Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Estimation of domain discontinuities using Hierarchical Bayesian Fay-Herriot models

by Jan A. van den Brakel and Harm-Jan Boonstra

Release date: June 24, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 151-189
Statistics Canada, Catalogue No. 12-001-X

Estimation of domain discontinuities using Hierarchical Bayesian Fay-Herriot models

Jan A. van den Brakel and Harm-Jan Boonstra[1]

Abstract

Changes in the design of a repeated survey generally result in systematic effects in the sample estimates, which are further referred to as discontinuities. To avoid confounding real period-to-period change with the effects of a redesign, discontinuities are often quantified by conducting the old and the new design in parallel for some period of time. Sample sizes of such parallel runs are generally too small to apply direct estimators for domain discontinuities. A bivariate hierarchical Bayesian Fay-Herriot (FH) model is proposed to obtain more precise predictions for domain discontinuities and is applied to a redesign of the Dutch Crime Victimization Survey. This method is compared with a univariate FH model where the direct estimates under the regular approach are used as covariates in a FH model for the alternative approach conducted on a reduced sample size and a univariate FH model where the direct estimates for the discontinuities are modeled directly. An adjusted step forward selection procedure is proposed that minimizes the WAIC until the reduction of the WAIC is smaller than the standard error of this criteria. With this approach more parsimonious models are selected, which prevents selecting complex models that tend to overfit the data.

Key Words: Area level models; Bivariate Fay-Herriot model; Small area estimation; Survey redesign; Measurement bias; MCMC; Gibbs sampler.

1. Introduction

Official statistics produced by national statistical institutes are generally based on repeated sample surveys. Much of their value lies in their continuity, enabling developments in society and the economy to be monitored, and policy actions decided. Survey samples contain besides sampling errors different sources of non-sampling errors that have a systematic effect on the outcomes of a survey. As long as the survey process is kept constant, this bias component is not visible. This is often an argument to keep survey processes of repeated surveys unchanged as long as possible. From time to time changes in surveys are needed to improve the efficiency, reduce the survey related costs, or meet new requirements, and this is seen strongly in the use of mixed-mode surveys including web-based questionnaires in official statistics. A redesign of the survey process generally has systematic effects on the survey estimates, since the biases induced by the aforementioned non-sampling errors are changed, disturbing comparability with figures published in the past.

Systematic differences in the outcomes of a repeated survey due to redesign of the survey process are called discontinuities. To avoid the implementation of a new survey process disturbing the comparability of estimates over time, it is important to quantify these discontinuities. This avoids confounding real change in the parameters of interest with changing measurement bias due to alteration of the survey process.

1. Jan A. van den Brakel, Statistics Netherlands, Department of Statistical Methods and Maastricht University, Department of Quantitative Economics. E-mail: ja.vandenbrakel@cbs.nl; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods.

Several methods to quantify discontinuities are proposed in the literature (van den Brakel, Smith and Compton, 2008). A reliable and straightforward approach is to conduct the old and new approach alongside of each other at the same time for some period of time, further referred to as a parallel run. Ideally this is based on a randomized experiment that can be embedded in the probability sample of the survey (van den Brakel, 2008). In this paper small area estimation methods for estimating domain discontinuities are proposed. We consider the situation where the regular survey, used for the production of official figures, is conducted at the full sample size and is conducted in parallel with an alternative approach. Due to budget limitations, the sample that is assigned to the alternative approach is often not sufficiently large to observe minimum detectable differences at prespecified significance and power levels using standard direct estimators, particularly for sub populations or domains.

To explain the problem addressed in this paper, some notation is introduced. Let θ_i denote the real population value of a variable of interest for domain i. Furthermore, $\hat{y}_i^r$ and $\hat{y}_i^a$ denote direct estimates of θ_i based on the regular survey and the alternative survey approach, respectively. Since the regular survey is conducted at the regular sample size, $\hat{y}_i^r$ is a reliable direct estimate for θ_i, at least for the planned domains. Due to the reduced sample size of the new survey in the parallel run $\hat{y}_i^a$, however, will be insufficiently precise. More precise domain estimates with the small sample available under the new approach can be obtained with the Fay-Herriot (FH) model (Fay and Herriot, 1979), which is defined as $\hat{y}_i^a = \mathbf{x}_i^t\beta + \nu_i + e_i^a$, with $\mathbf{x}_i$ a vector with covariates at the domain level, β the regression coefficients, ν_i the random domain effects and e_i^a the sampling error. To obtain more precise domain estimates for the alternative approach, van den Brakel, Buelens and Boonstra (2016) proposed an hierarchical Bayesian (HB) univariate FH model, where sample estimates of the regular survey are considered as potential auxiliary variables in a model selection procedure. This implies that $\hat{y}_i^r$ is used as a covariate in $\mathbf{x}_i$, besides the usual covariates that are available from registers or censuses. This results in an area level model, with measurement error (Ybarra and Lohr, 2008). The use of reliable direct estimates observed in the regular survey significantly increased the precision of the domain estimates for the alternative approach conducted at reduced sample size (van den Brakel et al., 2016).

