution the observed and expected Hesse matrices are identical (Fahrmeir, Tutz, & Hennevogl, 1994, p. 39), and hence, the Newton-Raphson and Fisher-Scoring algorithms give identical results.

Appendix B

Additional Graphs for Chapter 4

B.1 Additional Graphs for Latin Hypercube Sampling

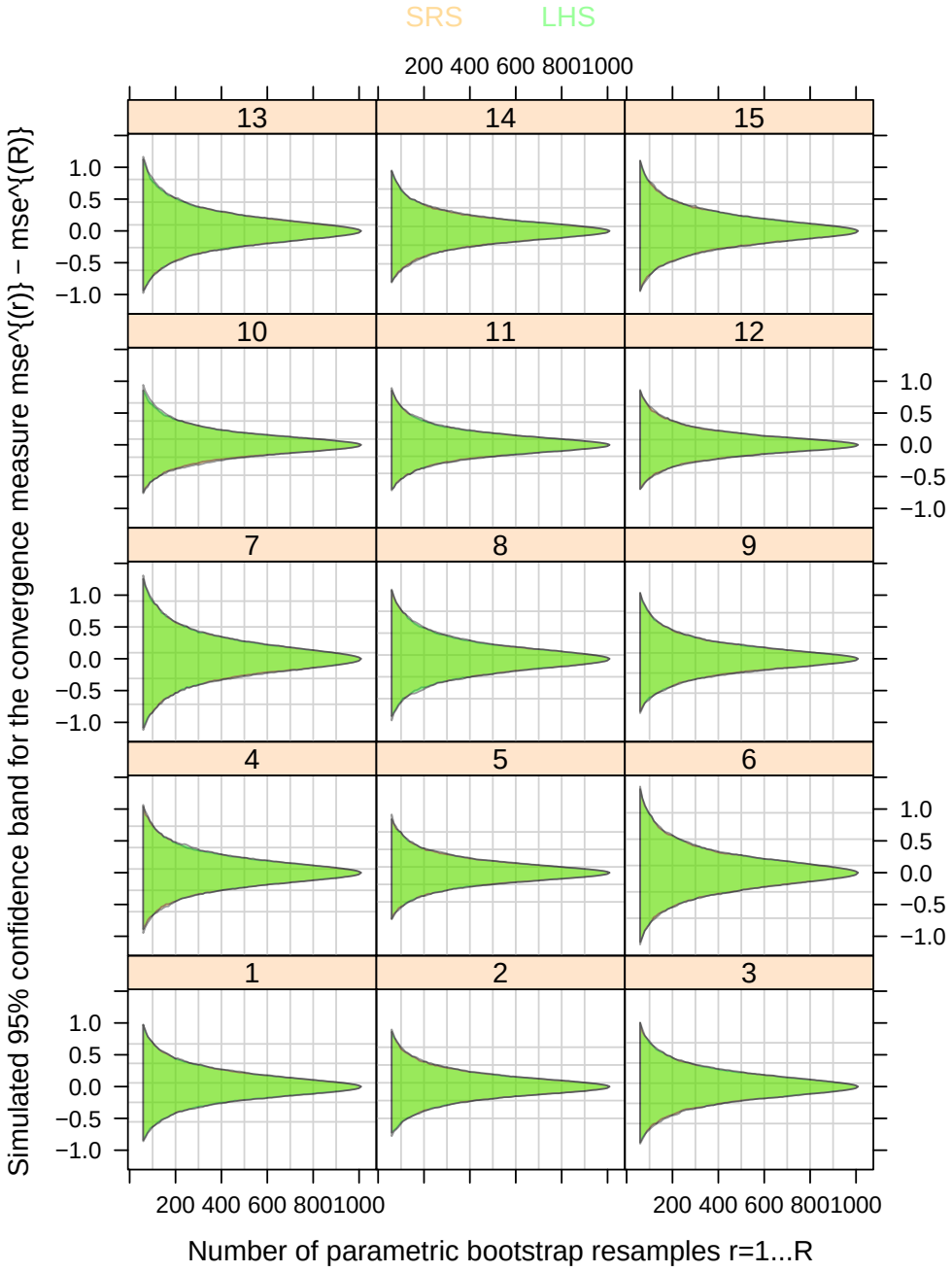

Figure B.1: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 2

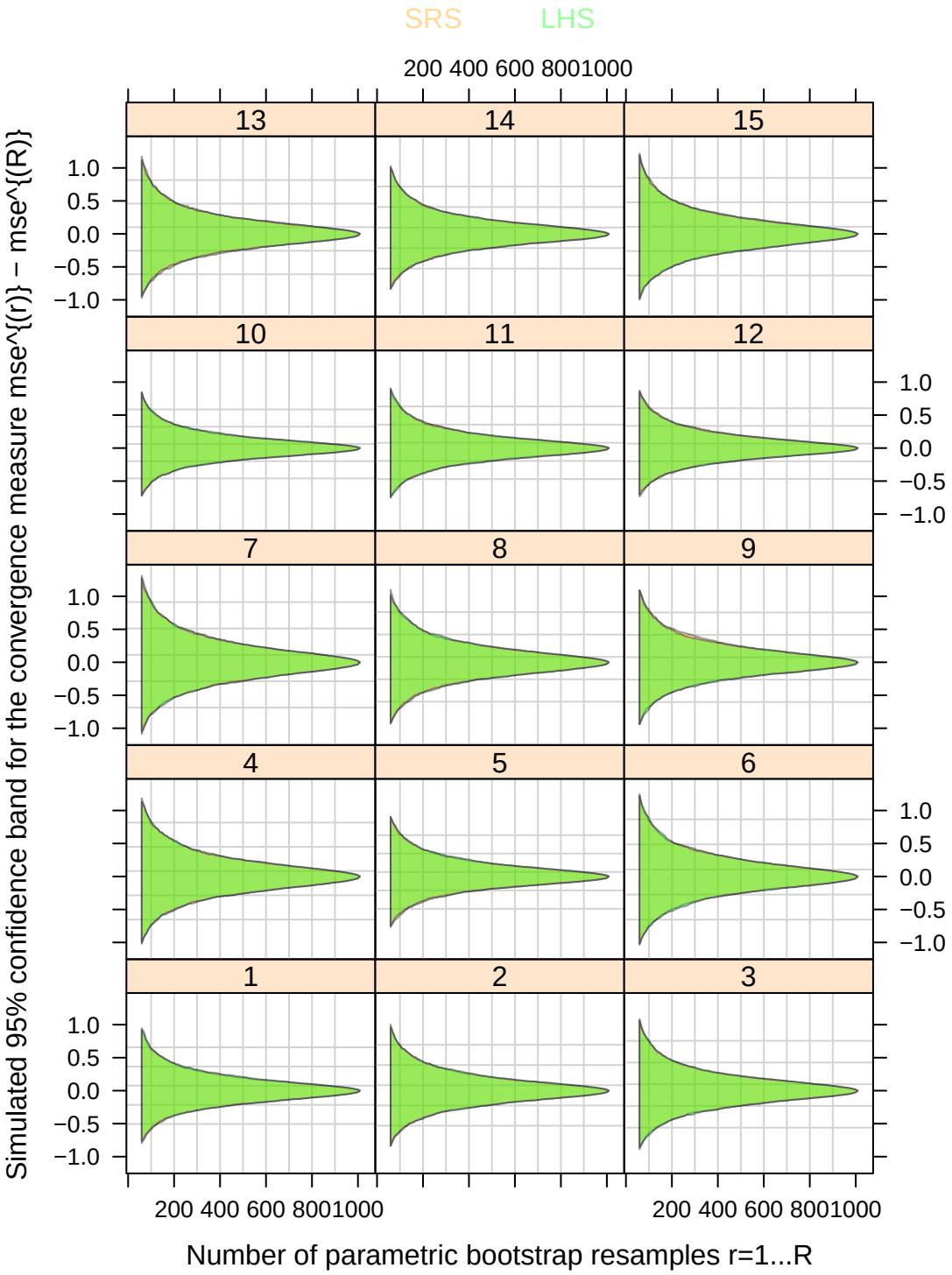

Figure B.2: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 3

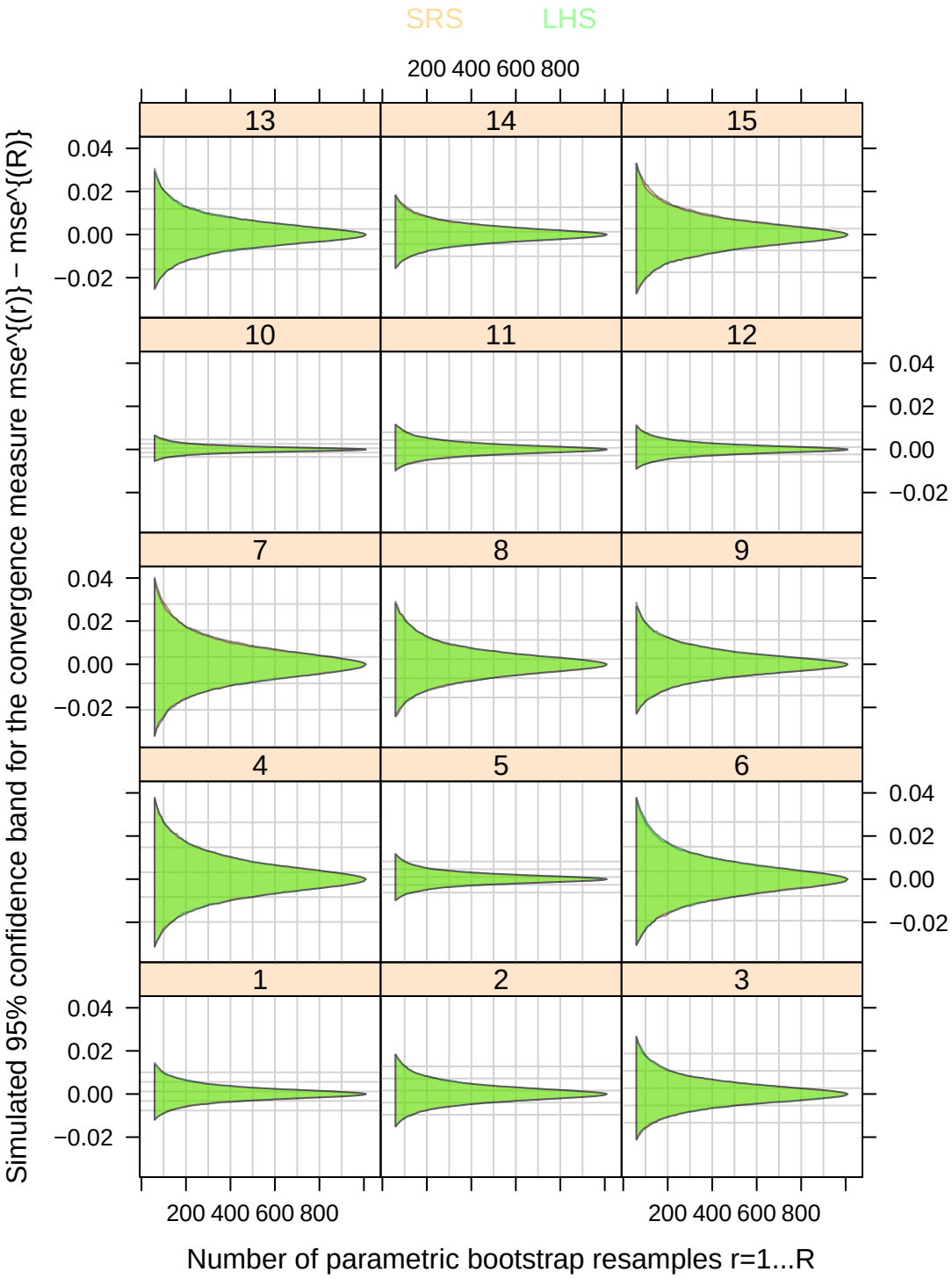

Figure B.3: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 4

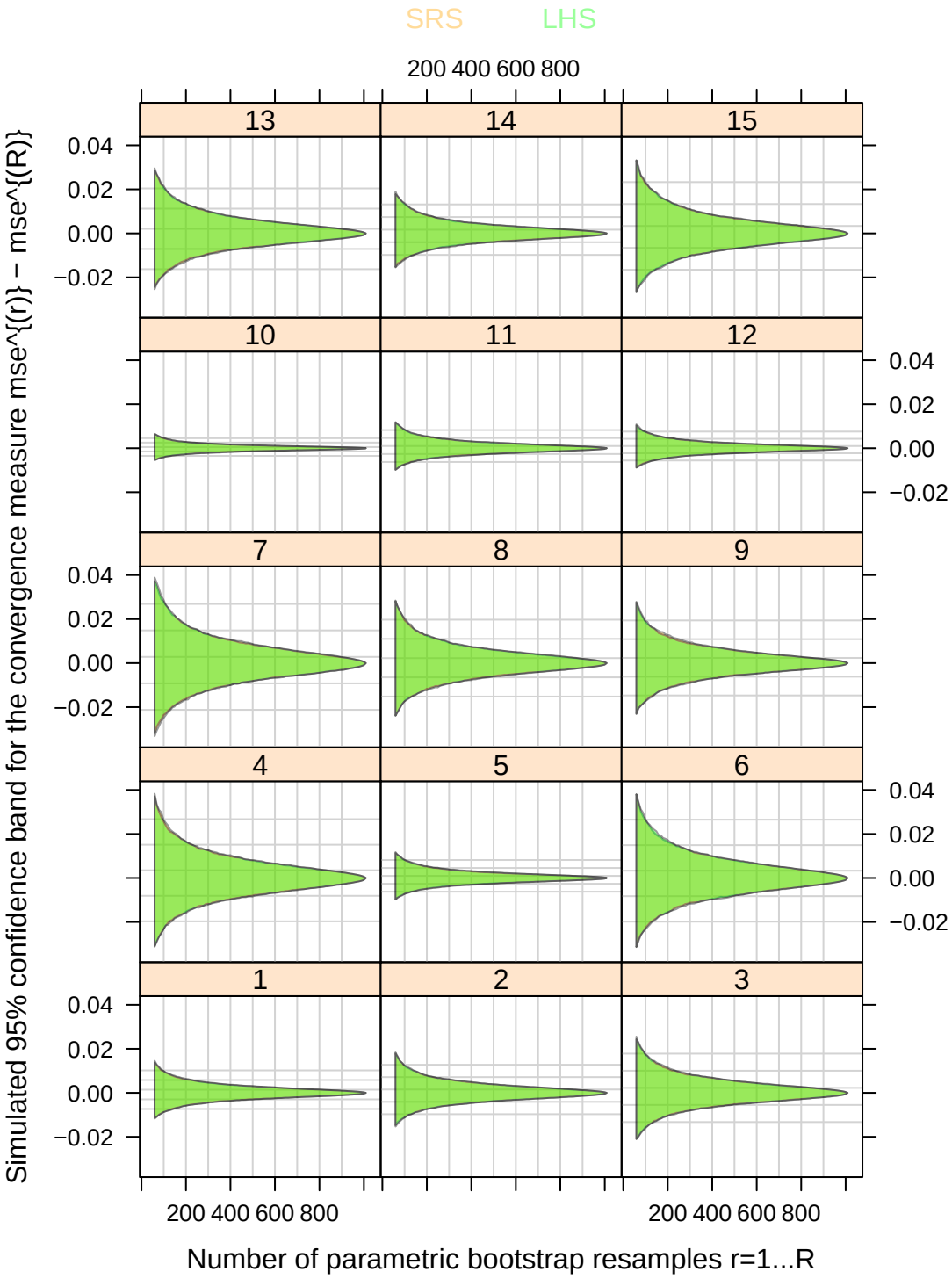