Let $\tilde{y}_i^a$ denote the small area prediction for θ_i based on the aforementioned FH model under the small sample assigned to the alternative survey approach. In the approach followed by van den Brakel et al. (2016), point estimates for domain discontinuities are obtained as the difference between the direct estimate obtained with the regular survey and the model based domain prediction obtained under the alternative approach, i.e., $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. The use of the direct estimate of the regular survey as an auxiliary variable in the small domain predictions of the alternative survey, results in strong positive correlations between both estimators, which cannot be ignored when computing the standard errors for the discontinuities. More precisely, $\mathrm{Var}(\tilde{\Delta}_i) = \mathrm{Var}(\hat{y}_i^r) + \mathrm{MSE}(\tilde{y}_i^a) - 2\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$. Since $\hat{y}_i^r$ is also used as a covariate in $\mathbf{x}_i$ in the FH model for $\tilde{y}_i^a$, $\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$ will be nonzero. To this end, two analytic approximations for the standard errors of the discontinuities are proposed. The first approach combines the design-based variance estimate of the direct estimator of the regular survey $(\mathrm{Var}(\hat{y}_i^r))$ with the posterior

variance of the HB domain predictions of the alternative survey $\left(\mathrm{MSE}\left(\tilde{y}_i^a\right)\right)$ and a design-based estimator for the covariance between both point estimates $\left(\mathrm{Cov}\left(\hat{y}_i^r,\ \tilde{y}_i^a\right)\right)$. This approach is unstable in the sense that even negative variance estimates occur in the case of strong positive covariance estimates. A related issue is that design-based and model-based variance approximations are combined in one uncertainty measure for the discontinuities. Therefore a second analytic approximation was proposed, where a design-based estimator for the variance of the HB domain predictions $\left(\mathrm{MSE}\left(\tilde{y}_i^a\right)\right)$ is derived and combined with the design-based variance for the direct estimator for the regular survey and the design-based covariance between both point estimates.

Several references to design-based mean squared error estimation in small area estimation can be found in the literature. Gonzalez and Waksberg (1973) introduced the concept of an average design-based mean squared error of a set of synthetic estimators and proposed an estimator that, however, can be unstable and take negative values. Marker (1995) proposed a more stable but biased estimator for the design-based mean squared error for small area estimates, which can also take negative values. Lahiri and Pramanik (2019) proposed a design-based estimator that cannot take negative values, following the concepts of an average design-based mean squared error, originally introduced by Gonzalez and Waksberg (1973). Rivest and Belmonte (2000) proposed an estimator for the mean squared error that measures the uncertainty with respect to the sampling design conditional on the random effects of the model and assuming normality of the sampling model. Rao, Rubin-Bleuer and Estevao (2018) and Pfeffermann and Ben-Hur (2018) also propose a model for the design-based mean squared error in small area estimators. Rao et al. (2018) estimate the model parameters through restricted maximum likelihood while Pfeffermann and Ben-Hur (2018) applies a bootstrap method.

The complications with variance estimation of domain discontinuities under a univariate FH model can also be circumvented by setting up a full Bayesian framework for the analysis of the domain discontinuities. Two approaches are proposed in this paper. The first approach is a bivariate FH model to model the direct estimates under the regular and alternative approach simultaneously, i.e., a bivariate area level model for the vector $\left(\hat{y}_i^r,\ \hat{y}_i^a\right)^t$. The random component of this model accounts for the correlation between the domain parameters under the regular and alternative approach. The precision of the estimated discontinuities is improved by increasing the effective sample size within the domains by means of cross-sectional correlations. In addition, a positive correlation between the random domain effects further decreases the standard error of the estimated discontinuities. The second approach uses a univariate FH model for the direct estimates of the discontinuities, i.e., a univariate FH model for $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$. This method is considered as a less complex alternative for the bivariate FH model. It is, however, anticipated that it is harder to construct good prediction models, since the available covariates from registers might be good predictors for the target variables of the sample survey but probably not for systematic differences between the differences of two estimates for the same variable obtained with different survey processes.

The univariate FH model proposed by van den Brakel et al. (2016) was applied to estimate domain discontinuities in five key target variable of the Dutch Crime Victimization Survey (CVS) using data

obtained in a parallel run where the regular survey is conducted at the regular sample size and the alternative survey at a sample size that is about one fourth of the regular sample size. In this paper the bivariate FH model and the univariate FH model for the domain discontinuities are also applied to the same redesign of the CVS. The results are compared with the univariate FH model proposed in van den Brakel et al. (2016).

Model selection in this paper is based on a step forward selection procedure that minimizes the WAIC criteria (Watanabe, 2010, 2013). To avoid selecting over-parameterized models, it is proposed to add covariates in a step forward selection procedure only if they decrease the WAIC by more than the standard error of the WAIC. This prevents selection of several covariates that only marginally improves the WAIC, resulting in models that tend to overfit the data.