Figure B.4: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 5

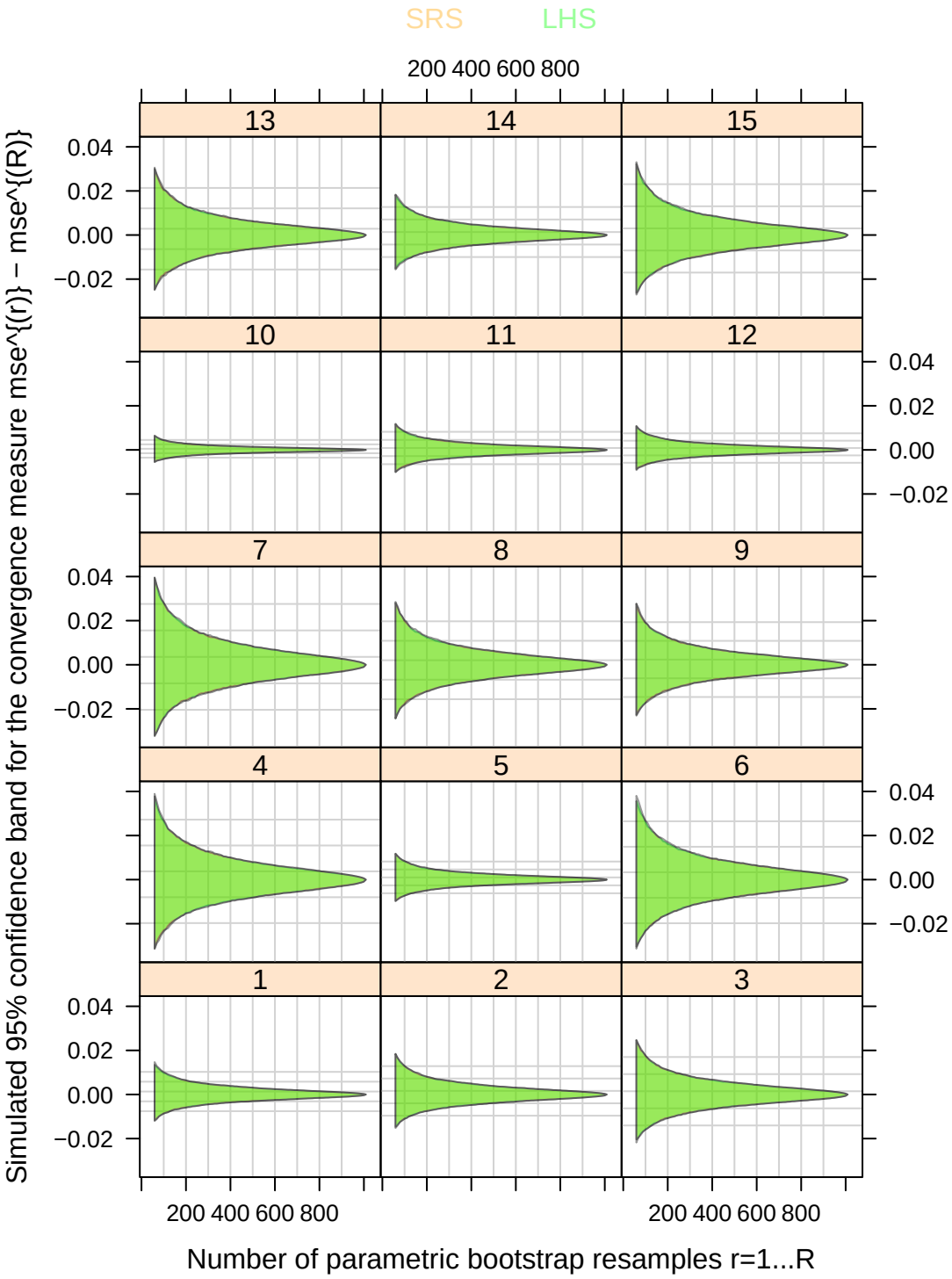

Figure B.5: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 6

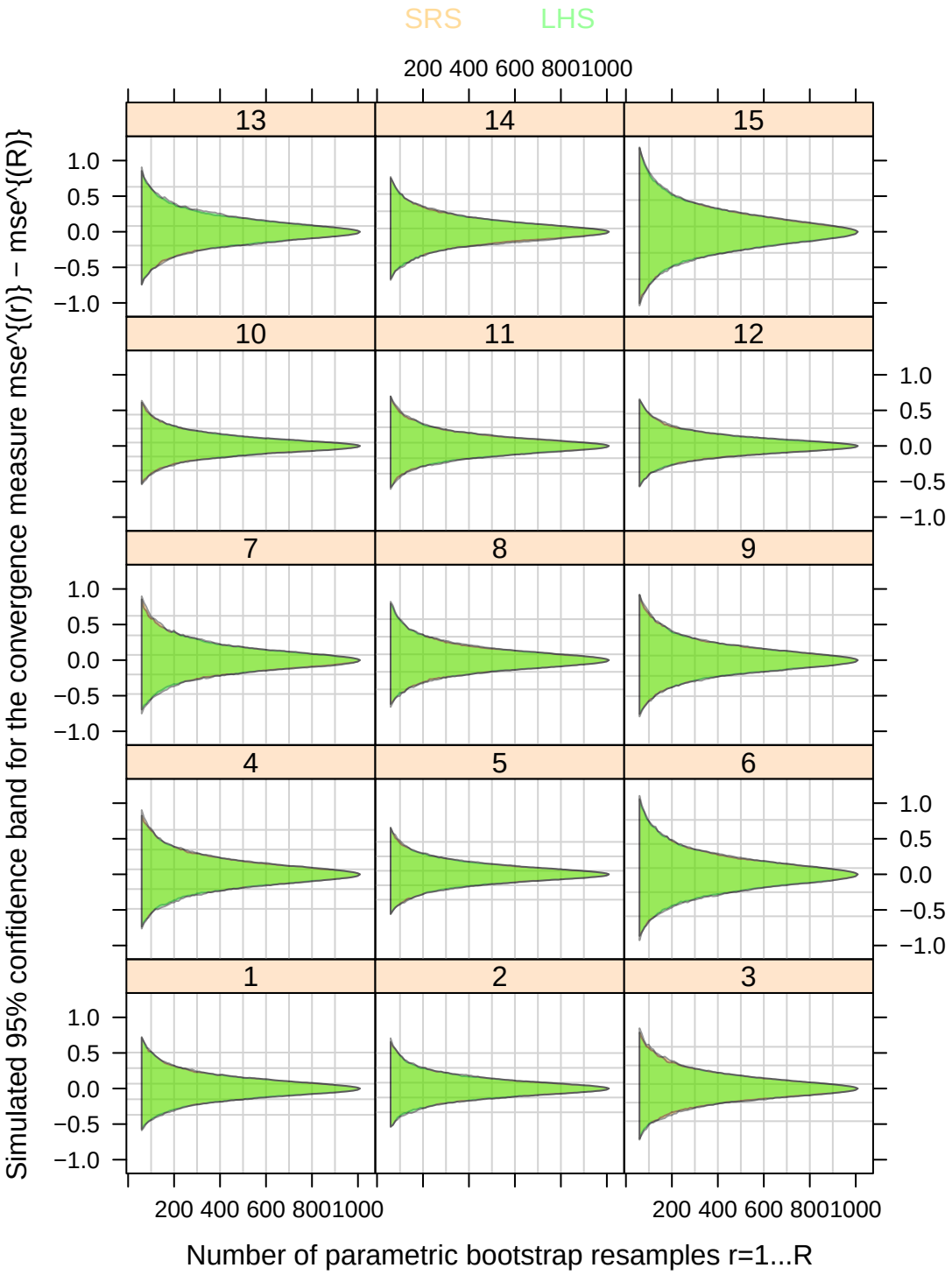

Figure B.6: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 7

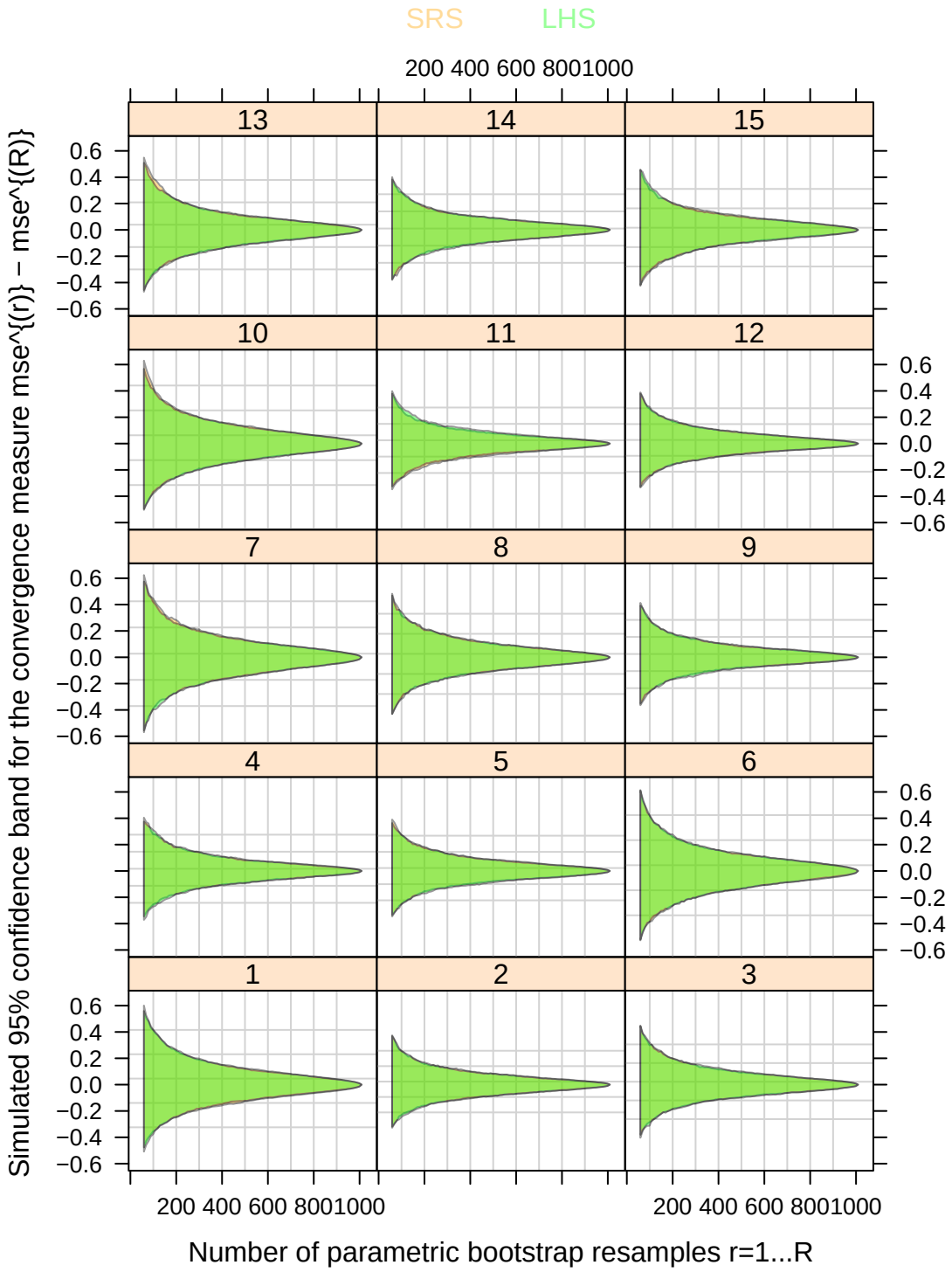

Figure B.7: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 8

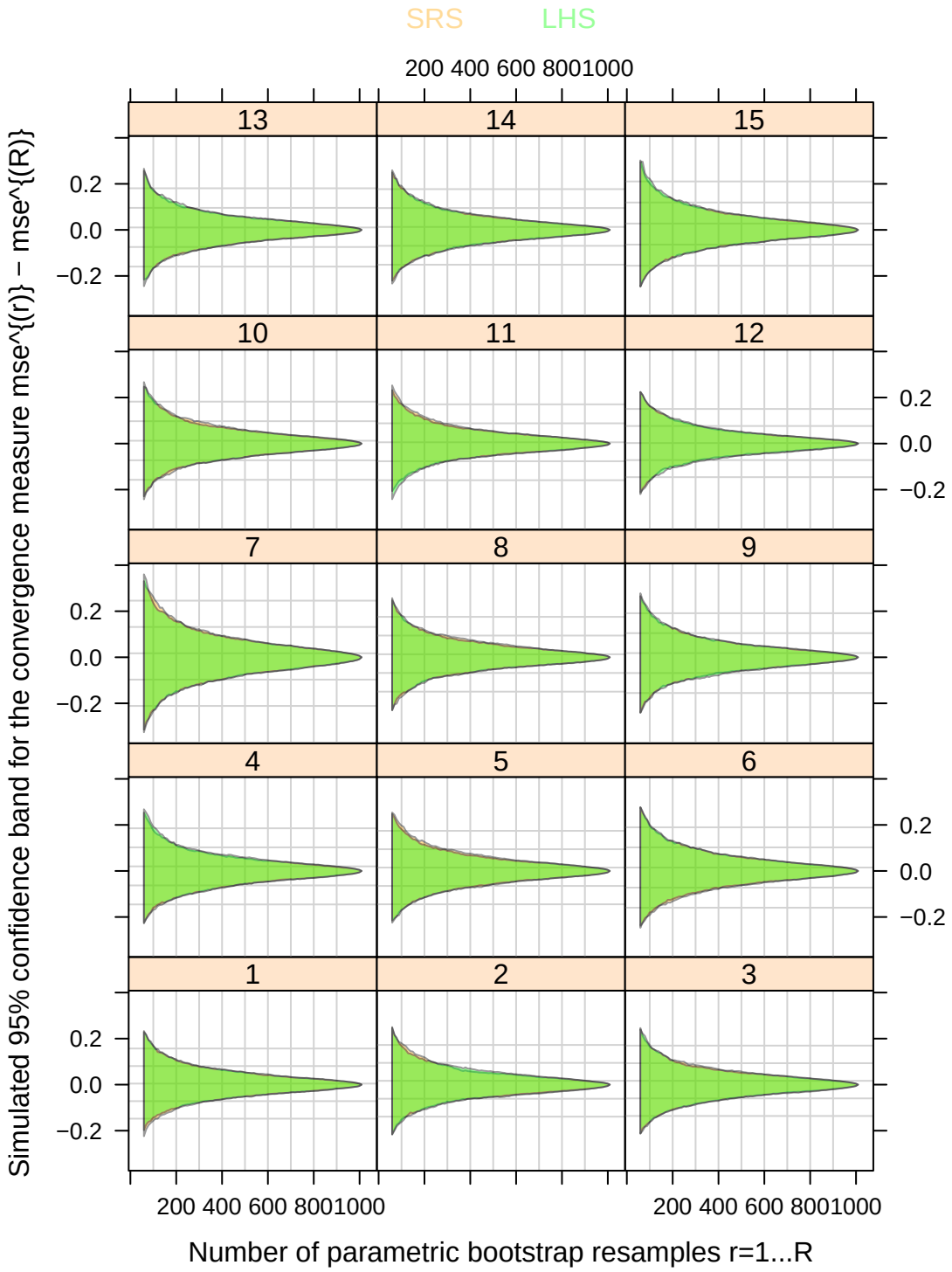

Figure B.8: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 9

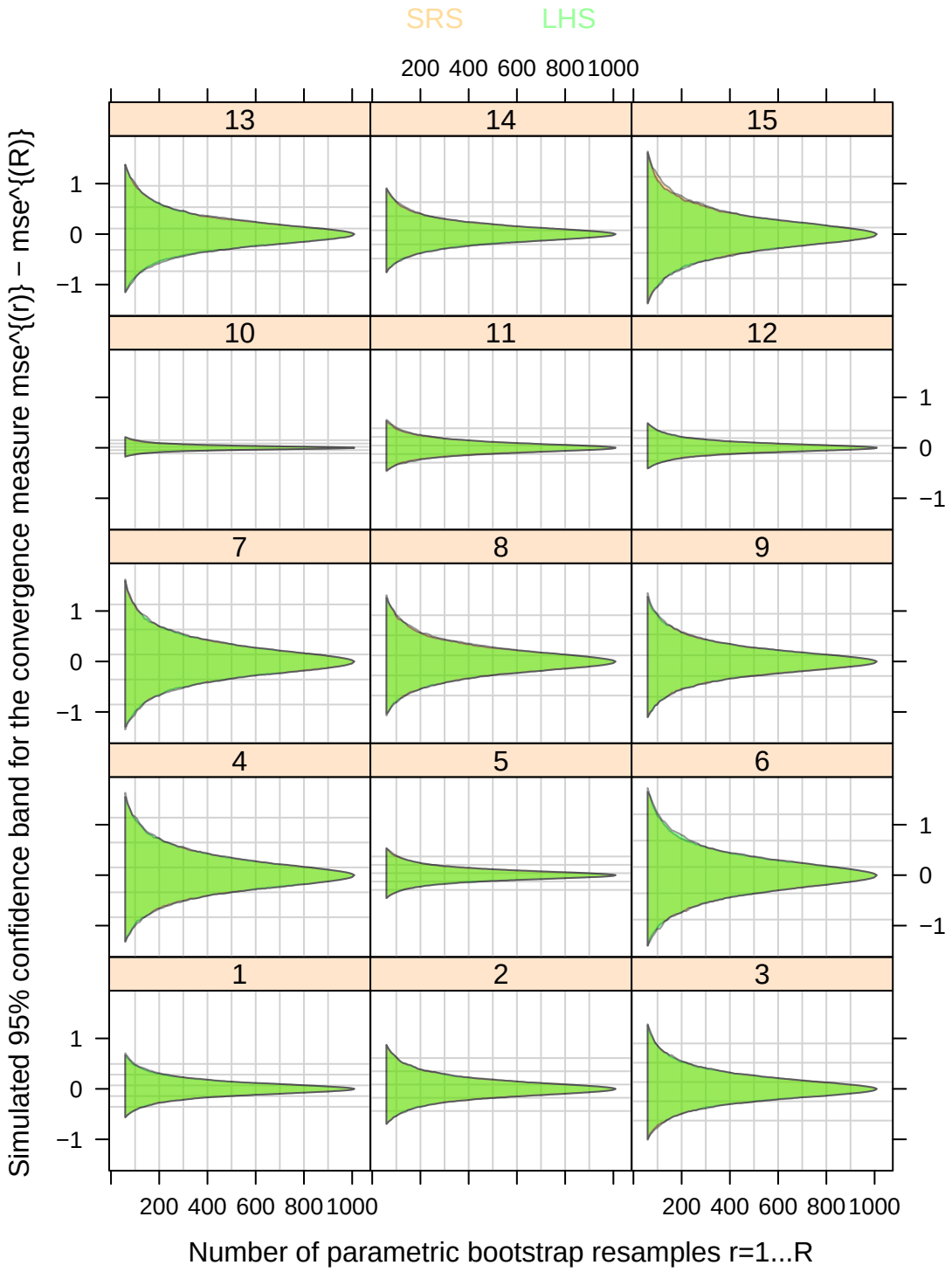

Figure B.9: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 10