The FH model (Fay and Herriot, 1979) is frequently applied in the context of small area estimation (Rao and Molina, 2015). FH models are particularly appropriate if auxiliary information is available at the domain level. Datta, Ghosh, Nangia and Natarjan (1996) employed a multivariate FH model fitted in an HB framework to estimate median income. Multivariate FH models fitted in a frequentist framework are considered in Gonzales-Manteiga, Lombardia, Molina, Morales and Santamaria (2008); Benavent and Morales (2016). Several authors provided time-series FH models to use sample information from previous editions of a survey as a form of small area estimation (Rao and Yu, 1994; Datta, Lahiri, Maiti and Lu, 1999; You and Rao, 2000; Estaban, Morales, Perez and Santamaria, 2012; Marhuenda, Molina and Morales, 2013). Pfeffermann and Burck (1990); Pfeffermann and Tiller (2006); van den Brakel and Krieg (2016); Bollineni-Balabay, van den Brakel, Palm and Boonstra (2017) are some examples of FH time-series models casted in a state-space framework. Boonstra and van den Brakel (2019) discuss how FH time series models can be expressed either in a state space frame work and fitted with the Kalman filter or alternatively expressed as time series multilevel models in an hierarchical Bayesian framework, and estimated using a Gibbs sampler.

The paper is structured as follows. In Section 2 the Crime Vicitimization Survey, the redesign and the set up of the parallel run are described. The bivariate FH model is explained in Section 3, including the HB framework and the model selection and evaluation approach. Results are presented in Section 4. The paper ends with a discussion in Section 5.

2. The crime victimization survey

The Dutch crime victimization survey (CVS) is a long-standing survey conducted by Statistics Netherlands at an annual frequency with the purpose to publish reliable figures about crime rates, safety feelings, and satisfaction about police performance in the Netherlands. The CVS is designed to provide reliable figures at the national level and at the level of police districts, which is a subdivision of the Netherlands in 25 regions. The CVS is based on a stratified simple random sampling design for people aged 15 years or older residing in the Netherlands. Strata are formed by police regions to control the

precision of these planned domain estimates. The sampling frame is based on the Dutch government's register of all residents in the Netherlands, called Municipal Basis Administration. The yearly sample of the regular CVS is designed such that about 19,000 respondents are observed. The sample is equally divided over the strata, such that about 760 observations are obtained in each stratum. The general regression (GREG) estimator (Särndal, Swensson and Wretman, 1992) is used to estimate population parameters at the national level and for police districts.

The CVS has been redesigned in 2008. The data collection changed from a mixed-mode design via computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) to a sequential mixed-mode design that starts with web interviewing (WI) and is followed up for nonrespondents with CAPI and CATI. In addition the questionnaire is changed to improve the wording as well as the order of the questions. To quantify discontinuities induced by this redesign, the regular survey used for official publication purposes was conducted in parallel with the alternative survey approach with a sample size of about 6,000 respondents. In this application, the regular approach was based on the new survey design using WI, CATI and CAPI and the alternative approach was based on the old design using CAPI and CATI data collection only. The sample design for the parallel run is based on stratified simple random sampling where police districts are the strata, using proportional allocation. This results in a sample design that is optimal to estimate figures at the national level but suboptimal for domain estimation.

This survey reports on many different outcome variables. In the present study five key survey variables are considered, see Table 2.1. Estimates for these variables at the national level under the regular and alternative survey are specified in Table 2.2. The sample size allocated to the alternative approach is sufficiently large to estimate discontinuites at the national level using the GREG estimator but insufficient to estimate discontinuities at the domain level of the 25 police districts. The direct estimates for the discontinuities at the national level are indeed significantly different from zero, contrary to the unweighted averages of the direct domain estimates and their standard errors. To obtain more precise predictions for domain discontinuities a model-based small area estimation method based on area level models (Fay and Herriot, 1979) is proposed in the next section.

Table 2.1
Five key CVS survey variables considered in the present study

Variable	Description
`nuisance`	Perceived nuisance in the neighborhood on a ten point scale; this includes nuisance by drunk people, neigbours, or groups of youngsters, harassment, and drug related problems.
`unsafe`	Percentage of people feeling unsafe at times.
`propvict`	Percentage of people saying to have been victim to property crime in the last 12 months.
`offtot`	Total number of offenses per 100 people.
`satispol`	Percentage of people satisfied with police at their last contact (if contact in last 12 months).

Table 2.2
GREG estimates for the regular and alternative survey approach averaged over districts and national level. Standard errors between brackets

Variable	Average over 25 police districts						National level					
	regular		alternative		$\hat{\Delta}$		regular		alternative		$\hat{\Delta}$	
offtot	42.29	(4.73)	33.28	(5.73)	9.01	(7.69)	43.79	(1.07)	34.09	(1.04)	9.70	(1.49)
unsafe	24.38	(2.03)	19.86	(2.87)	4.52	(3.57)	25.07	(0.44)	20.48	(0.52)	4.59	(0.68)
nuisance	1.61	(0.11)	1.28	(0.13)	0.33	(0.17)	1.67	(0.02)	1.34	(0.02)	0.33	(0.03)
satispol	60.61	(4.23)	55.58	(6.88)	5.04	(8.21)	59.88	(0.92)	55.10	(1.25)	4.78	(1.55)
propvict	12.55	(1.60)	9.78	(2.19)	2.78	(2.77)	13.02	(0.36)	10.32	(0.39)	2.70	(0.53)

3. Methods

3.1 Small area estimation for domain discontinuities

Testing hypotheses about differences between estimates of a finite population parameter observed under different survey processes implies the existence of measurement errors. Therefore a measurement error model is required to explain systematic differences between survey estimates for the same population parameter observed under two different survey approaches. Let θ_i denote the population parameter of domain $i = 1, \ldots, m$. Let y_i^r and y_i^a denote the observed value for θ_i in the case of a complete enumeration under the regular approach and alternative approach, respectively. Direct estimates for y_i^r and y_i^a are obtained with the GREG estimator based on the samples assigned to the regular and alternative survey and are denoted as $\hat{y}_i^r$ and $\hat{y}_i^a$ respectively.

The relation between the observed values under a complete enumeration and the real population parameter is: $y_i^q = \theta_i + \lambda_i^q, i = 1, \ldots, m, q = r, a,$ with λ_i^q the real measurement bias if θ_i is measured with survey approach q. Without any external information, it is not possible to estimate λ_i^q. From the sample data only the relative bias, say $\Delta_i = y_i^r - y_i^a = \lambda_i^r - \lambda_i^a$ is identifiable. Direct estimates for these discontinuities are obtained from the survey data as the contrast between the GREG estimates, i.e., $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$.

In the case of the Dutch CVS the sample size of the regular survey is large enough to obtain sufficiently precise direct estimates for the planned domains, since the sample is designed to publish official statistics for these domains. The sample assigned to the alternative survey for the parallel run has only a size of one third of the regular sample size, which is insufficient to obtain precise direct estimates for the planned domains. In an earlier paper (van den Brakel et al., 2016) univariate FH models were developed to obtain more precise predictions for the domain parameters observed with the small sample size assigned to the alternative survey approach using auxiliary variables derived from three different sources. The first source contains demographic variables derived from the Municipal Basis Administration (MBA), which is an administration of all people residing in the Netherlands. The second source contains related variables available in the Police Register of Reported Offences (PRRO). The third source, which is unique in the case of a parallel run, contains direct estimates for the same variables observed under the

regular survey, which are sufficiently precise at least for the planned domains like police districts. The direct estimates from the regular survey are often selected as auxiliary variables for these univariate FH models. This comes not as a surprise since these are survey estimates for the same population parameters. Although measured with a different survey process, strong positive correlations can be expected. Strong improvements of the precision of small domain prediction are indeed found if the set of potential auxiliary variables, i.e., from MBA and PRRO, is extended with the direct estimates from the regular CVS.

In this application the sampling error in the auxiliary variables that come from the regular CVS can be ignored in the FH model, since the sample size and therefore the sampling error for these domains is more or less equal for the domains (Ybarra and Lohr, 2008). This implies that the variance component of the random domain effects is inflated with the sampling error of the auxiliary variables, which is fine as long as the sampling error does not differ between domains. In most applications this is not the case and the methods proposed by Ybarra and Lohr (2008) should be used to account for sampling error in the auxiliary variables.

FH multilevel models can be fitted under a frequentist approach using EBLUP or under an HB approach (Rao and Molina, 2015). In van den Brakel et al. (2016) the HB approach is preferred over the EBLUP, since the strong auxiliary information in the fixed effect part of the model often results in zero estimates for the variance component of the random domain effects, giving too much weight to the synthetic regression part and too little weight to the direct estimates in the EBLUP, (Bell, 1999; Rao and Molina, 2015). This problem can also be overcome with adjusted maximum likelihood estimation, see e.g., Li and Lahiri (2010) and Hirose and Lahiri (2018).

Let $\tilde{y}_i^a$ denote the HB prediction for domain i under the alternative approach. Now domain discontinuites are obtained by $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. Using direct estimates of the regular survey as auxiliary variables in the fixed part of the FH model for the alternative survey considerably increases the complexity of the variance estimation for the discontinuities. The variance of $\tilde{\Delta}_i$ can be expressed as $\mathrm{Var}(\tilde{\Delta}_i) = \mathrm{Var}(\hat{y}_i^r) + \mathrm{MSE}(\tilde{y}_i^a) - 2\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$. The use of $\hat{y}_i^r$ or related sample estimates as auxiliary variables to predict $\tilde{y}_i^a$, results in non-zero values for $\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ that cannot be ignored. To approximate $\mathrm{Var}(\tilde{\Delta}_i)$, van den Brakel et al. (2016) proposed an approximately design-unbiased estimator for $\mathrm{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ and $\mathrm{Var}(\hat{y}_i^r)$, while the $\mathrm{MSE}(\tilde{y}_i^a)$ is approximated with the posterior variance of the HB domain predictions. A major disadvantage of this approach is that model-based and design-based uncertainty measures are intertwined. On the one hand, the MSE's for $\tilde{y}_i^a$ are approximated with their posterior variances. On the other hand, the covariances between $\hat{y}_i^r$ and $\tilde{y}_i^a$ are approximated from a design-based perspective. Consequently, naive application of this approach may give negative variance estimates for the discontinuities. This drawback has been solved using a design-based approximation for the $\mathrm{MSE}(\tilde{y}_i^a)$, resulting in a fully design-based approximation for the uncertainty of the estimated domain discontinuities.