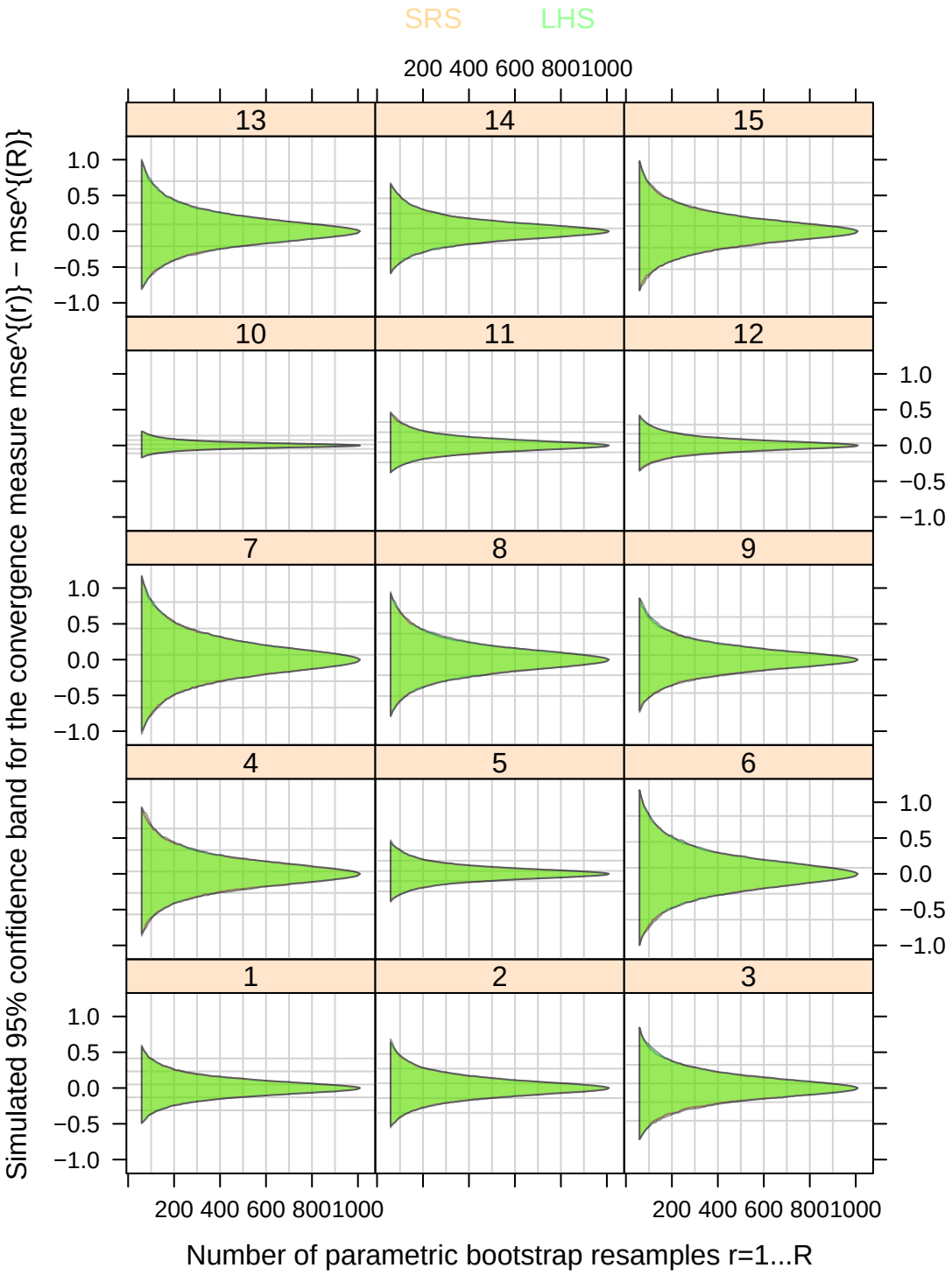

Figure B.10: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 11

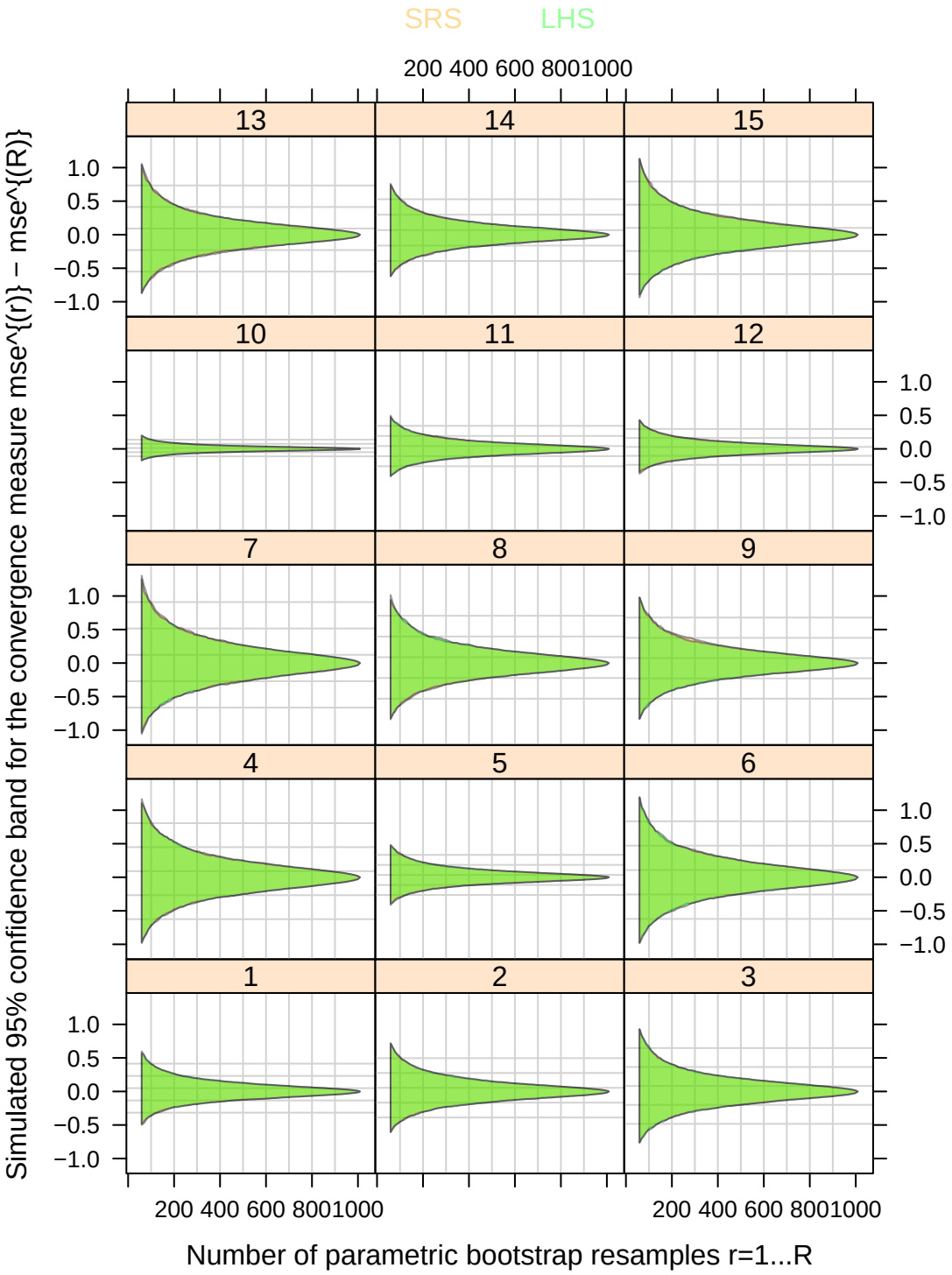

Figure B.11: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 12

B.2 Additional Graphs for Control Variate $g^{(2)}$

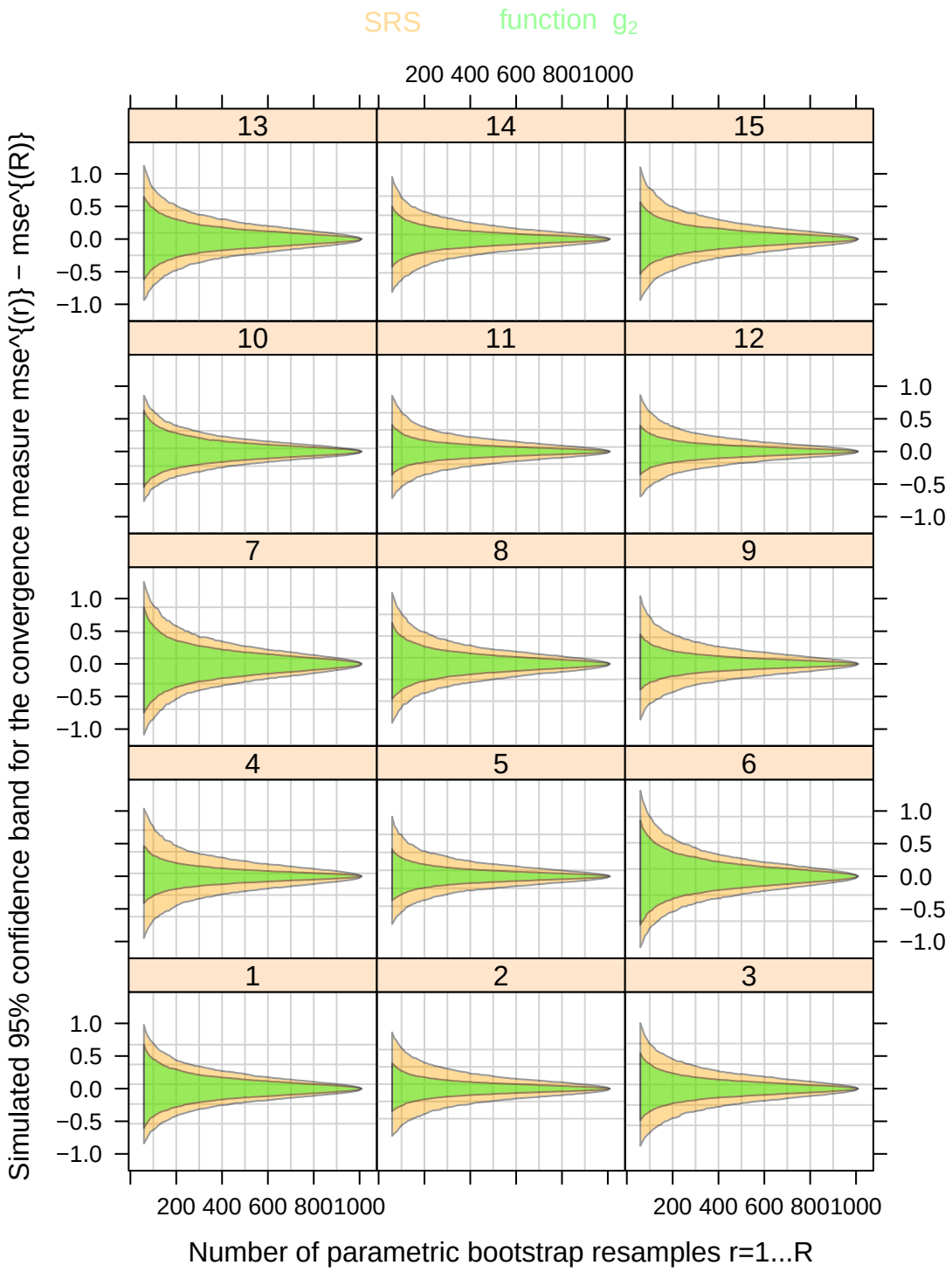

Figure B.12: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 2

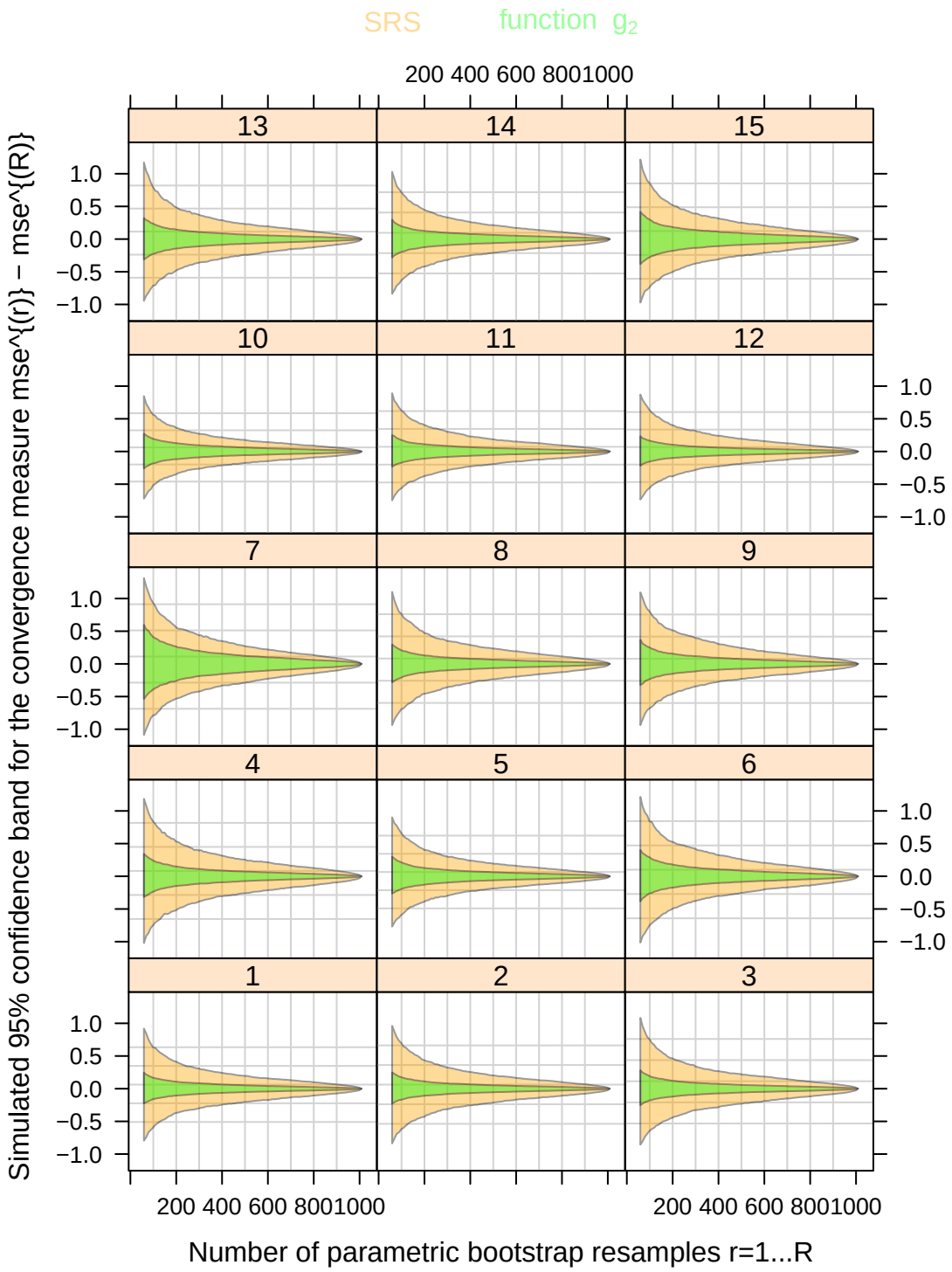

Figure B.13: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 3

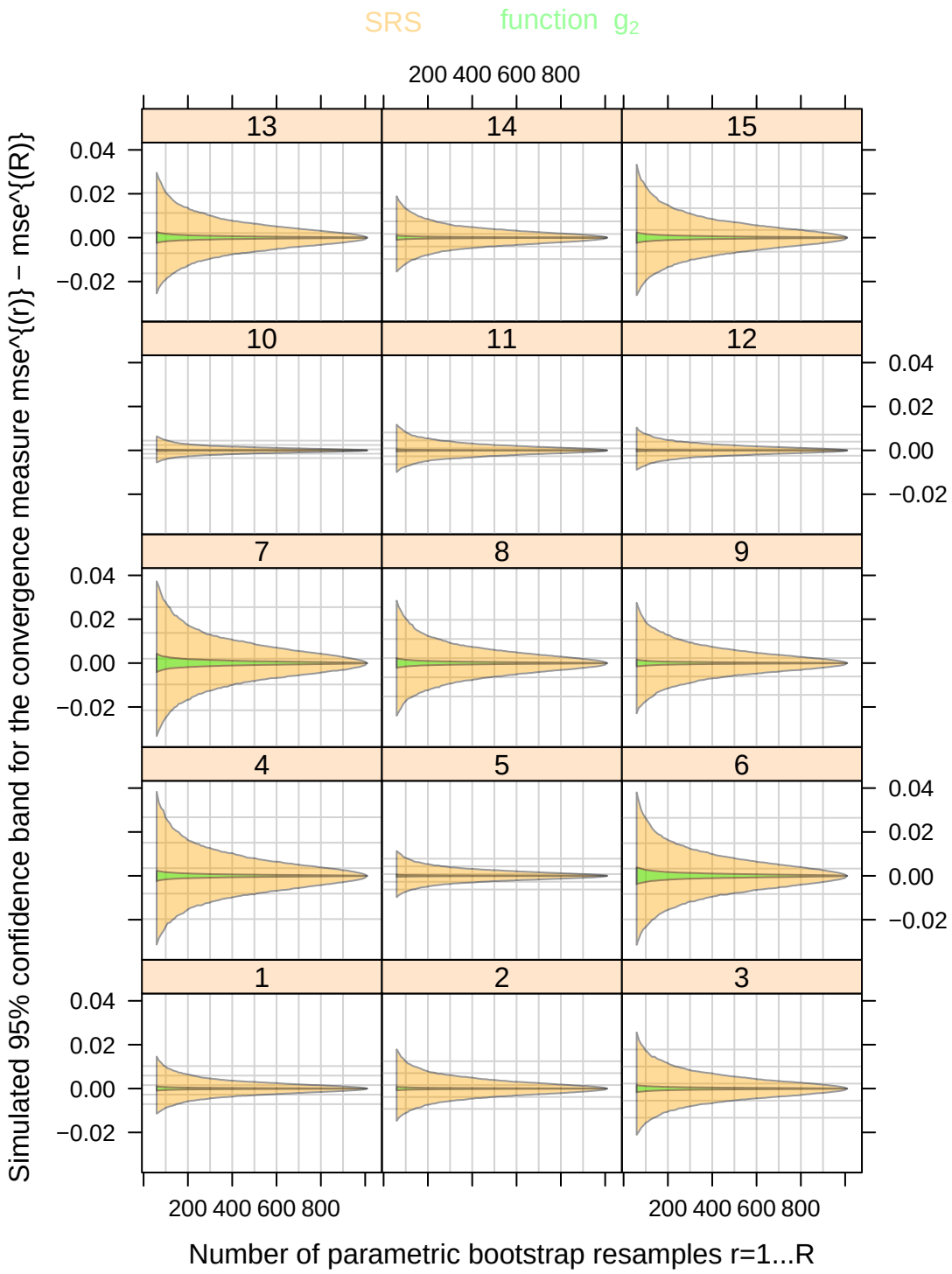

Figure B.14: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 5

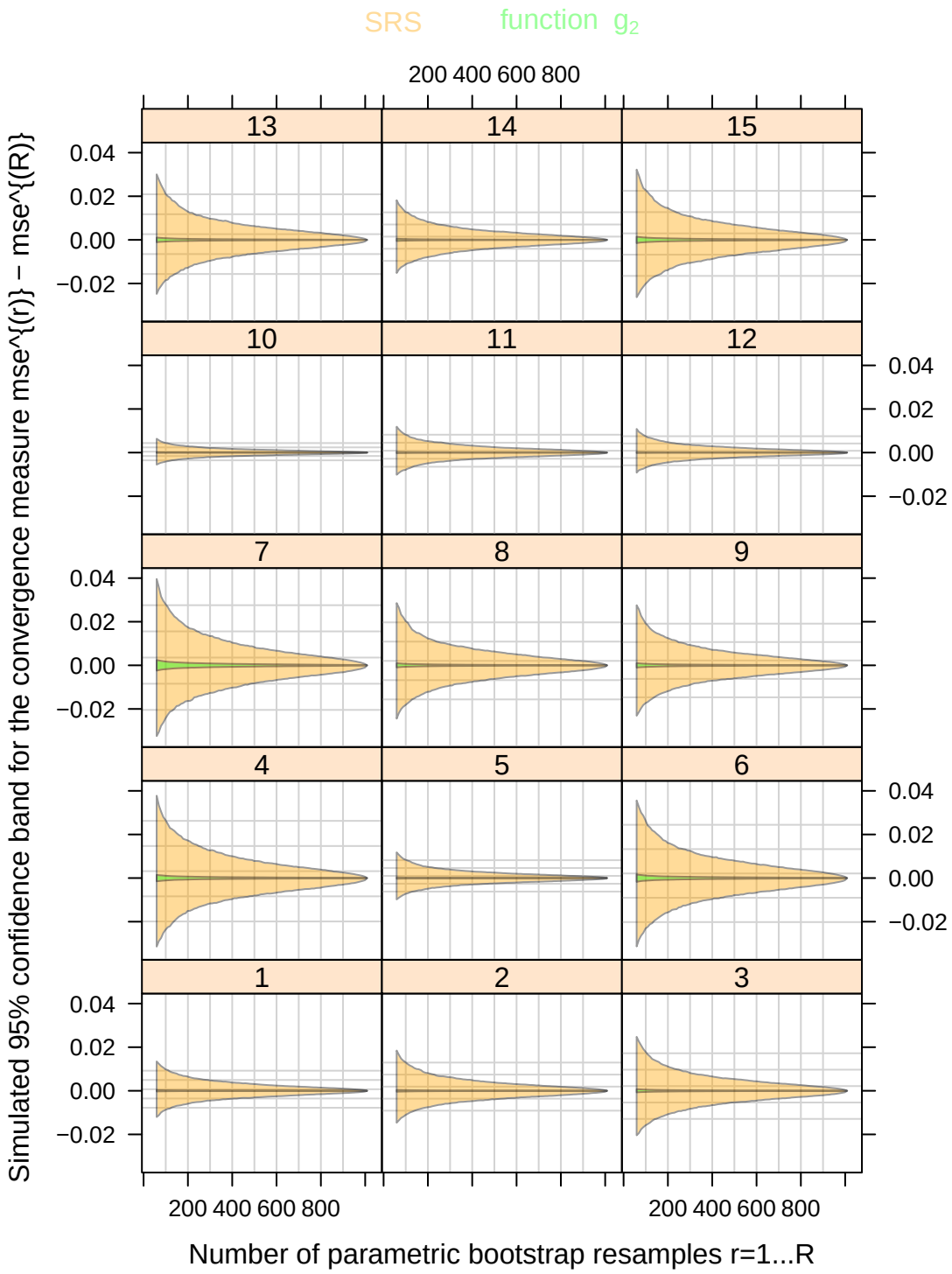