In this paper a full HB framework for estimating domain discontinuities is proposed as an alternative by developing a bivariate FH model for the domain parameters observed under both the regular and

alternative approach. The advantage of this approach is that it improves the precision of both the direct estimates of the regular and alternative domain estimates by borrowing strength from other domains and both surveys. Negative variance estimates for the estimated domain discontinuities are precluded by definition under this multivariate HB framework. This method is compared with a simple alternative, namely a univariate FH model for the direct estimates of the discontinuities. As mentioned in the introduction, it is anticipated that it might be hard to find covariates in the available registers that explain the discontinuities. An advantage of both models is that they avoid the complications of accounting for sampling error in the auxiliary variables, which is necessary if the survey estimates of the regular survey are used as covariates in univariate FH models and the sampling error differs between domains.

3.2 Bivariate Fay-Herriot model

A bivariate version of the FH model (Fay and Herriot, 1979) starts with a measurement model for the two GREG estimates observed in each domain:

$$\hat{\mathbf{y}}_i = \mathbf{y}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.1}$$

with $\mathbf{y}_i = \left(y_i^r, y_i^a\right)^t$, $\hat{\mathbf{y}}_i$ a vector containing the GREG estimates of $\mathbf{y}_i$ and $\mathbf{e}_i = \left(e_i^r, e_i^a\right)^t$ a vector with the sampling errors of $\hat{\mathbf{y}}_i$ for which it is assumed that

$$\mathbf{e}_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \mathbf{\Psi}_i\right),\ i = 1, \ldots, m. \tag{3.2}$$

Here $\mathbf{0}_2$ is a 2 dimensional column vector with each element equal to zero. Since the sample for the regular and alternative survey are drawn independently, it is assumed that $\mathbf{\Psi}_i = \text{Diag}\left(\psi_i^r, \psi_i^a\right)$ where ψ_i^q is the design variance of $\hat{y}_i^q$. It is also assumed that these design variances are known although they are replaced by their estimates in practice. The true domain parameters are modelled with a multilevel model. For the fixed effects it is assumed that the regular and alternative approach share the same covariates. In the most general case the regression coefficients for the fixed part are different for both variables i.e., $y_i^q = \mathbf{x}_i^t \boldsymbol{\beta}^q + \nu_i^q$, with $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^t$ a p-vector with covariates of domain i, $\boldsymbol{\beta}^q$ a p-vector of regression coefficients, which are equal over the domains but might be different between the two survey approaches. It is assumed that $x_{i1} = 1$ corresponds to the intercept. Furthermore ν_i^q are random domain effects. This gives rise to the following bivariate multilevel model for the two domain parameters:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i,\ i = 1, \ldots, m, \tag{3.3}$$

where $\mathbf{X}_i = \mathbf{I}_2 \otimes \mathbf{x}_i^t$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{r^t}, \boldsymbol{\beta}^{a^t}\right)^t$, $\mathbf{I}_2$ a 2 dimensional identity matrix, and $\mathbf{v}_i = \left(\nu_i^r, \nu_i^a\right)^t$. For the random domain effects it is assumed that

$$\mathbf{v}_i \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \boldsymbol{\Sigma}\right),\ i = 1, \ldots, m, \tag{3.4}$$

with $\boldsymbol{\Sigma}$ a general 2×2 covariance matrix for the random domain effects. Inserting (3.3) into (3.1) gives:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.5}$$

with model assumptions (3.2) and (3.4).

Since the number of domains in this application is small, it is important to select parsimonious models. One way to reduce model complexity is to assume that the regression coefficients are equal for both survey approaches. In this case a dummy indicator, say δ_i, is introduced, which is equal to zero for the regular survey and equal to one for the alternative survey. In this case $x_{i1} = 1$ corresponds to the overall intercept and $x_{i2} = \delta_i$ is the indicator whose coefficient measures the differences between intercepts of the variables observed under both surveys. So $y_i^q = \mathbf{x}_i^t\boldsymbol{\beta} + \nu_i^q$ and in (3.3) $\mathbf{X}_i = \mathbf{x}_i^t$, and $\boldsymbol{\beta}$ a vector with the corresponding regression coefficients. As a result, two versions for the fixed effects are considered:

- `FE_uq`: A fixed effect model where the regular and alternative approach share the same covariates, but have different regression coefficients. In this case, domain discontinuities are given by

$$\Delta_i = \sum_{j=1}^{p} x_{i,j}\left(\beta_j^r - \beta_j^a\right) + \left(\nu_i^r - \nu_i^a\right). \tag{3.6}$$

- `FE_eq`: A more parsimonious version for the fixed effect component by assuming that the regression coefficients are equal for the regular and alternative approach. In this case domain discontinuities are given by

$$\Delta_i = -\beta_2 + \left(\nu_i^r - \nu_i^a\right), \tag{3.7}$$

  with β_2 the regression coefficient for $x_{i2} = \delta_i$.

The following covariance structures for the random domain effects are considered:

- `RE_f`: A full covariance matrix $\boldsymbol{\Sigma}$ for the random domain effects. Positive correlation between the random domain effects will further increase the precision of the estimates for the domain discontinuities since domain estimates borrow strength not only from different domains but also across the two surveys.
- `RE_d`: A diagonal covariance matrix with separate variances for the regular and alternative approach, i.e.: $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\sigma_r^2, \sigma_a^2\right)$. This covariance structure in combination with model `FE_uq` comes down to applying a univariate FH model to both surveys separately. In this case models only use sample information from other domains within the same survey but not across the two surveys to improve the precision of the estimates for domain discontinuities.
- `RE_s`: A diagonal covariance matrix with equal variances for the regular and alternative approach, i.e.: $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_2$.