Figure B.15: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 6

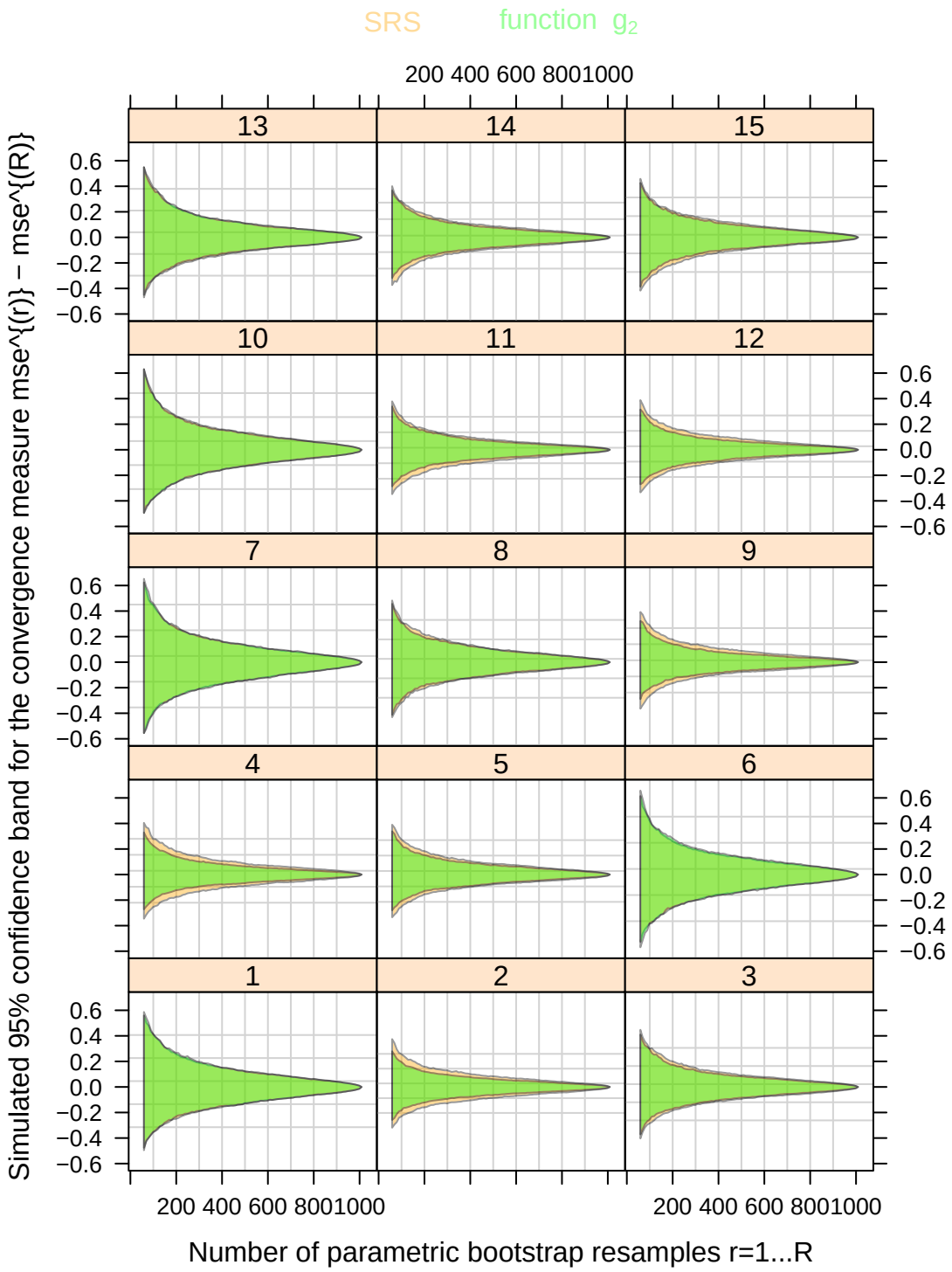

Figure B.16: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 8

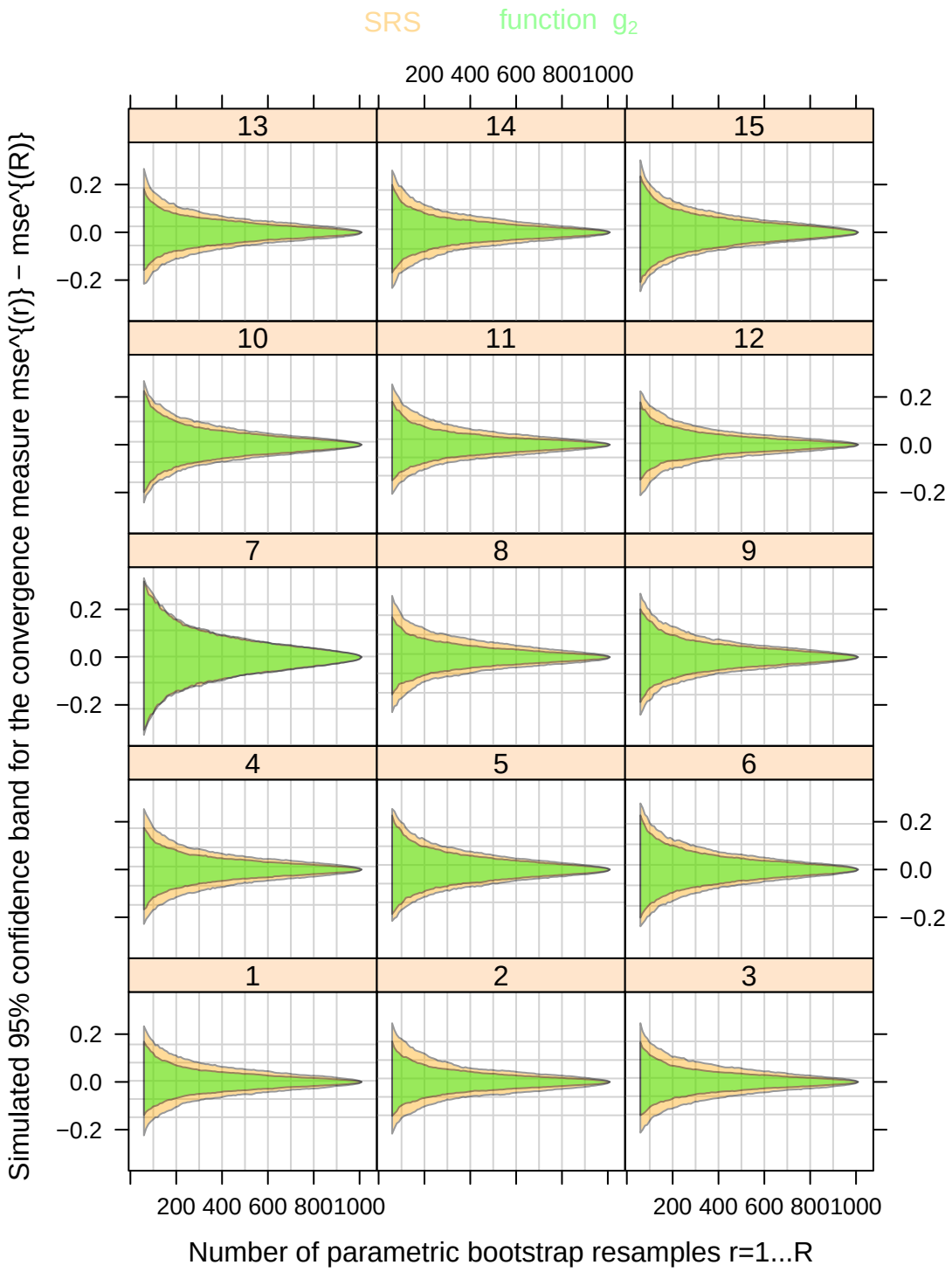

Figure B.17: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 9

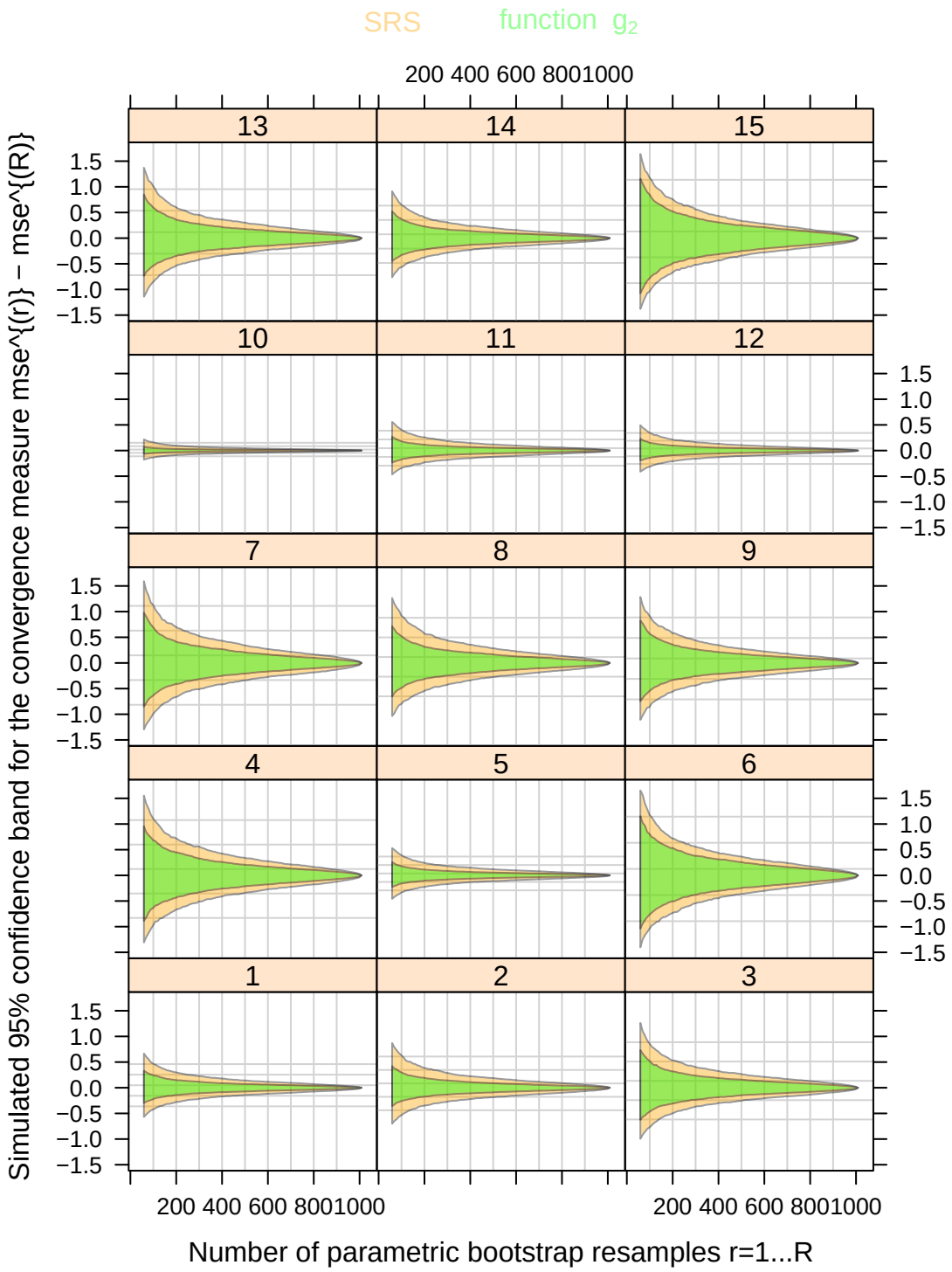

Figure B.18: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 10

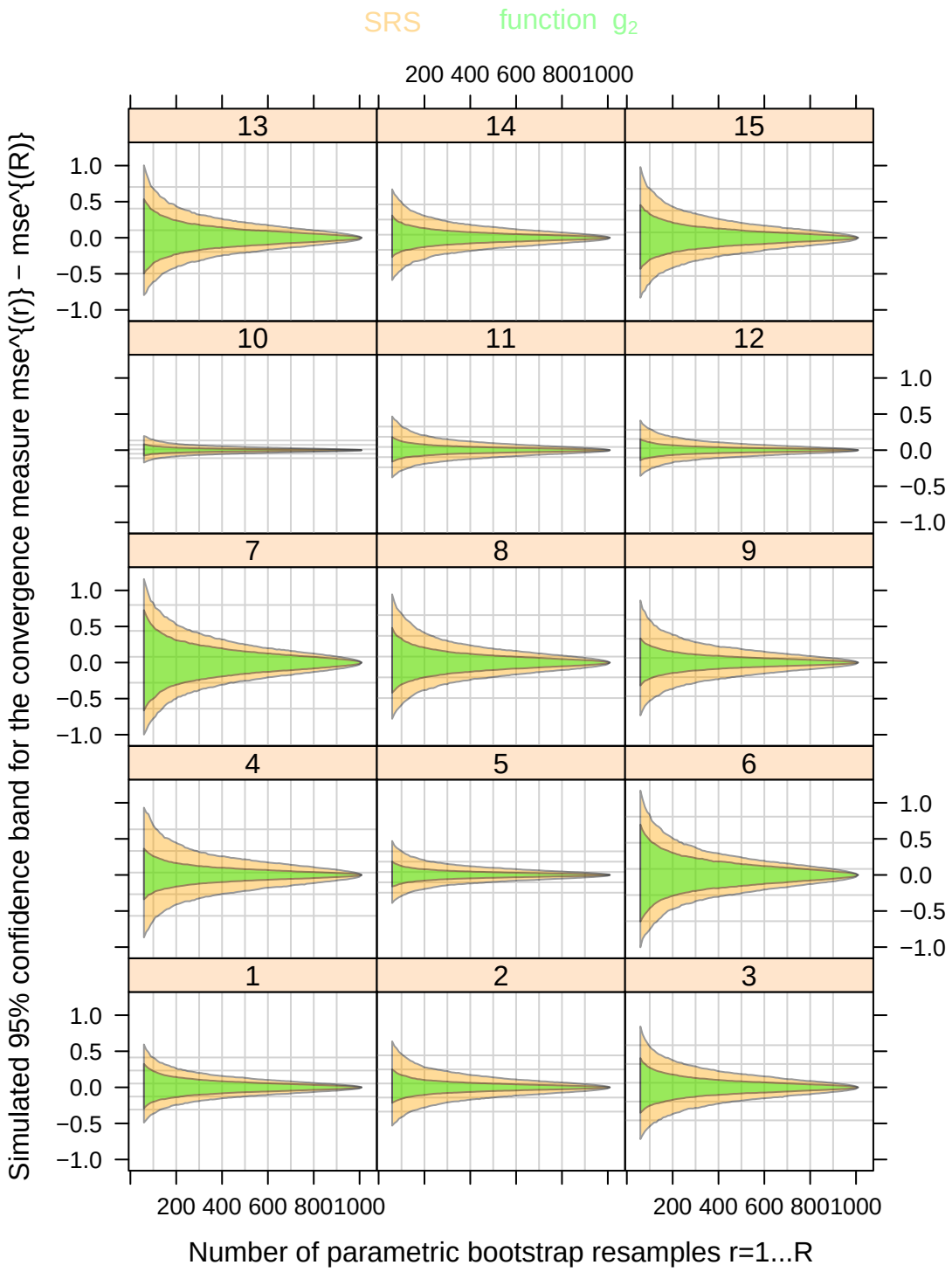

Figure B.19: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 11

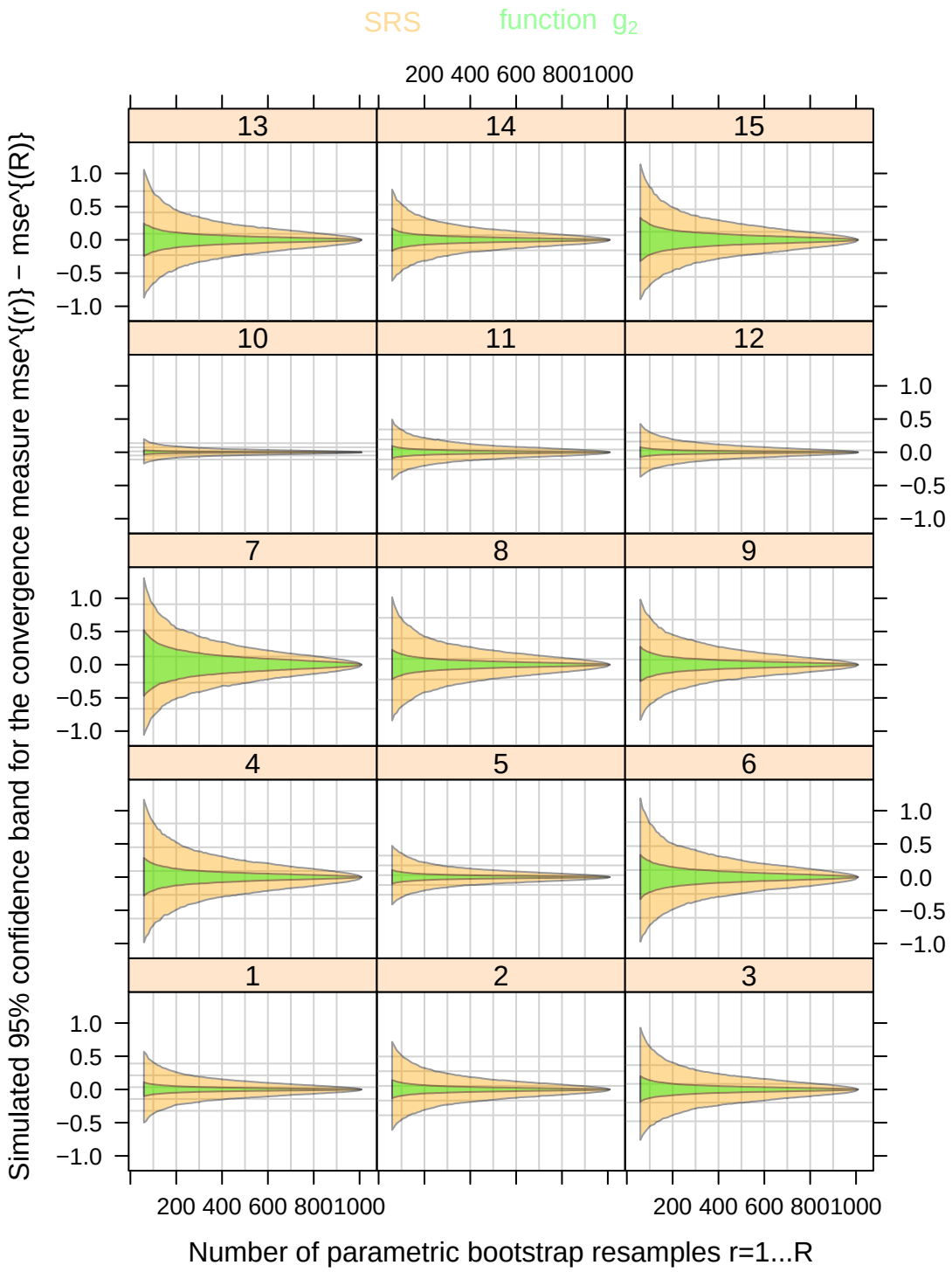

Figure B.20: Usingcontrol variate function $g^{(2)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 12