3.3 Univariate Fay-Herriot model for domain discontinuities

The univariate FH model for the direct estimates of the discontinuities starts with defining a measurement error model for the GREG estimates of the discontinuities:

$$\hat{\Delta}_i = \Delta_i + z_i \tag{3.8}$$

with $\Delta_i = y_i^r - y_i^a$ the true discontinuity of domain i under a complete enumeration of the population under both approaches, $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$ the GREG estimate for Δ_i based on the parallel run and $z_i = e_i^r - e_i^a$ the sampling error of $\hat{\Delta}_i$. It is assumed that $z_i \simeq \mathcal{N}\left(0, \psi_i^r + \psi_i^a\right)$ and that the design variances of the sampling errors are known. For Δ_i the following linear model is assumed:

$$\Delta_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i, \tag{3.9}$$

with $v_i \simeq \mathcal{N}\left(0, \sigma_v^2\right)$. Inserting (3.9) into (3.10) gives:

$$\hat{\Delta}_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i + z_i. \tag{3.10}$$

3.4 Estimation of the bivariate Fay-Herriot model

The models developed in Subsections 3.2 and 3.3 are fitted with a HB approach using Markov Chain Monte Carlo (MCMC) sampling. In particular the Gibbs sampler is used. The following priors are used for the model parameters and hyperparameters. For the regression coefficients uniform improper priors are assumed, i.e., $\boldsymbol{\beta} \sim 1$. For the random domain effects a multivariate normal prior is used: $\mathbf{v} \,|\, \boldsymbol{\Sigma} \simeq \mathcal{N}\left(\mathbf{0}_{2m}, \mathbf{I}_m \otimes \boldsymbol{\Sigma}\right)$.

In the case of a full covariance matrix for the random domain effects, the prior for $\boldsymbol{\Sigma}$ is taken to be a scaled inverse Wishart distribution (O'Malley and Zaslavsky, 2008). This distribution is obtained by writing $\boldsymbol{\Sigma} = \mathrm{Diag}(\boldsymbol{\xi})\,\tilde{\boldsymbol{\Sigma}}\mathrm{Diag}(\boldsymbol{\xi})$, with $\boldsymbol{\xi} = \left(\xi^r, \xi^a\right)^t$ and assuming a standard normal distribution for ξ^r and ξ^a, i.e., $\xi^x \simeq \mathcal{N}(0, 1)$, $(x \in (a, r))$ and an inverse Wishart distribution for $\tilde{\boldsymbol{\Sigma}}$, i.e., $\tilde{\boldsymbol{\Sigma}} \simeq \mathrm{Inv-Wish}\left(v_v, \boldsymbol{\Phi}_v\right)$, with $v_v = d + 1$ degrees of freedom, with d the dimension of $\tilde{\boldsymbol{\Sigma}}$ which is equal to 2 in this application, and scale parameter $\boldsymbol{\Phi}_v = \mathbf{I}_2$. In the case of a diagonal covariance matrix for the random domain effects, the priors for σ_q (in the case of unequal variances) or σ (in the case of equal variances) are half-Cauchy distributions. These are more robust prior distributions than the more commonly used inverse chi-squared distribution (Gelman, 2006). The inverse chi-squared distribution might be informative, even in the case of small scale and shape parameters. In addition convergence problems might occur with the Gibbs sampler. Both problems are largely avoided with a redundant multiplicative parametrization of the random effects (Gelman, 2006; Gelman, Van Dyk, Huang and Boscardin, 2008; Polson and Scott, 2012). See van den Brakel and Boonstra (2018) for more details on the priors of this model.

Let $\hat{\mathbf{y}}$ denote the $2m$ column vector obtained by stacking the m column vectors $\hat{\mathbf{y}}_i$, and $\mathbf{X}$ the matrix obtained by stacking the matrices $\mathbf{X}_i$. In the case of unequal regression coefficients, $\mathbf{X}$ is a $2m \times 2p$ matrix. In the case of equal regression coefficients, $\mathbf{X}$ is a $2m \times p$ matrix. The likelihood function can be written as

$$p\left(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \boldsymbol{\Psi}\right), \tag{3.11}$$

with $\boldsymbol{\Psi} = \oplus_{i=1}^m \boldsymbol{\Psi}_i$ a $2m \times 2m$ diagonal matrix with the design variances of the direct estimates $\hat{\mathbf{y}}$ and $\boldsymbol{\theta}$ a vector containing all model parameters. The joint prior distribution $p(\boldsymbol{\theta})$ equals the product of the aforementioned priors. The posterior distribution of $\boldsymbol{\theta}$ is proportional to the joint density, i.e., $p\left(\boldsymbol{\theta} \,|\, \hat{\mathbf{y}}\right) \propto p(\boldsymbol{\theta})\, p\left(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta}\right)$. The model is fitted using the Gibbs sampler Geman and Geman (1984);

Gelfand and Smith (1990). The full conditional distributions used in the Gibbs sampler are specified in van den Brakel and Boonstra (2018).