B.3 Additional Graphs for Control Variate $g^{(3)}$

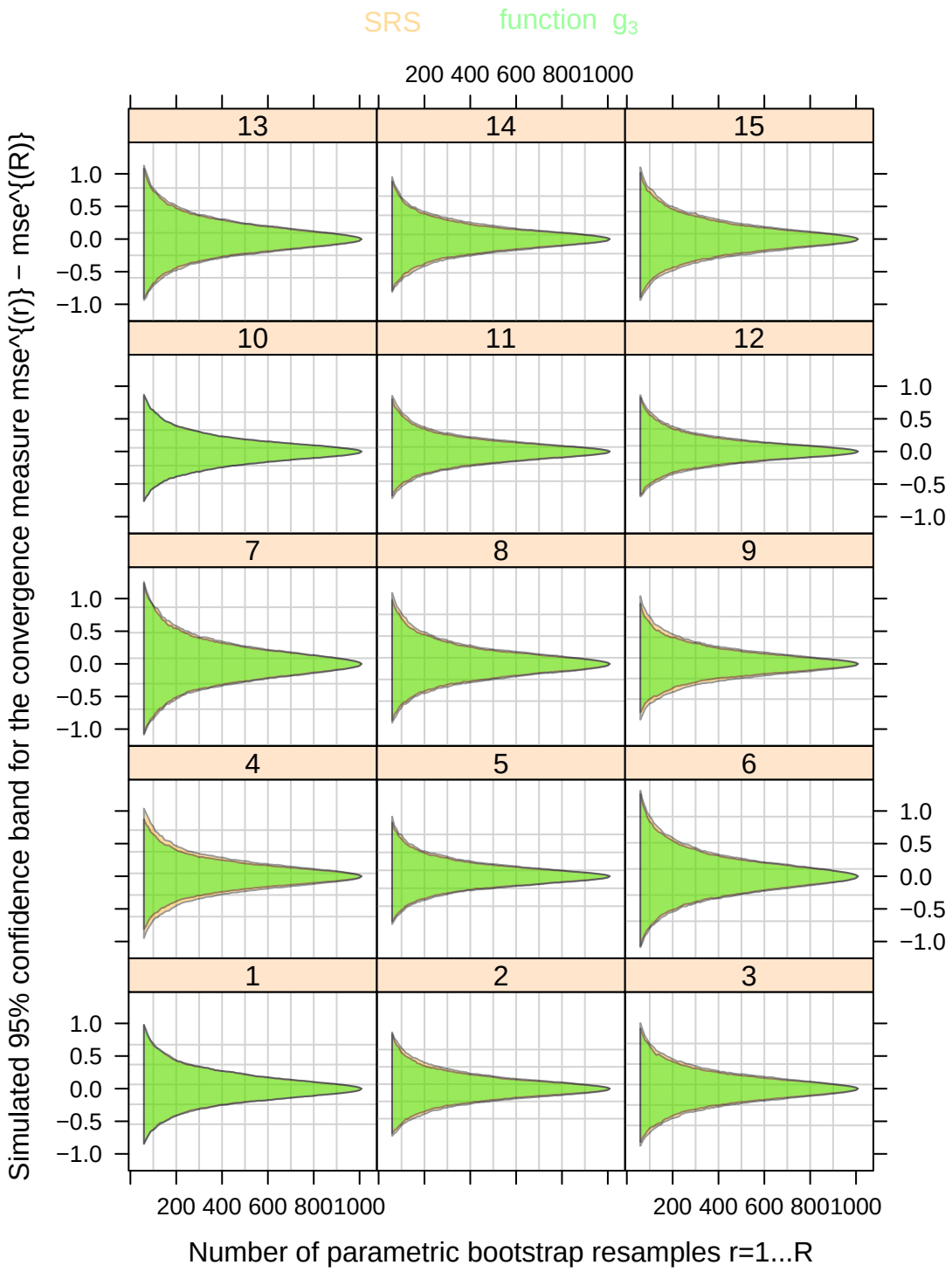

Figure B.21: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 2

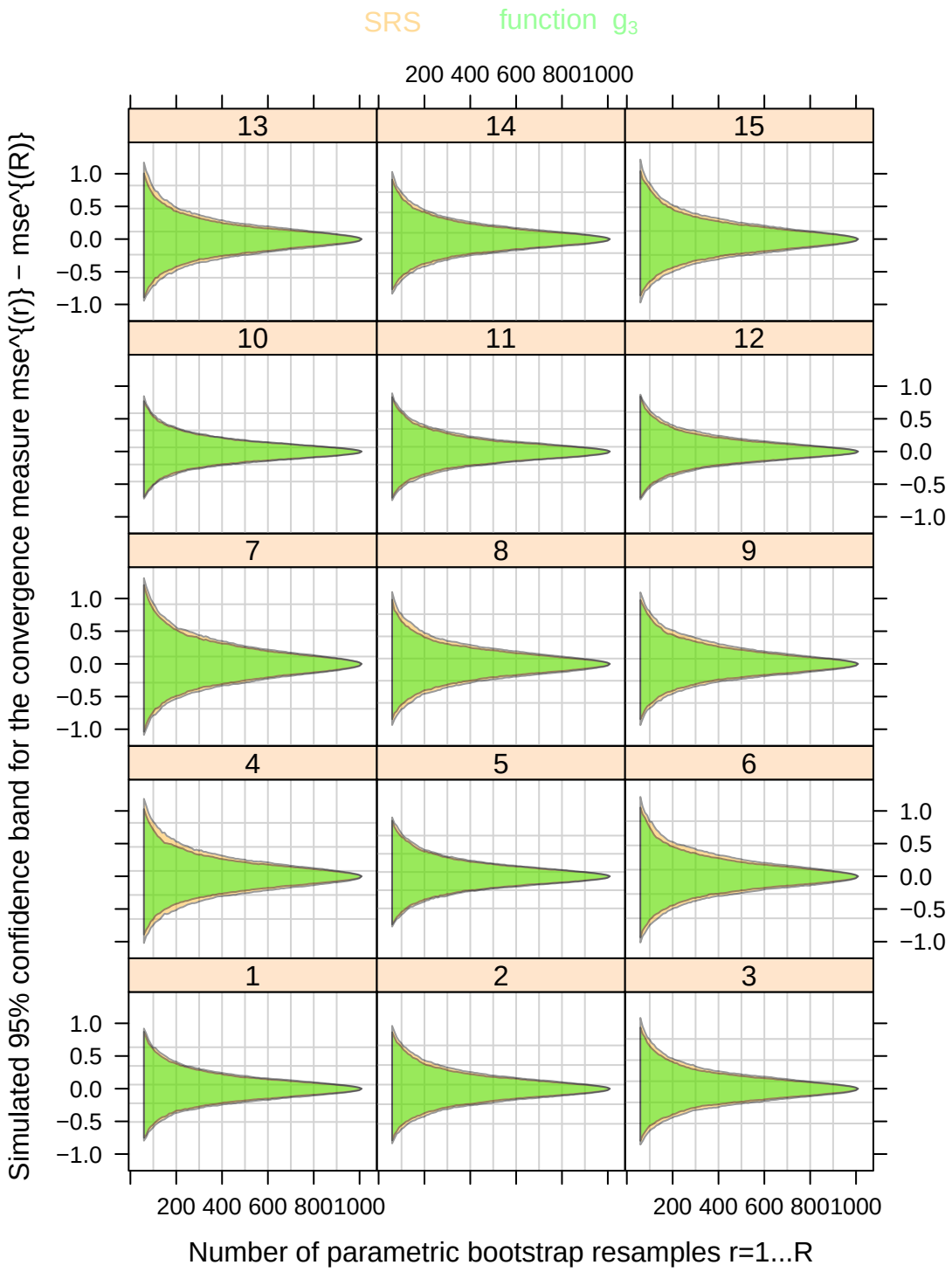

Figure B.22: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 3

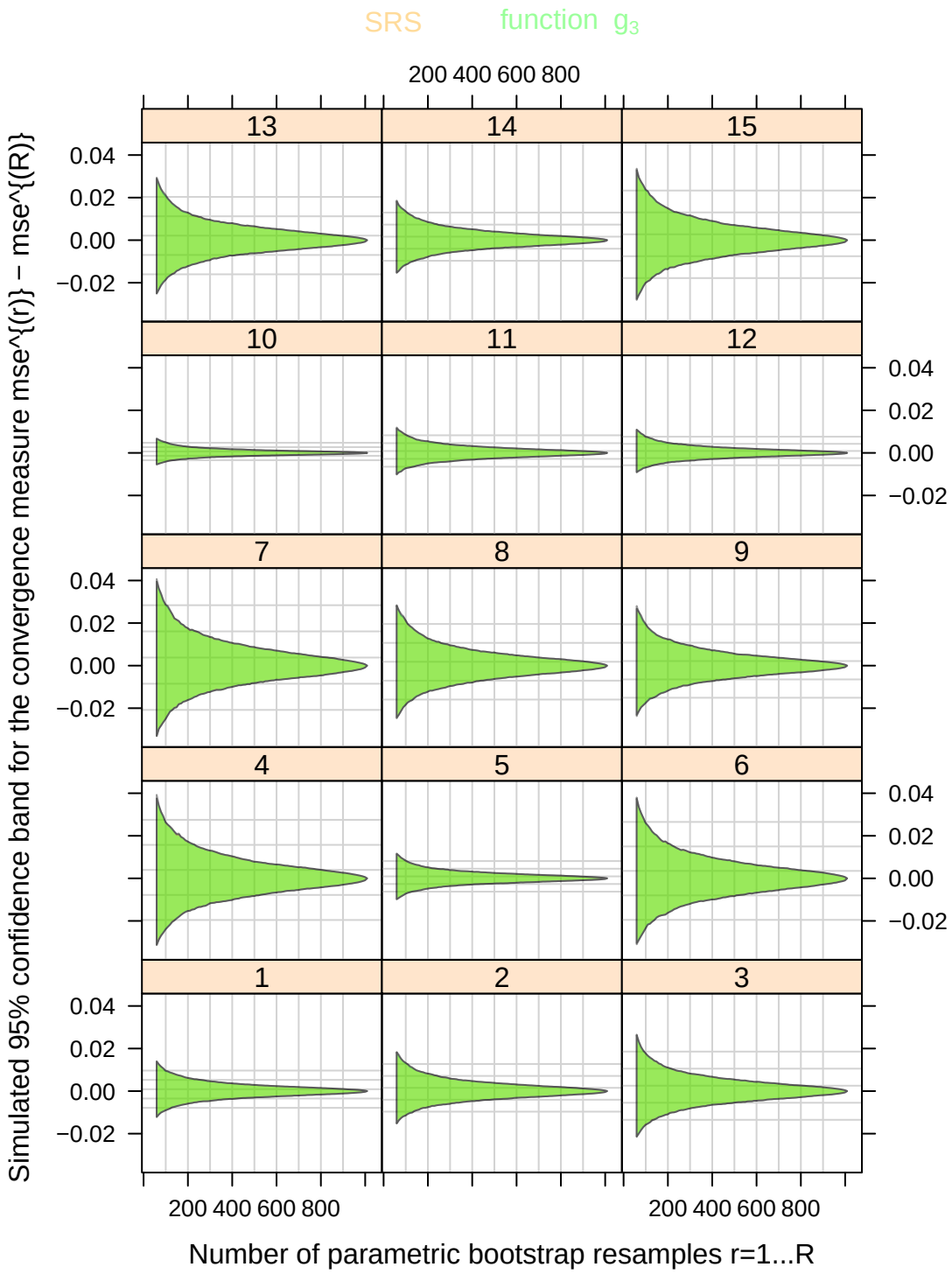

Figure B.23: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 4

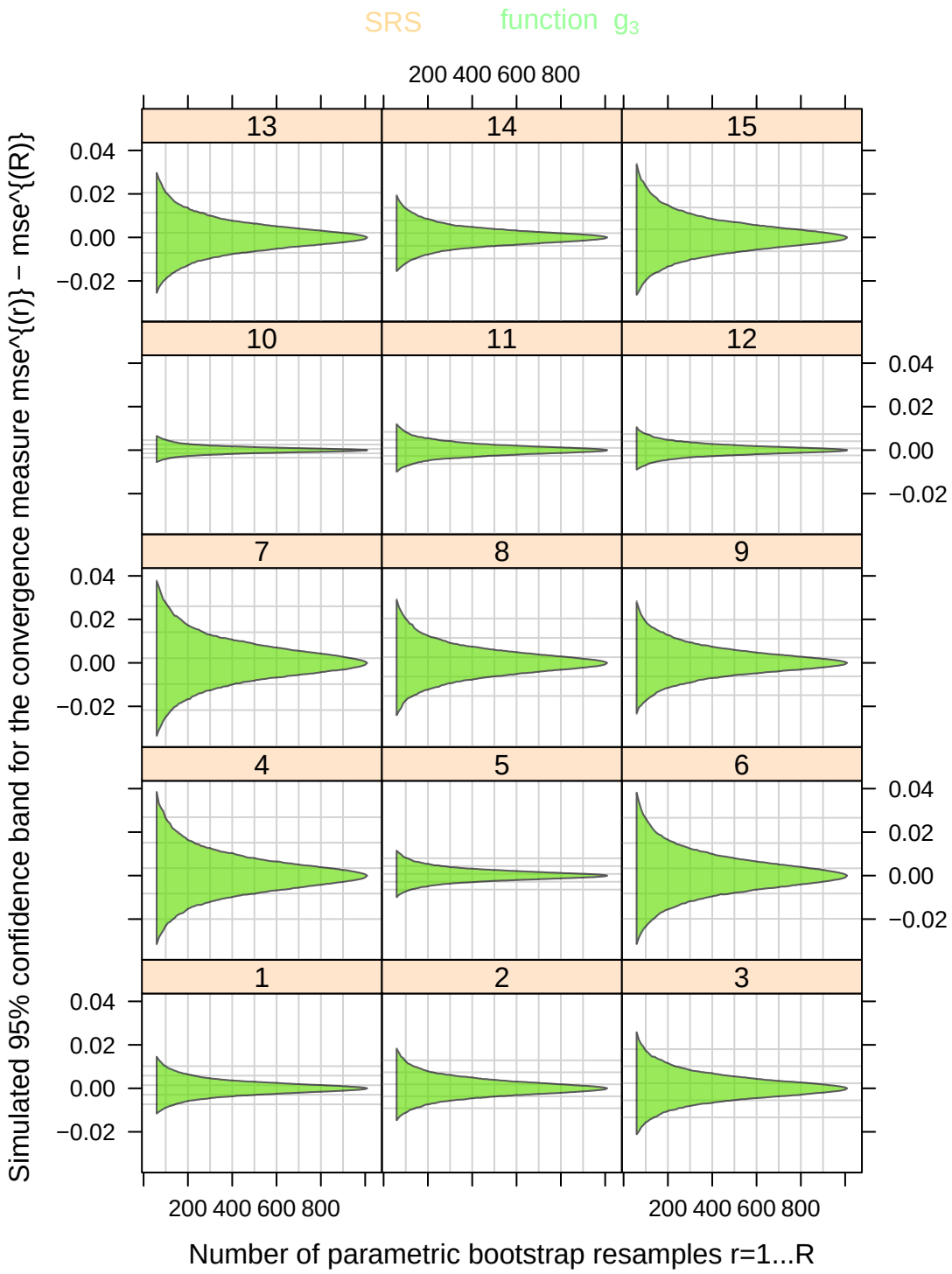

Figure B.24: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 5

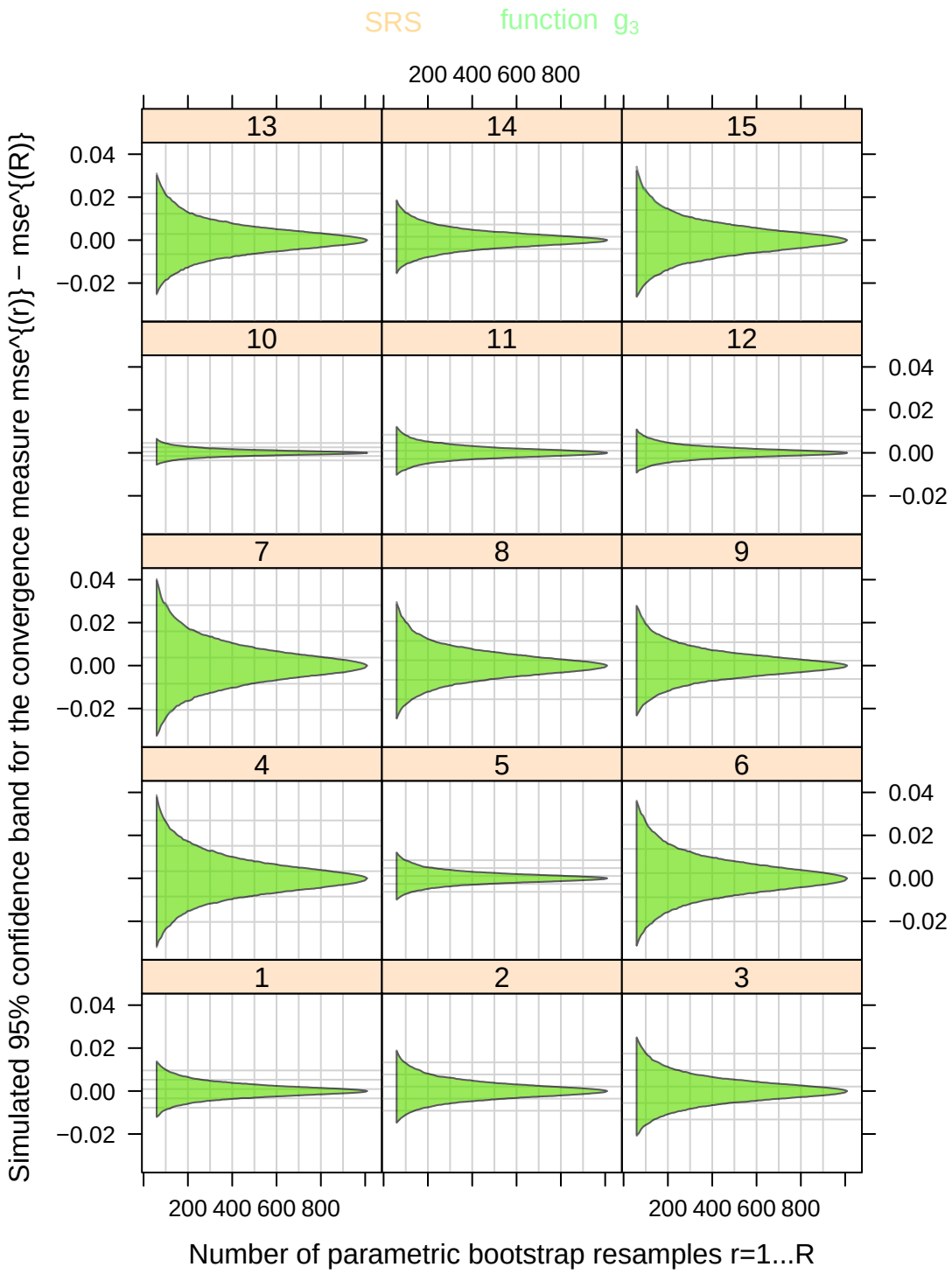

Figure B.25: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 6

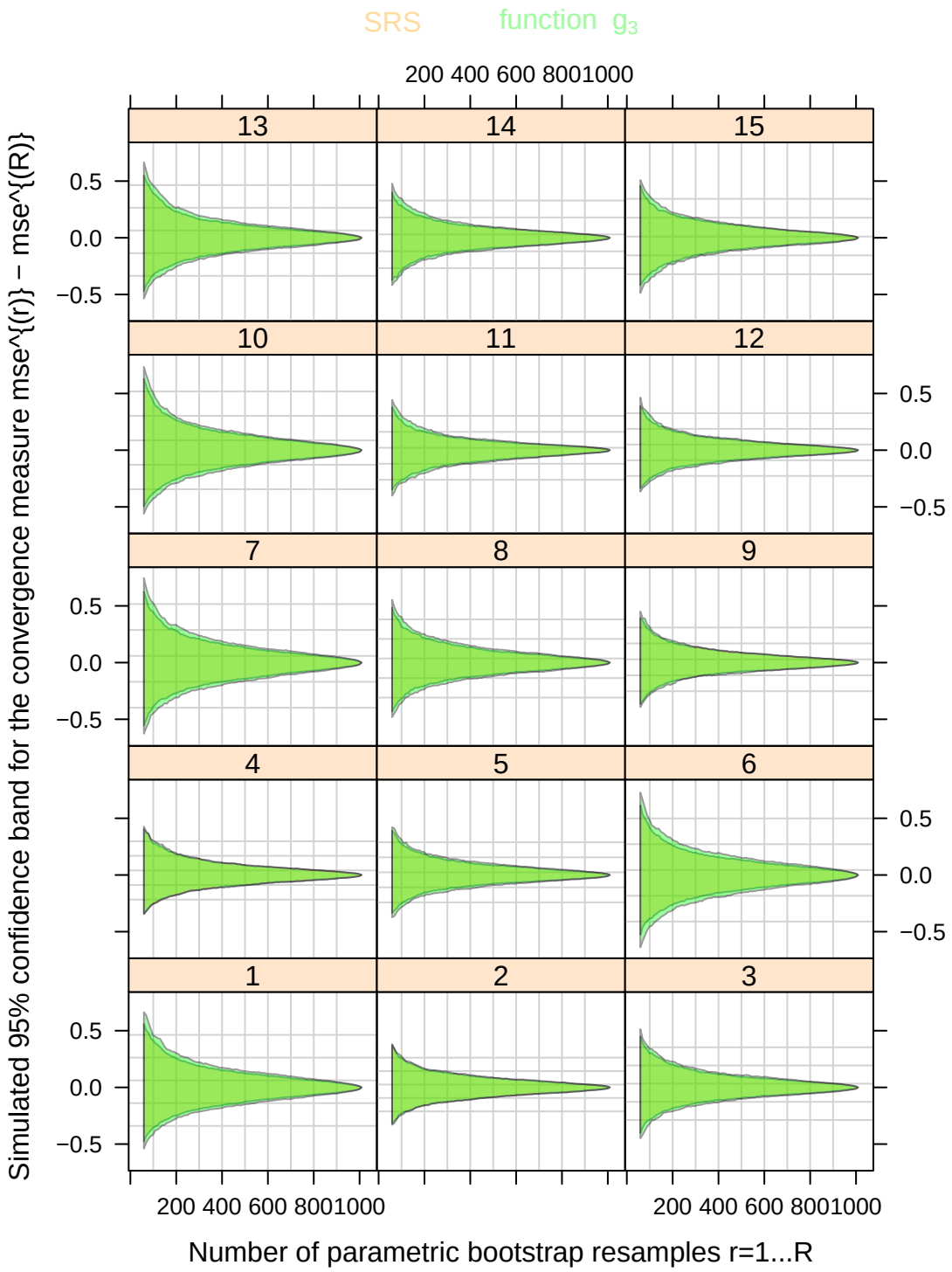

Figure B.26: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 8

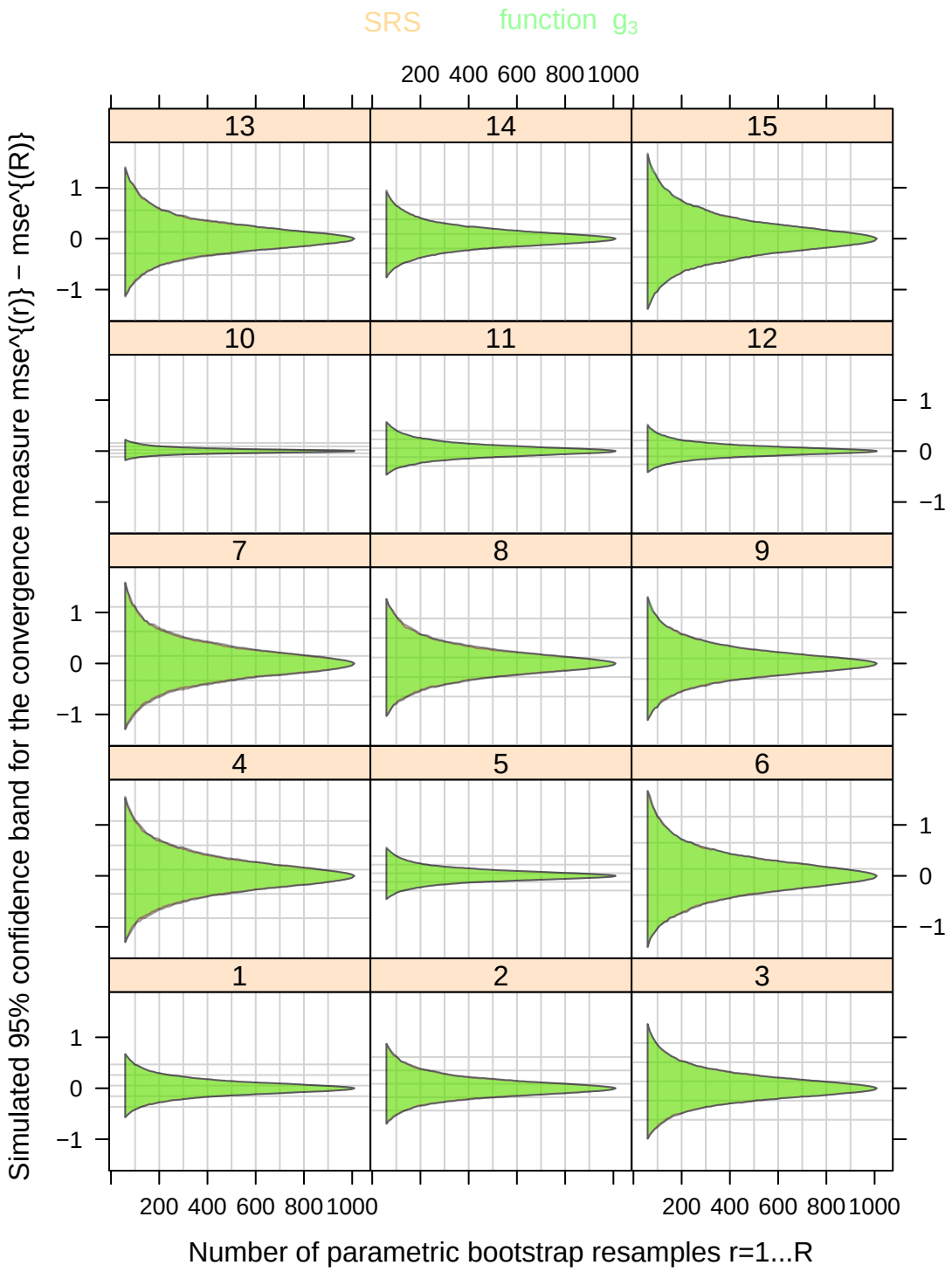

Figure B.27: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 10

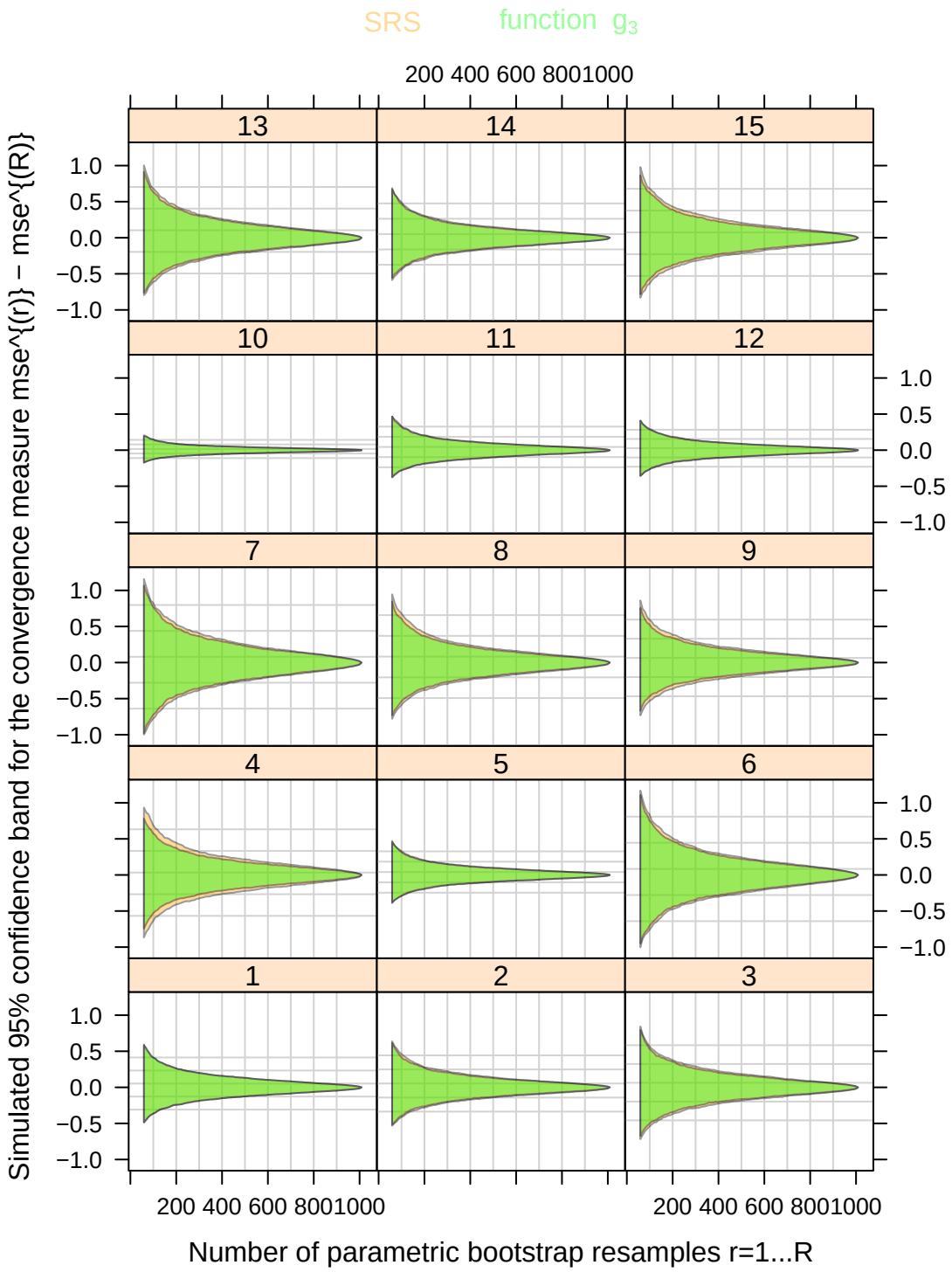

Figure B.28: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 11

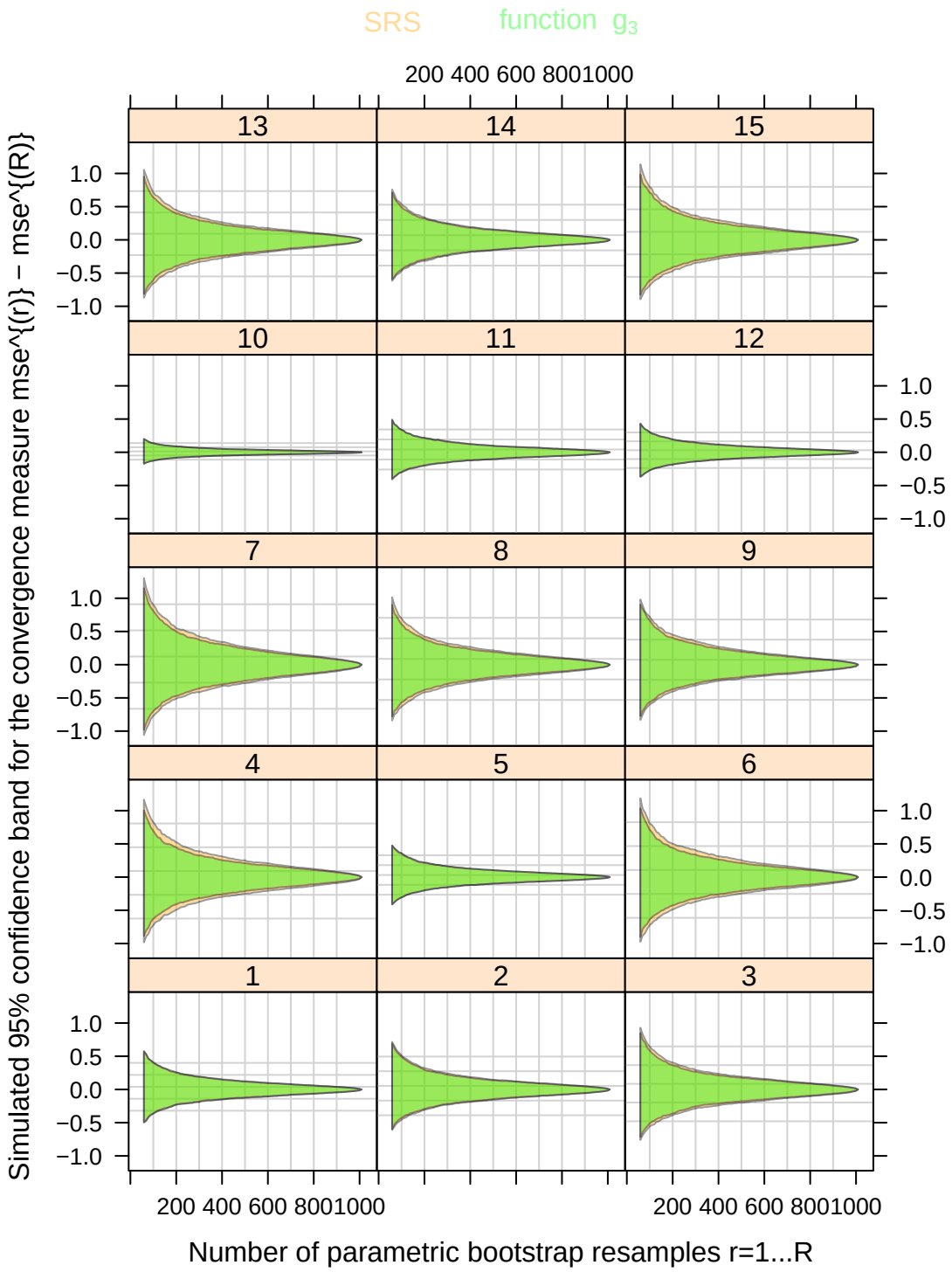

Figure B.29: Usingcontrol variate function $g^{(3)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 12

B.4 Additional Graphs for Control Variate $g^{(4)}$

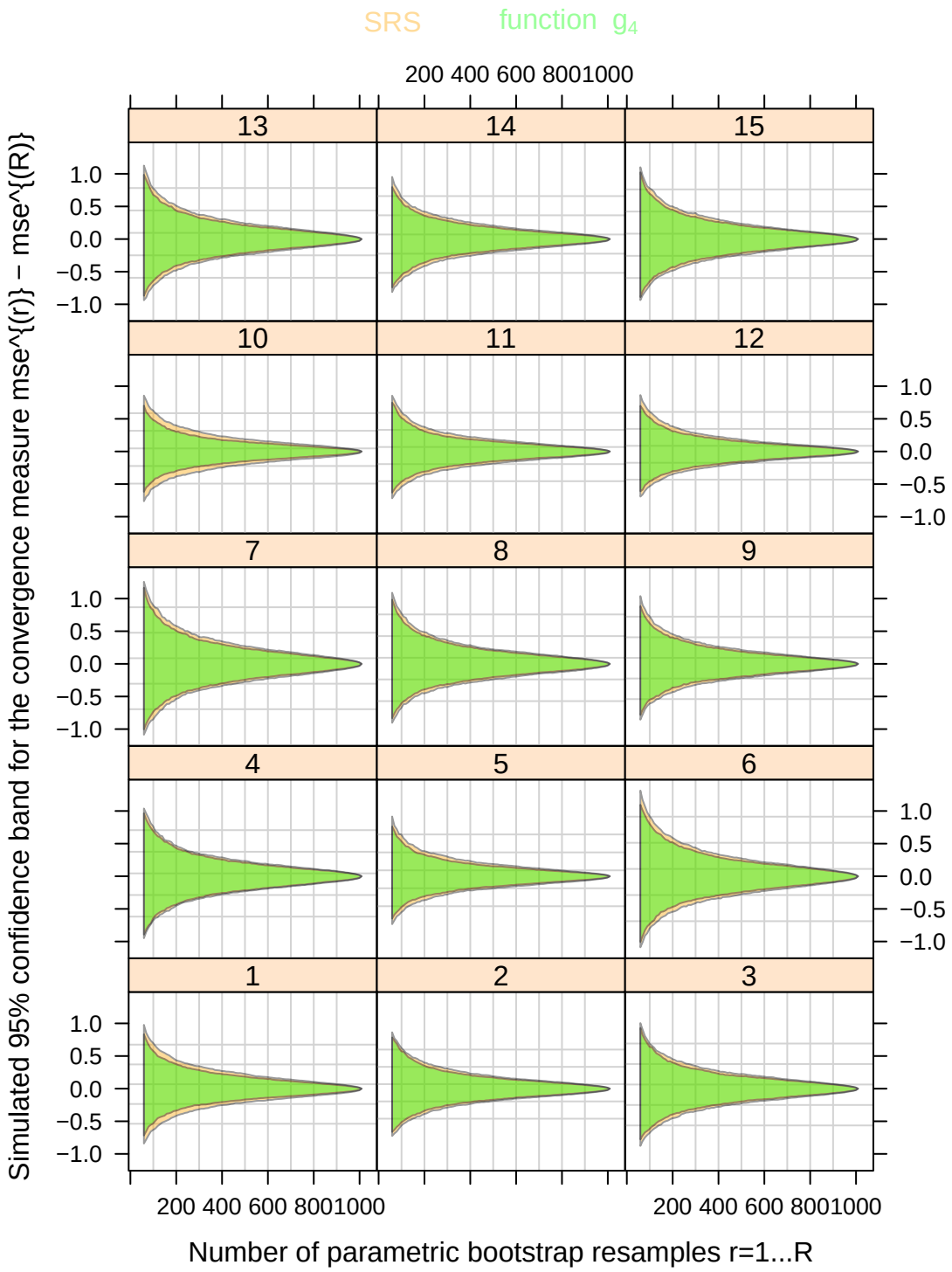

Figure B.30: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 2

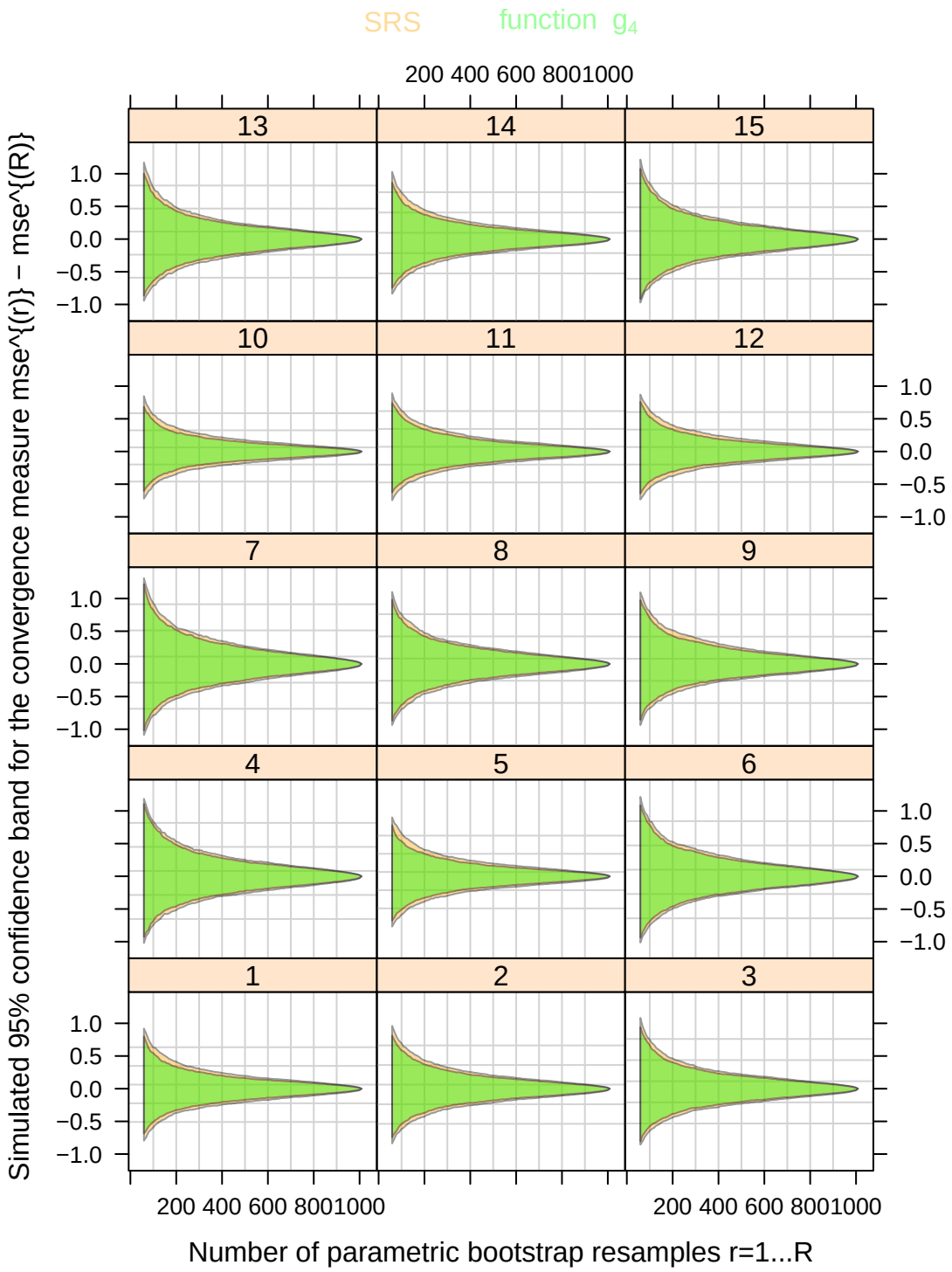

Figure B.31: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 3

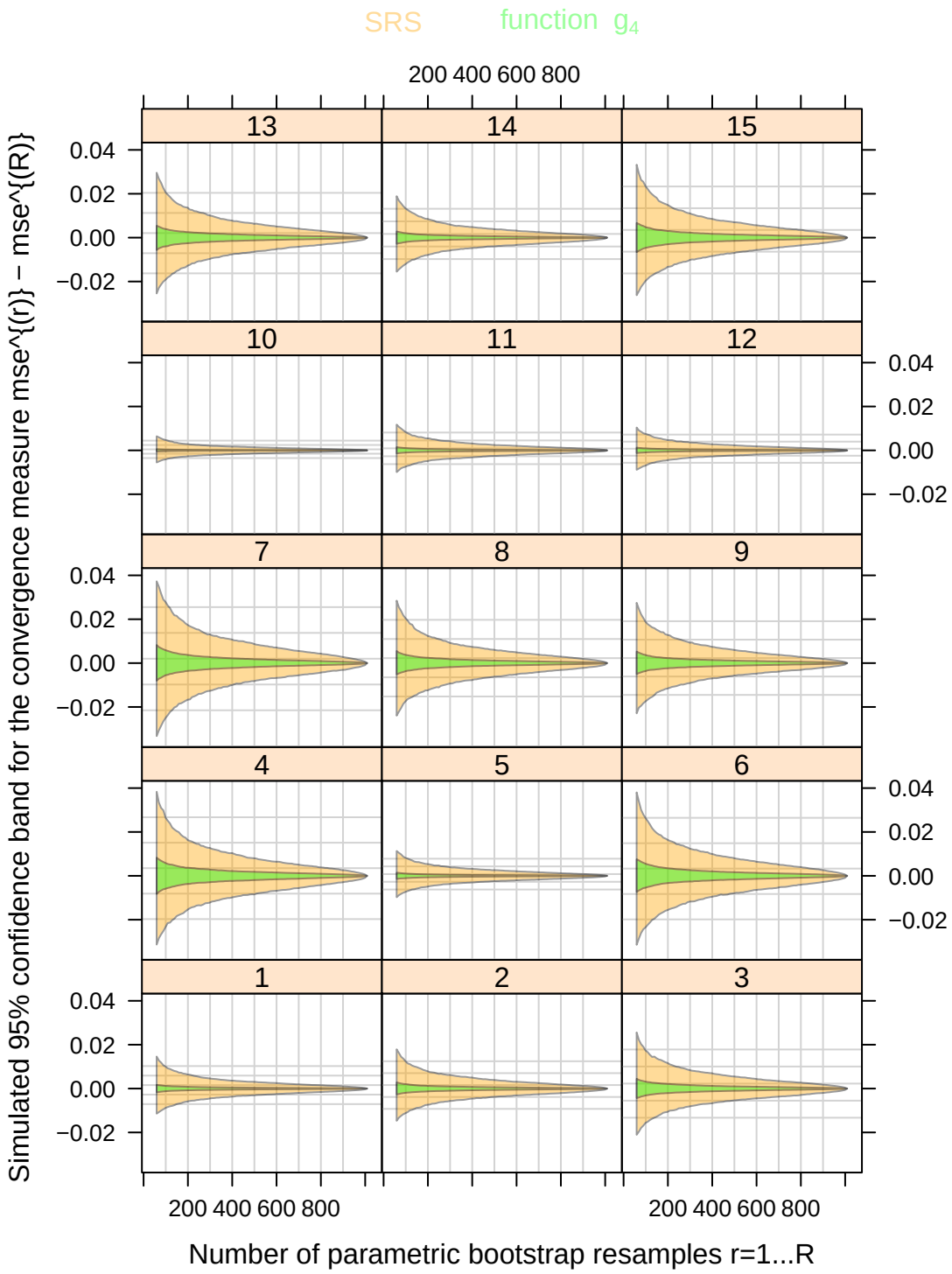

Figure B.32: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 5

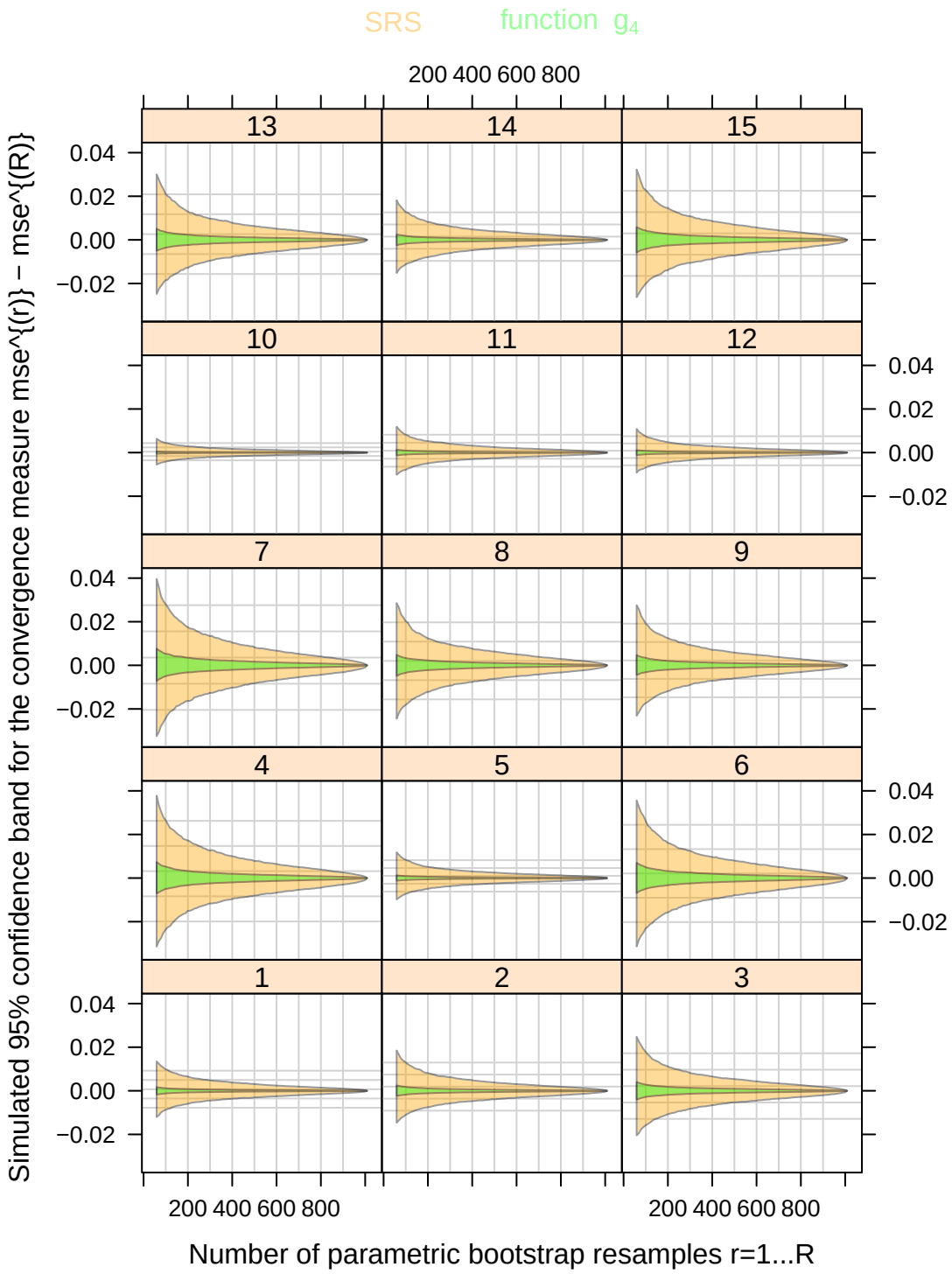

Figure B.33: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 6

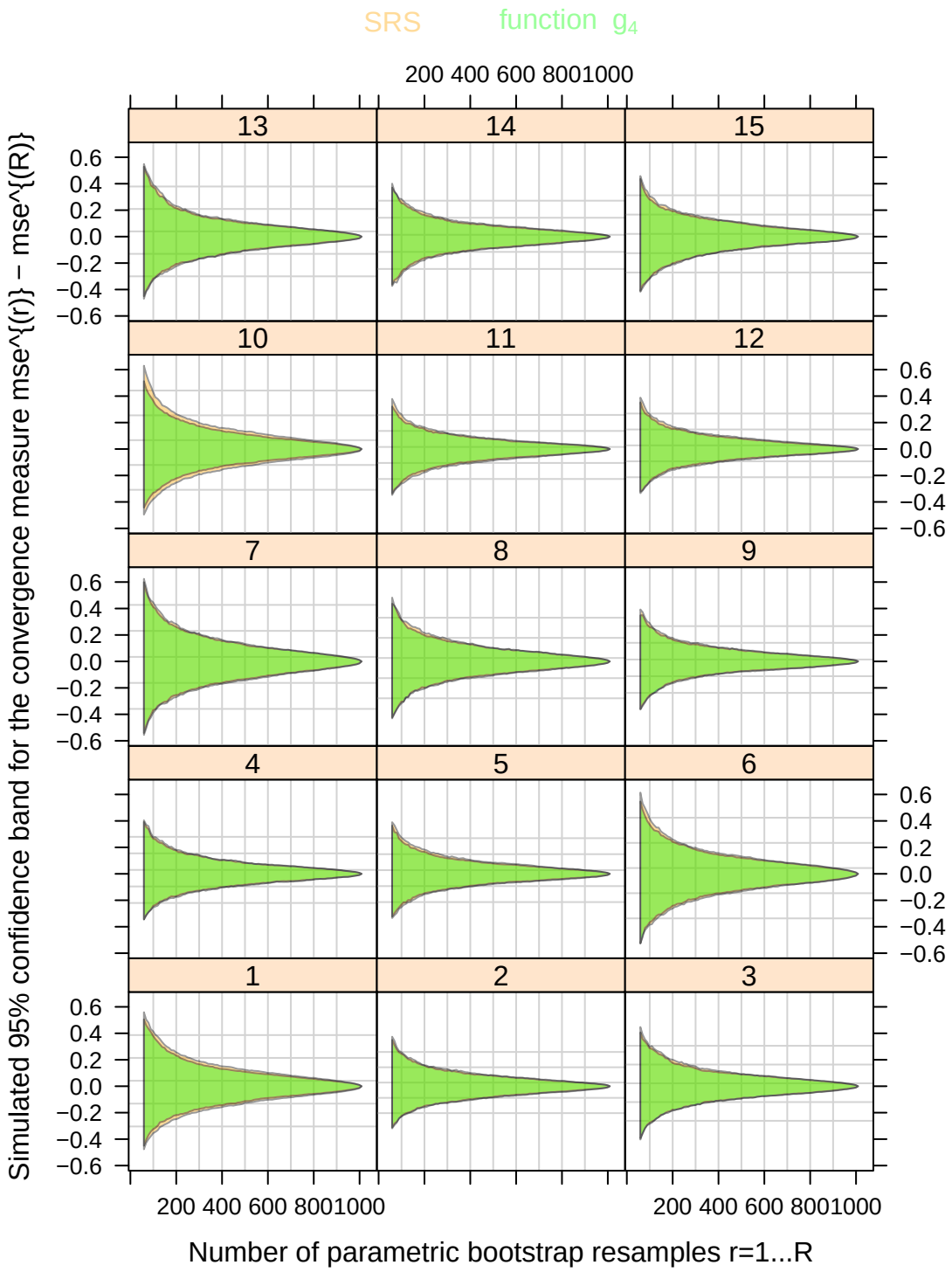

Figure B.34: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 8

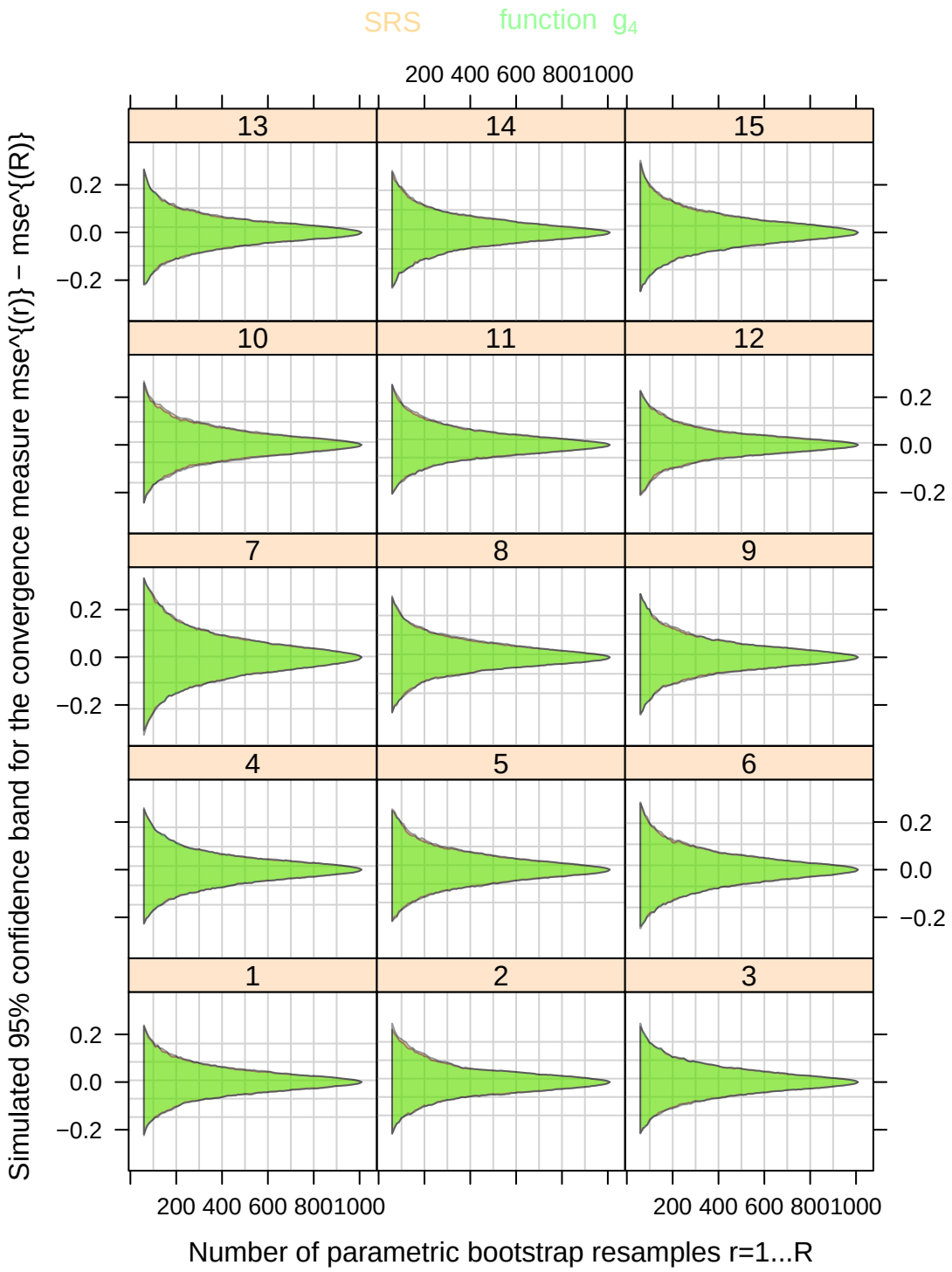

Figure B.35: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 9

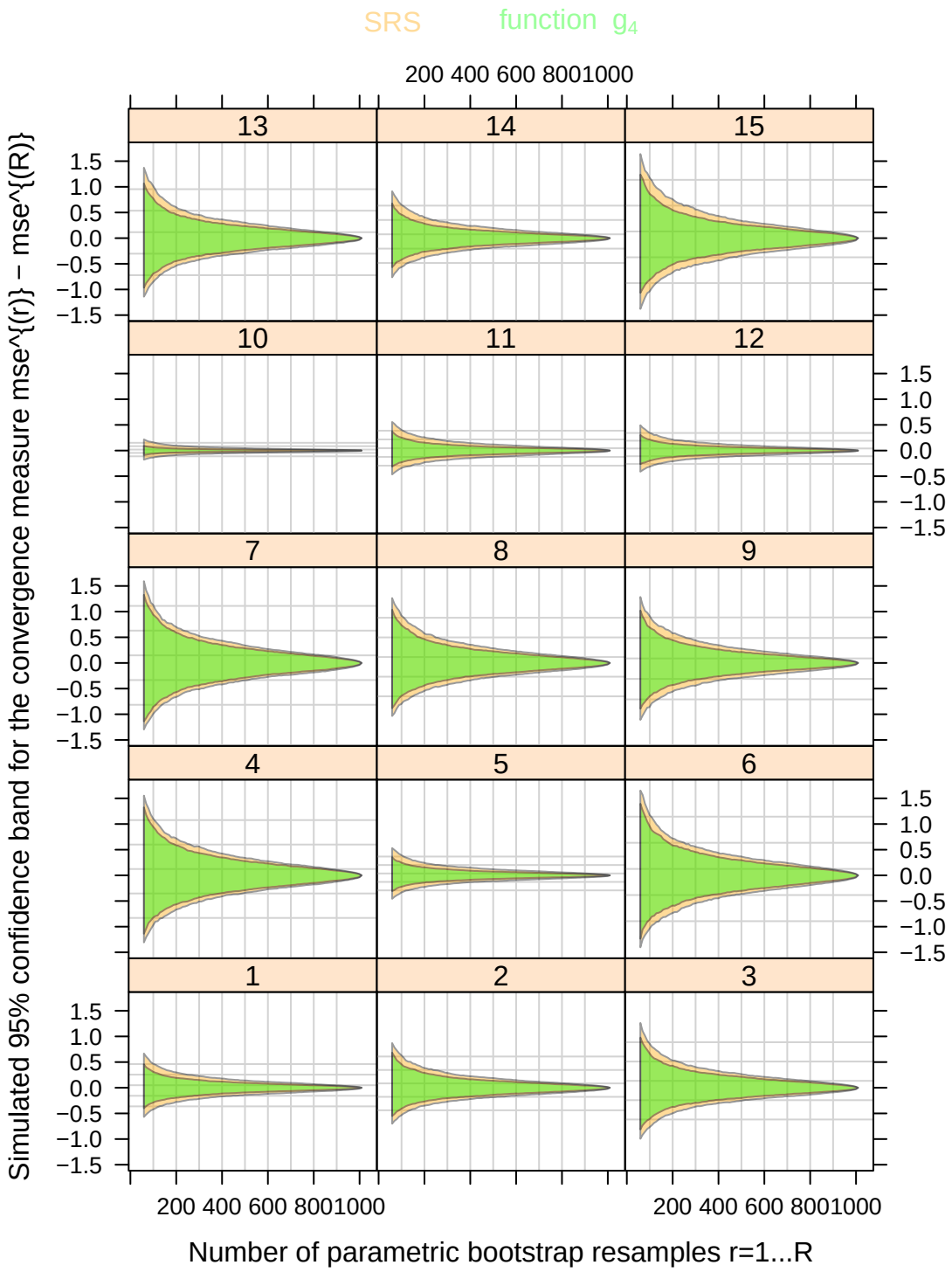

Figure B.36: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 10

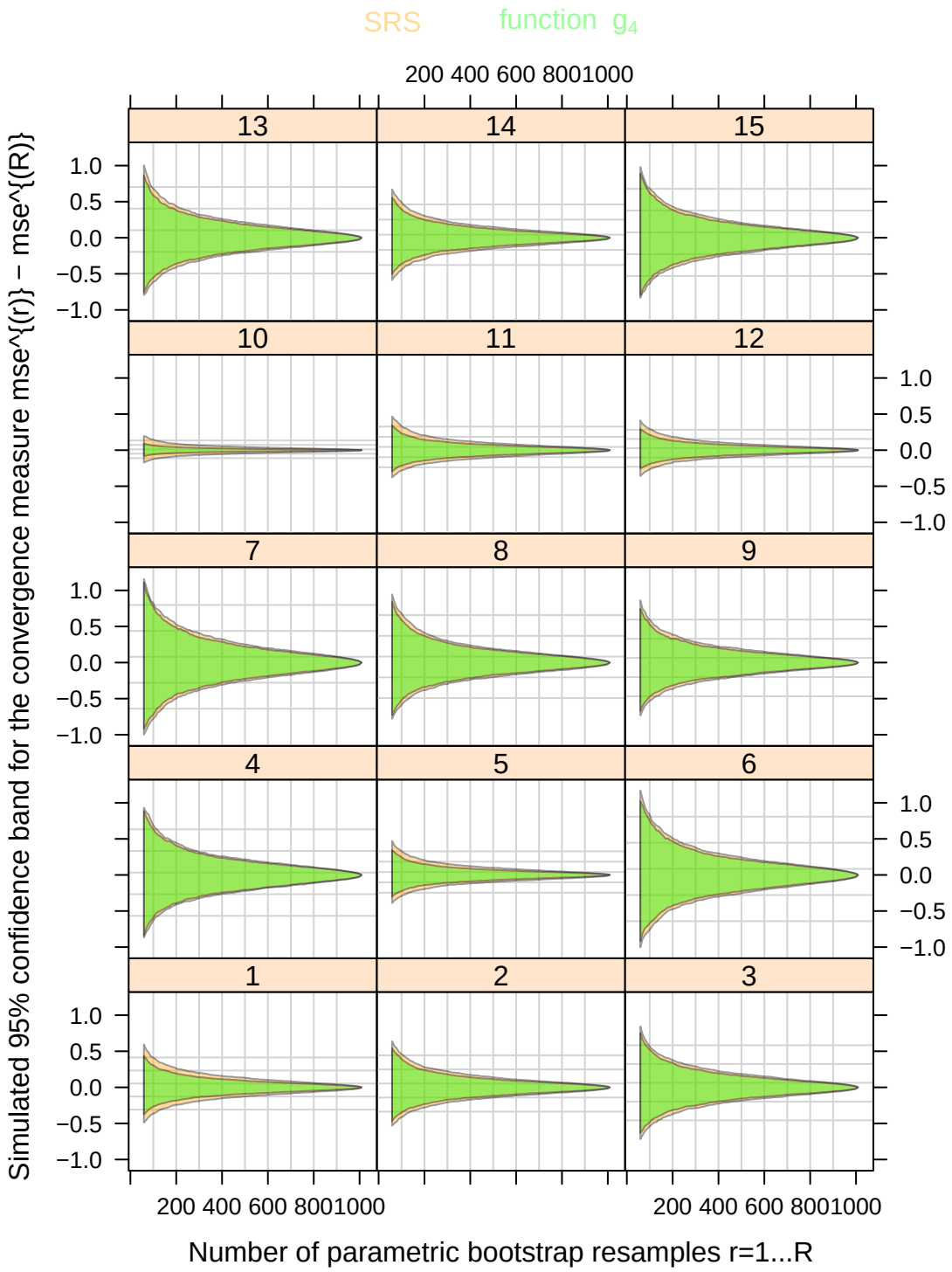

Figure B.37: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 11

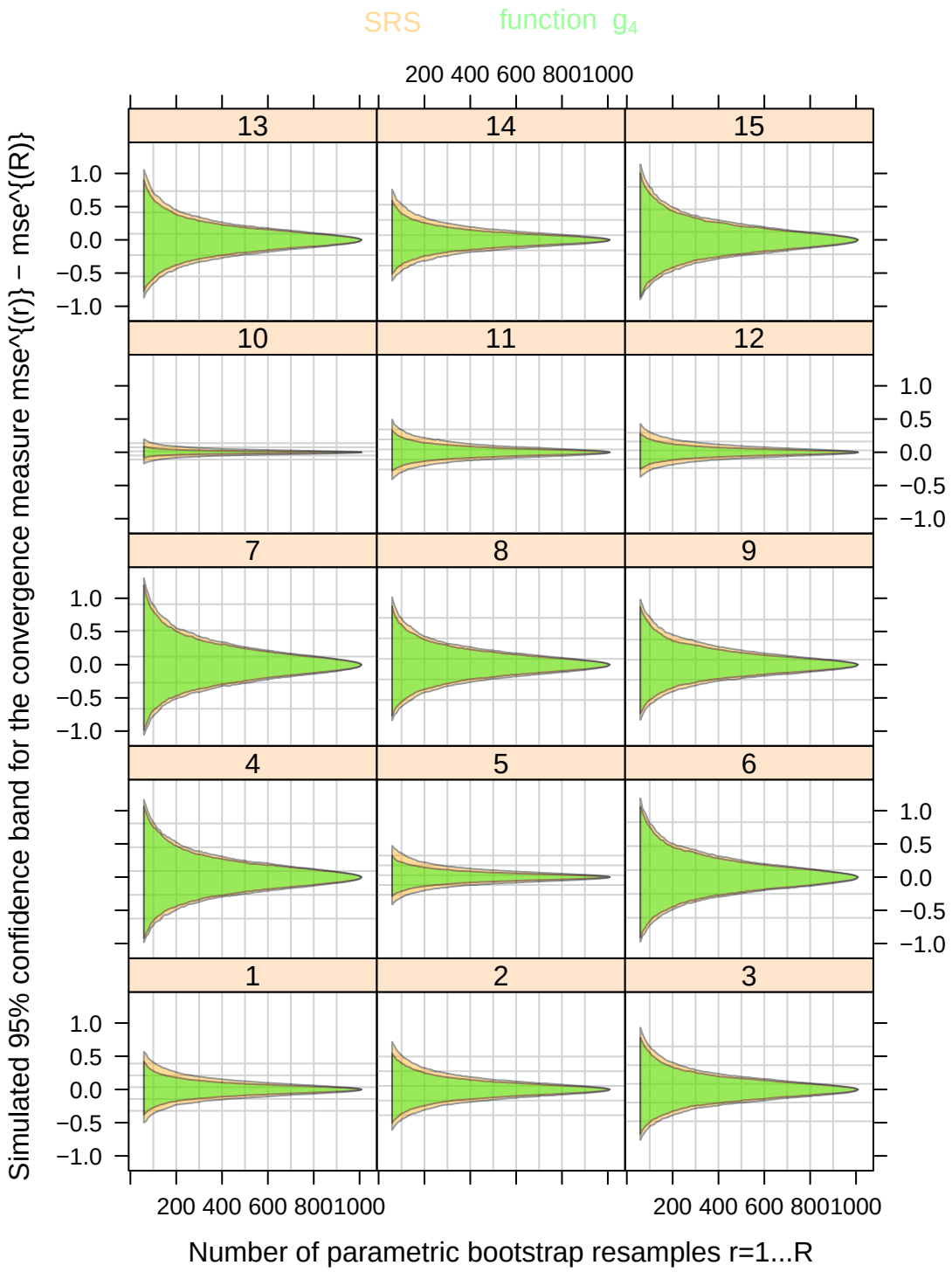

Figure B.38: Usingcontrol variate function $g^{(4)}$ for variance reduction of the parametric bootstrap MSE estimate for the Fay-Herriot estimator on the model based population 12

Appendix C

Bibliography

Babu, G., & Singh, K. (1983). Inference on means using the bootstrap. *The Annals of Statistics*, *11*(3), 999–1003.

Bailey, N. T. (1961). Introduction to the mathematical theory of genetic linkage. *Oxford University Press, London*.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28 – 36.

Björck, Å. (1996). *Numerical methods for least squares problems*. Philadelphia: Siam.