For each model considered, the Gibbs sampler is run in three independent chains with randomly generated starting values. The length of each chain after the burn-in period for each run is 10,000 iterations. This gives 30,000 draws to compute estimates and standard errors. The convergence of the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. The diagnostics suggest that all chains converge well within 500 draws. The estimated Monte Carlo simulation errors are small compared to the posterior standard errors for all parameters, so that the number of draws are more than sufficient for our purposes.

The estimands of interest are expressed as functions of the parameters, and applying these functions to the MCMC output for the parameters results in draws from the posteriors for these estimands. Domain predictions for the target variables under the bivariate FH model are obtained as the posterior means approximated by the Gibbs sampler output and are denoted as $\tilde{y}_i^{q,b\mathrm{FH}}$. Domain predictions for the discontinuities are obtained as the posterior means of (3.6) or (3.7) approximated by the Gibbs sampler output and are denoted as $\tilde{\Delta}_i^{b\mathrm{FH}}$. Mean squared errors for $\tilde{y}_i^{q,b\mathrm{FH}}$ and $\tilde{\Delta}_i^{b\mathrm{FH}}$ are obtained as posterior variances approximated from the Gibbs sampler output.

The methods are implemented in R using the `mcmcsae` R-package (Boonstra, 2020).

3.5 Pooling design variances

Estimates for the design variances ψ_i^r and ψ_i^a are available from the GREG estimator and are used as if the true design variances are known. This is a standard assumption in small area estimation. Therefore it is important to provide reliable estimates for these design variances. For the regular survey the variance estimates of the GREG estimates are considered to be reliable enough to be used in the FH model. For the alternative survey the estimates of the design variances are unreliable and therefore smoothed to improve their stability of the estimates of ψ_i^a. Under the assumption that the population variances of the GREG residuals under the alternative approach are equal accross domains, the analysis-of-variance type of pooled variance estimator is used:

$$\psi_i^a = \frac{1 - f_i^a}{n_i^a} \frac{1}{n^a - m} \sum_{i=1}^{m} \left(n_i^a - 1\right) S_{i;\mathrm{GREG}}^{a^2}$$

with f_i^a the sample fraction in domain i of the alternative survey, n_i^a the number of respondents in domain i under the alternative survey, $n^a = \sum_{i=1}^{m} n_i^a$, and $S_{i;\mathrm{GREG}}^{a^2}$ the estimated population variance of the GREG residuals.

Alternatively, variance estimates of the direct estimates can be smoothed by modeling the variance estimates along with the GREG estimates themselves, (You and Chapman, 2006) and Sugasawa, Tamae and Kubokawa (2017). Their approach can be traced back to Arora and Lahiri (1997). Another possibility

is to smooth variance estimates by applying generalized variance functions, (Wolter (2007), Chapter 7, and Hawala and Lahiri (2018)).

3.6 Model selection and evaluation

Frequently applied model selection criteria in HB settings are the Widely Applicable Information Criterion or Watanabe-Akaike Information Criteria (WAIC) (Watanabe, 2010, 2013) and the Deviance Information Criteria (DIC) (Spiegelhalter, Best, Carlin and van der Linde, 2002). They are popular because they are easy to compute from MCMC simulation output and because of their ability to make a reasonable tradeoff between model fit and model complexity. The WAIC is seen as an improvement on the DIC since the latter can produce negative estimates for the effective number of parameters and it is not defined for singular models (Vehtari, Gelman and Gabry, 2017). The penalty used for model complexity in DIC and WAIC is closely related to the effective number of parameters proposed by Hodges and Sargent (2001) for linear multilevel models where each fixed effect contributes one degree of freedom and the random effects contribute a value in the range between zero and m, depending on the size of the variance component. As follows from the definition of WAIC, models with lower WAIC values are preferred. The WAIC estimates are uncertain and an approximation of its standard error is provided by Vehtari et al. (2017) equation (23) and can be computed using R package `loo` (Vehtari, Gelman and Gabry, 2015).

Covariates are selected from the set of auxiliary variables listed in van den Brakel and Boonstra (2018) using a step-forward selection procedure. Various models are compared using the aforementioned WAIC estimates. From the set of potential covariates, the covariate with the lowest WAIC value is selected in the model. This selection process is iteratively repeated as long as adding a new covariate further decreases the WAIC value. In this application, this step-forward selection procedure, further abbreviated as `step-WAIC`, often results in models with a large number of covariates. Since the WAIC values are estimates that contain error, it appears that it might not be desirable to minimize the WAIC by adding covariates to the model as long as it reduces the point estimates of the WAIC. As an alternative we applied a step-forward selection procedure where covariates are added to the model as long as a new covariate decreases the WAIC with a value that exceeds the estimated standard error of the WAIC. This method will be referred to as `step-WAIC-se`.