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M., & White, J. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Booth, J. G., Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, *89*(428), 1282–1289.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, *93*(441), 262–272.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(1), 265–285.

Bose, R. C., Chakravarti, I. M., & Knuth, D. E. (1960). On methods of constructing sets of mutually orthogonal latin squares using a computer. *Technometrics*, *2*(4), 507–516.

Box, J. F. (1980). R. A. Fisher and the design of experiments, 1922-1926. *The American Statistician*, *34*(1), 1–7.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Bundesamt für Statistik Schweiz. (2013a). *Einwohnerzahlen.* Retrieved 16 Juli 2013, from http://www.bfs.admin.ch/bfs/portal/de/index/themen/01/01/key.html

Bundesamt für Statistik Schweiz. (2013b). *Strukturerhebung.* Retrieved 16 Juli 2013, from http://www.bfs.admin.ch/bfs/portal/de/index/news/02/03/02.html

Bungartz, H. J., & Dirnstorfer, S. (2003). Multivariate quadrature on adaptive sparse grids. *Computing*, *71*(1), 89–114.

Burgard, J. P. (2009). *Erstellung von Karteileichen- und Fehlbestandsmodellen durch Multilevel-Modelle.* (Universität Trier, Diplomarbeit)

Burgard, J. P., & Münnich, R. T. (2012). Modelling Over- and Undercounts for Design-Based Monte Carlo Studies in Small Area Estimation: An Application to the German Register-Assisted Census. *Computational Statistics & Data Analysis*, *56*(10), 2856–2863.

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*(1–2), 63 – 76. (Special issue II: Model Selection, Model Diagnostics, Empirical Bayes and Hierarchical Bayes)

Cajori, F. (1911). Historical note on the newton-raphson method of approximation. *The American Mathematical Monthly*, *18*(2), 29–32.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, *9*(1), 49. Retrieved 16. Juli 2013, from http://www.biomedcentral.com/1471-2288/9/49

Chatterjee, S., & Lahiri, P. (2007). A simple computational method for estimating mean squared prediction error in general small-area model. In *Proceedings of the section on survey research methods* (pp. 3486–3493).

Chatterjee, S., Lahiri, P., & Li, H. (2007). On small area prediction interval problems. In *Proceedings of the section on survey research methods* (pp. 3494–3505).

Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of eblup and related prediction intervals in linear mixed models. *The Annals of Statistics*, *36*(3), 1221–1245.

Chattopadhyay, M., Lahiri, P., Larsen, M., & Reimnitz, J. (1999). Composite estimation of drug prevalences for sub-state areas. *Survey Methodology*, *25*(1), 81–86.

Chen, S., & Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems. *Communications in Statistics-Theory and Methods*, *37*(11), 1792–1798.

Choudhry, G. H., Rao, J. N. K., & Hidiroglou, M. A. (2011). On sample allocation for domains. In *Sae2011 workshop in trier.*

Choudhry, G. H., Rao, J. N. K., & Hidiroglou, M. A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, *38*(1), 23–29.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley & Sons.

Concordet, D., & Nunez, O. G. (2002). A simulated pseudo-maximum likelihood estimator for nonlinear mixed models. *Computational statistics & data analysis*, *39*(2), 187–201.

Cook, N. (1982). *A FORTRAN Program for Random-effects Models* (technical report). Boston: Harvard School of Public Health, Dept. of Biostatistics.

Costa, À., Satorra, A., & Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT*, *28*(1), 69-86.

Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *Comptes rendus de l'AcadÃ©mie des Sciences Paris*, *200*, 1265-1266.

Das, K., Jiang, J., & Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, *32*(2), 818–840.

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, *10*, 613–627.

Datta, G. S., Rao, J. N. K., & Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, *92*(1), 183–196.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. New York: Cambridge University Press.

Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science*, *18*(2), 141–157.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics*, *27*(1), 94–128.

Dempfle, L. (1977). Comparison of several sire evaluation methods in dairy cattle breeding. *Livestock Production Science*, *4*(2), 129–139.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

DESTATIS. (2012). *Zensus2011*. Retrieved 31-May-2012, from `http://www.zensus2011.de`

DiCiccio, T. J., & Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika*, *79*(2), 231–245.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–212.

DiCiccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(3), 338–354.

Djordjević, V., & Lepojević, V. (2003). Henderson's approach to variance components estimation for unbalanced data. *Facta Universitatis Series: Economics and Organization*, *2*(1), 59–64.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.

Dougherty, C. (2011). *Introduction to econometrics*. New York: Oxford University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, *7*(1), 1–26.

Efron, B. (1980, December). *The jackknife, the bootstrap and other resampling plans* (Tech. Rep. No. 63). Stanford: Stanford University, California.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans* (Vol. 38). Philadelphia: Society for Industrial and Applied Mathematics, Philadelphia.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). New York: Chapman & Hall/CRC.

Ehrenfeld, S., & Ben-Tuvia, S. (1962). The efficiency of statistical simulation procedures. *Technometrics*, *4*(2), 257–275.

Engel, B., & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, *48*(1), 1–22.

Entacher, K. (1998). Bad subsequences of well-known linear congruential pseudorandom number generators. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 61–70. (Special issue on uniform random number generation)

Euler, L. (1782). Recherches sur une nouvelle espece de quarres magiques. *Verhandelingen uitgegeven door het zeeuwsch Genootschap der Wetenschappen te Vlissingen*, *9*, 85–239.

Fahrmeir, L., Tutz, G., & Hennevogl, W. (1994). *Multivariate statistical modelling based on generalized linear models* (Vol. 2). New York: Springer New York.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, *Vol. 74, No. 366*, 269–277.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 22, pp. 700–725).

Fisher, R. A. (1935). *The design of experiments*. Oxford: Oliver & Boyd.

Fishman, G. S. (1996). *Monte carlo: concepts, algorithms, and applications*. Berlin: Springer.

Foulley, J.-L., & Van Dyk, D. A. (2000). The px-em algorithm for fast stable fitting of Henderson's mixed model. *Genetics Selection Evolution*, *32*(2), 1–21.

Frey, H. C., & Rhodes, D. S. (1999). *Theory and methodology based upon bootstrap simulation* (Tech. Rep. No. DOE/ER/30250 Vol. 1). Washington D.C.: U.S. Department of Energy Office of Energy Research.

Fuller, W. A. (1990). Prediction of true values for the measurement error model. *Statistical Analysis of Measurement Error Models and Applications*, *112*, 41–57.

Fuller, W. A. (2009). *Sampling statistics* (Vol. 560). Hoboken: John Wiley & Sons Inc.

Gabler, S., Ganninger, M., & Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, *75*(2), 151–161.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Hamburg: Perthes et Besser. (english translation by C. H. Davis, reprinted 1963 by Dover, New York)

Gaver, D. (1969). Statistical methods for improving simulation efficiency. In *Proceedings of the third conference on applications of simulation* (pp. 38–46).

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.

Gershunskaya, J., Jiang, J., & Lahiri, P. (2009). *Resampling methods in surveys* (Vol. 29).

Gerstner, T., & Griebel, M. (1998). Numerical integration using sparse grids. *Numerical algorithms*, *18*(3), 209–232.

Ghosh, M., Natarajan, K., Stroud, T. W. F., & Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, *93*(441), 273–282.

Giesbrecht, F. G., & Burrows, P. M. (1978). Estimating variance components in hierarchical structures using minque and restricted maximum likelihood. *Communications in Statistics-Theory and Methods*, *7*(9), 891–904.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*(1), 45–51.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2007, February). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, *51*, 2720–2733.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(2), 149–192.

Green, P. J. (1987). Penalized Likelihood for General Semi-parametric Regression Models. *International Statistical Review*, *55*(3), 245–259.

Griebel, M., Zenger, C., & Zimmer, S. (1992). *Improved multilevel algorithms for full and sparse grid problems* (SFB-Report 342/15/92 A). München: Technische Universität München.

Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 221–228.

Hall, P. (1997). *The bootstrap and edgeworth expansion*. New York: Springer Verlag.

Hall, P. (2003). A short prehistory of the bootstrap. *Statistical Science*, *18*(2), 158–167.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal Of The Royal Statistical Society Series B*, *68*(2), 221–238.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*(1-2), 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Harville, D. A. (1990). Blup (best linear unbiased prediction) and beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock* (pp. 239–276). Springer.

Henderson, C. R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, *21*, 309–310.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*(2), 226–252.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical genetics and plant breeding*, *982*, 141–163.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423–447.

Hesterberg, T. (1996). Control variates and importance sampling for efficient bootstrap simulations. *Statistics and Computing*, *6*(2), 147–157.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Jiang, J. (1997). A derivation of blupâ€"best linear unbiased predictor. *Statistics & Probability Letters*, *32*(3), 321–324.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, *93*(442), 720–729.

Jiang, J. (2000). A matrix inequality and its statistical application. *Linear Algebra and its Applications*, *307*(1), 131–144.

Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, *53*(2), 217–243.

Jiang, J., & Lahiri, P. (2006a). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, *101*(473), 301–311.

Jiang, J., & Lahiri, P. (2006b). Mixed model prediction and small area estimation. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, *15*(1), 1–96.

Jiang, J., Lahiri, P., & Wan, S. (1998). Jackknifing the mean squared error of empirical best predictor. *Technical Report*.

Jiang, J., Lahiri, P., & Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, *30*(6), 1782–1810.

Kackar, R. N., & Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in statistics-theory and methods*, *10*(13), 1249–1261.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association*, *79*(388), 853–862.

Kahaner, D., Moler, C. B., Nash, S., & Forsythe, G. E. (1989). *Numerical methods and software*. Englewood Cliffs, NJ: Prentice-Hall.

Kale, B. (1961). On the solution of the likelihood equation by iteration processes. *Biometrika*, *48*(3/4), 452–456.

Kale, B. (1962). On the solution of likelihood equations by iteration processes. the multiparametric case. *Biometrika*, *49*(3/4), 479–486.

Kendall, M. G., & Babington-Smith, B. (1938). Randomness and random sampling numbers. *Journal of the Royal Statistical Society*, *101*(1), 147–166.

Kleffe, J., & Seifert, B. (1986). Computation of variance components by the minque method. *Journal of multivariate analysis*, *18*(1), 107–116.

Knight, E. (2008). *Improved Iterative Schemes for REML Estimation of Variance Parameters in Linear Mixed Models* (Phd thesis). School of Agriculture, Food and Wine. The University of Adelaide.

Knuth, D. E. (1981). *The art of programming* (Vol. 2: Semi-Numerical Algorithms). Amsterdam: Addison Wesley, Reading, MA.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–409.

Kott, P. S. (2001). The Delete-a-Group Jackknife. *Journal of Official Statistics*, *17*(4), 521–526.

Kowalsky, A. (1924). *Basic theory of sampling methods*. Saratov: State University of Saratov.

Kuhn, E., & Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, *8*, 115–131.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, *18*(2), 199–210.

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American statistical association*, *100*(469), 222–230.

Lahiri, P., & Li, H. (2009a). An adaptive hierarchical bayes quality measurement plan. *Applied Stochastic Models in Business and Industry*, *25*(4), 468–477.

Lahiri, P., & Li, H. (2009b). Generalized maximum likelihood method in linear mixed models with an application in small-area estimation. In *Proceedings of the federal committee on statistical methodology research conference.*

Lahiri, P., & Mukherjee, K. (2007). On the design-consistency property of hierarchical bayes estimators in finite population sampling. *Annals of Statistics*, *35*, 724–737.

Lahiri, P., & Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, *90*(430), 758–766.

Laird, N. (1978). Empirical bayes methods for two-way contingency tables. *Biometrika*, *65*(3), 581–590.

Laird, N., Lange, N., & Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the em algorithm. *Journal of the American Statistical Association*, *82*(397), 97–105.

L'Ecuyer, P. (1994). Uniform random number generation. *Annals of Operations Research*, *53*(1), 77–120.

Legendre, A. M. (1805). Sur la méthode des moindres quarrés. *Nouvelles méthodes pour la détermination des orbites des comètes*. Retrieved 16. July 2013, from `https://play.google.com/store/books/details?id=JYQ_AAAAcAAJ&rdid=book-JYQ_AAAAcAAJ&rdot=1`

Lehmer, D. H. (1951). Mathematical methods in large-scale computing units. In *Proceedings of the 2nd symposium on large-scale digital calculating machinery* (pp. 141–146). Harvard University Press, Cambridge, MA.

Lehtonen, R., & Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, *24*(1), 51–55.

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, *101*, 882–892.

Li, Y., & Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, *102*(478), 664–673.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lindley, D. V., & Smith, A. F. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(1), 1–41.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Liu, B. (2009). *Hierarchical bayes estimation and empirical best prediction of small-area proportions* (Unpublished doctoral dissertation). University of Maryland (College Park, Md.).

Liu, C., & Rubin, D. B. (1994). The ecme algorithm: A simple extension of em and ecm with faster monotone convergence. *Biometrika*, *81*(4), 633–648.

Liu, C., Rubin, D. B., & Wu, Y. N. (1998). Parameter expansion to accelerate em: the px-em algorithm. *Biometrika*, *85*(4), 755–770.

Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 1696–1708.

Lohr, S. L. (2010). *Sampling: design and analysis* (2nd ed.). Boston: Brooks/Cole.

Lohr, S. L., & Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, *96*(2), 457–468.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*(4), 817–827.

Longford, N. T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, *32*(1), 87.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics*. *Economic Record*, *82*, S2–S18.

Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the indian statistical institute. *Journal of the Royal Statistical Society*, *109*(4), 325–378.

Malec, D., Sedransk, J., Moriarity, C. L., & LeClere, F. B. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, *92*(439), 815–826.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, *21*(1), 255–285.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 3–30. (Special issue on uniform random number generation)

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245.

McLachlan, G. J., & Krishnan, T. (2007). *The em algorithm and extensions* (Vol. 382). Hoboken: Wiley-Interscience.

Meng, X.-L., & Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, *80*(2), 267–278.

Meza, C., Jaffrézic, F., & Foulley, J.-L. (2007). Reml estimation of variance parameters in nonlinear mixed effects models using the saem algorithm. *Biometrical Journal*, *49*(6), 876–888.

Miller, R. G. (1964). A trustworthy jackknife. *The Annals of Mathematical Statistics*, *35*(4), 1594–1605.

Miller, R. G. (1974). The jackknife - a review. *Biometrika*, *61*(1), 1–15.

Morozova, E. Y. (2008). A multidimensional bisection method for unconstrained minimization problem. In *Proceedings of the fourteenth symposium on computing: the australasian theory* (Vol. 77, pp. 57–62). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Morris, C., & Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science*, *26*(2), 271–287.

Moura, F. A. d. S., & Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, *25*, 73–80.

Münnich, R. T. (1997). *Gebundene Hochrechnung bei Stichprobenerhebungen mit Hilfe von Splines* (No. 43). Göttingen: Vandenhoeck & Ruprecht.

Münnich, R. T., & Burgard, J. P. (2012a). On the influence of sampling design on small area estimates. *Journal of the Indian Society of Agricultural Statistics*, *66*, 145–156.

Münnich, R. T., & Burgard, J. P. (2012b). *Simulation der Strukturerhebung und Kleingebiet-Schätzungen.* Retrieved 16 Juli 2013, from `http://www.uni-trier.de/index.php?id=45166`

Münnich, R. T., & Burgard, J. P. (2012c). *Simulation der Strukturerhebung und Kleingebiet-Schätzungen.* Retrieved 16 Juli 2013, from `http://www.bfs.admin.ch/bfs/portal/de/index/news/02/07/01.parsys.76923.downloadList.69899.DownloadFile.tmp/berichtsmallareaestimationbfsunitrier.pdf`

Münnich, R. T., Burgard, J. P., & Vogt, M. (2013). Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *6*(3), 149–191.

Münnich, R. T., Burgard, J. P., & Zimmermann, T. (2012). *Small area modelling under complex survey designs for business data.* (Contributed paper at the Fourth International Conference on Establishment Surveys, Montreal)

Münnich, R. T., Sachs, E. W., & Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Advances in Statistical Analysis*, *96*(3), 435–450.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, *3*, 169–174.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Orchard, T., & Woodbury, M. A. (1972). A Missing Information Principle: Theory and Applications. In *Proceedings of the 6th berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 697–715).

Owen, A. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, *2*(2), 439–452.

Panneton, F., L'Ecuyer, P., & Matsumoto, M. (2006). Improved long-period generators based on linear recurrences modulo 2. *ACM Transactions on Mathematical Software (TOMS)*, *32*(1), 1–16.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Peixoto, J. L. (1988). Prediction in a class of mixed models with two variance components. *Journal of statistical planning and inference*, *19*(1), 81–94.

Peixoto, J. L., & Harville, D. A. (1986). Comparisons of alternative predictors under the balanced one-way random model. *Journal of the American Statistical Association*, *81*(394), 431–436.

Pérez, B., Peña, D., & Molina, I. (2011). Robust henderson iii estimators of variance components in the nested error model. In *Modern mathematical tools and techniques in capturing complexity* (pp. 329–339). Springer.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *60*(1), 23–40.

Pfeffermann, D., & Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, *102*(480), 1427-1439.

Pinheiro, J., & Bates, D. (2000). *Mixed effects models in s and s-plus*. New York.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 32, pp. 567–579).

Pokropp, F. (1996). *Stichproben: Theorie und Verfahren*. München: Oldenbourg Verlag.

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, *52*, 145–158.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171.

Prasad, N. G. N., & Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, *25*, 67–72.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, *11*(1), 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3/4), 353–360.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Raj, D. (1968). *Sampling theory*. New York: McGraw-Hill New York.

Ranalli, M. G., & Mecatti, F. (2012). Comparing recent approaches for bootstrapping sample survey data: A first step toward a unified approach. In *Joint statistical meeting (jsm) 2012*.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, *65*(337), 161–172.

Rao, C. R. (1971a). Estimation of variance and covariance components–minque theory. *Journal of multivariate analysis*, *1*(3), 257–275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, *1*(4), 445–456.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, *67*(337), 112–115.

Rao, J. N. K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Rao, P. S. R. S. (1977). Theory of the minque: A review. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002)*, *39*(3), 201–210.

Rapanos, N. (2008). Latin squares and their partial transversals. *Harvard College Mathematics Review, SD Kominers, Ed. Harvard College*, *2*, 4–12.

Raphson, J. (1690). *Analysis aequationum universalis*. London.

Rizzo, M. L. (2008). *Statistical computing with R*. Boca Raton: Chapman & Hall.

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer.

Saei, A., & Chambers, R. (2003). *Small area estimation under linear and generalized linear mixed models with time and area effects* (Tech. Rep. No. M03/15). Southampton: Southampton Statistical Sciences Research Institute. Retrieved 16. Juli 2013, from `http://eprints.soton.ac.uk/8165/`

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

Sarraj, R., & Rosen, D. v. (2009). Improving Henderson's Method 3 Approach when Estimating Variance Components in a Two-way Mixed Linear Model. In B. Schipp & W. Kräer (Eds.), *Statistical Inference, Econometric Analysis and Matrix Algebra* (pp. 125–142). Physica-Verlag HD.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, *78*(4), 719–727.

Schmid, T. (2011). *Spatial Robust Small Area Estimation applied on Business Data* (Unpublished doctoral dissertation). Universität Trier, Universitätsring 15, 54296 Trier.

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance Components*. New York: Wiley Online Library.

Shao, J., & Tu, D. (1996). *The Jackknife and Bootstrap*. New York: Springer.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory* (Vol. 9). Philadelphia: Society for Industrial Mathematics.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer.

Singh, R., & Mangat, N. S. (1996). *Elements of survey sampling* (Vol. 15). Dordrecht: Kluwer Academic Publishers.

Skinner, C. J. (1989). Analysis of complex surveys. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), (pp. 59–87). New York: Wiley & Sons.

Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Doklady of the academy of sciences of the ussr* (Vol. 4, pp. 240–243).

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, *40*(4), 961–971.

Swallow, W. H., & Monahan, J. F. (1984). Monte carlo comparison of anova, mivque, reml, and ml estimators of variance components. *Technometrics*, *26*(1), 47–57.

Tippett, L. H. C. (1927). Tables of random sampling numbers. *Cambridge University Tracts for computers*(15).

Torabi, M., & Rao, J. (2010). Mean squared error estimators of small area means using survey weights. *Canadian Journal of Statistics*, *38*(4), 598–608.

Tschuprow, A. (1923a). On the mathematical expectation of the moments of frequency distributions in the case of. *Metron*, *2*(3), 461–493.

Tschuprow, A. (1923b). On the mathematical expectation of the moments of frequency distributions in the case of. *Metron*, *2*(3), 646–680.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, *29*(2), 614–623.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

van Dyk, D. A. (2000). Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms. *Journal of Computational and Graphical Statistics*, *9*(1), 78–98.

Verbeek, M. (2005). *A Modern Guide to Econometrics*. Chichester: Wiley.

Vogt, M. (2007). *Small Area Estimation: Die Schätzer von Fay-Herriot und Battese-Harter-Fuller* (diploma-thesis). Universität Trier.

von Auer, L. (2011). *Ökonometrie: Eine Einführung*. Berlin: Springer.

Wallis, J. (1685). *A treatise of algebra, both historical and practical: shewing the original, progress, and advancement thereof, from time to time, and by what steps it hath attained to the heighth at which now it is ; with some additional treatises*. London: Richard Davis, London. Retrieved 16. Juli 2013, from `http://echo.mpiwg-berlin.mpg.de/ECHOdocuViewfull?howpublished=/mpiwg/online/permanent/library/H3GRV5AU/pageimg&tocMode=thumbs&viewMode=images&pn=1&mode=imagepath`

Wichura, M. J. (1988). Algorithm as 241: The percentage points of the normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *37*(3), 477–484.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, *48*(3-4), 233–243.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295.

Wu, C.-t., Gumpertz, M. L., & Boos, D. D. (2001). Comparison of gee, minque, ml, and reml estimating equations for normally distributed data. *The American Statistician*, *55*(2), 125–130.

Yoshimori, M., & Lahiri, P. (2012). A New Adjusted Residual Likelihood Method for the Fay-Herriot Small Area Model. In *Section on Survey Research Methods â€“ Joint Statistical Meeting 2012*.

You, Y., & Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, *30*(3), 431–439.

Žarković, S. S. (1956). Note on the history of sampling methods in russia. *Journal of the Royal Statistical Society. Series A (General)*, *119*(3), 336–338.

Žarković, S. S. (1962). A supplement to "note on the history of sampling methods in russia". *Journal of the Royal Statistical Society. Series A (General)*, *125*(4), 580–582.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, *44*(4), 1049–1060.