The step-forward selection procedure is applied to each of the six different combinations of the two fixed effect versions (`FE_uq` and `FE_eq`) and the three covariance structures of the random component (`RE_f`, `RE_d`, and `RE_s`). From the resulting six models the one with the lowest WAIC value is selected. Model adequacy of these six selected models is evaluated with posterior predictive checks. This implies that replicate data sets, simulated from the posterior predictive distribution are compared with the originally observed data to study systematic discrepancies and to evaluate how well the selected model fits the observed data (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin, 2004). Posterior predictive p-values are calculated for six different tests that evaluate particular aspects of the posterior predictive

distribution. Posterior predictive p-values for the domain discontinuities are defined as $p = P\left(T\left(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta}\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}\right) \middle| \hat{\mathbf{y}}\right)$, where $\hat{\boldsymbol{\Delta}}^{\text{sim}} = \left(\hat{\Delta}_1^{\text{sim}}, \ldots, \hat{\Delta}_m^{\text{sim}}\right)^t$ are replicates of the observed discontinuities for the m domains under the posterior predictive distribution, $\hat{\boldsymbol{\Delta}} = \left(\hat{\Delta}_1, \ldots, \hat{\Delta}_m\right)^t$ the observed direct estimates for the m domain discontinuities and $T(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta})$ (or $T(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta})$) a test statistic that depends on $\hat{\boldsymbol{\Delta}}^{\text{sim}}$ (or $\hat{\boldsymbol{\Delta}}$) and unknown true values for the m domain discontinuities $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_m)^t$. Posterior predictive p-values are estimated from the Gibbs sampler output as the average over the S Monte Carlo samples

$$\hat{p} = \frac{1}{S} \sum_{s=1}^{S} I\left(T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)\right), \tag{3.12}$$

with $I(A)$ the indicator function with value one if the condition A is fulfilled and zero otherwise, $\hat{\boldsymbol{\Delta}}^s = \left(\hat{\Delta}_1^s, \ldots, \hat{\Delta}_m^s\right)^t$ the m observed domain discontinuities in the sth replicate of the MCMC simulation and $\boldsymbol{\Delta}^s = \left(\Delta_1^s, \ldots, \Delta_m^s\right)^t$ the true values of the m domain discontinuities in the sth replicate of the MCMC simulation. If a model fits the observed data adequately, then it is expected that $T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)$ is in the bulk of the histogram of the replicates $T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right)$. Therefore p-values close to zero or one are indications of a poor fit with respect to that test statistic. In the expressions below, it is understood that T_x, for $x = 1, \ldots, 6$, is a function of $\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right)$ or $\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)$, depending on the component that is evaluated in (3.12). The following posterior predictive tests are defined (You, 2008):

1. A general goodness-of-fit test statistic $T_1 = \sum_{i=1}^m \left(\hat{\Delta}_i - \Delta_i\right)^2 \Big/ \text{Var}\left(\hat{\Delta}_i \middle| \Delta_i\right)$. Here $\text{Var}\left(\hat{\Delta}_i \middle| \Delta_i\right) = \psi_i^r + \psi_i^a$.
2. $T_2 = \max\left(\hat{\Delta}_i\right)$ and $T_3 = \min\left(\hat{\Delta}_i\right)$ which are sensitive for deviations in the tails of the distribution.
3. $T_4 = \frac{1}{m} \sum_{i=1}^m \hat{\Delta}_i \equiv \bar{\hat{\Delta}}$, i.e., the mean which is sensitive for bias in the domain predictions.
4. $T_5 = \frac{1}{m-1} \sum_{i=1}^m \left(\hat{\Delta}_i - \bar{\hat{\Delta}}\right)^2$, i.e., the variance of the domain estimates, which is sensitive for e.g., overshrinkage.
5. $T_6 = \left|\max\left(\hat{\Delta}_i\right) - \bar{\Delta}\right| - \left|\min\left(\hat{\Delta}_i\right) - \bar{\Delta}\right|$, with $\bar{\Delta} = \frac{1}{m} \sum_{i=1}^m \Delta_i$ which is sensitive to asymmetry in the distribution.

4. Results

4.1 Model selection

In Subsection 3.2, two different versions for the fixed effects (`FE_uq` and `FE_eq`) and three different covariance structures of the random effects (`RE_f`, `RE_d` and `RE_s`) are considered for the bivariate FH model. The step-forward selection procedure from Subsection 3.6 is applied to each of these six combinations separately to select covariates. Recall from Subsection 3.1 that for the bivariate FH model

and the univariate FH model for the discontinuities, potential covariates are available from the Municipal Base Administration and the Police Register of Reported Offences. Names of these covariates start with MBA_ and PR_ respectively. For the univariate FH model for the alternative survey, direct estimates from the regular CVS are also considered as covariates (van den Brakel et al., 2016). Names of these covariates start with CVSR_. See the appendix for an overview of the covariates.

The finally selected models for the bivariate FH model are summarized in Table 4.1. The models presented in Table 4.1 are selected with the step-WAIC-se procedure. The step-WAIC procedure selects models with a substantially larger amount of covariates which improve the WAIC only marginally. For offtot and unsafe the step-WAIC result in a model with 4 covariates with unequal regression coefficients and diagonal covariance matrices for the random effects (RE_d and RE_s). For satispol and propvict the step-WAIC result in a model with respectively 3 and 2 covariates with unequal regression coefficients, also with diagonal covariance matrices with equal variances for the random effects (RE_s). With only 25 domains, there is a substantial risk that these models overfit the data. An exception is nuisance, where both selection procedures result in the same model.

With the step-WAIC-se procedure more parsimonious models are obtained as follows from Table 4.1. For total offences, offtot and nuisance, a model with only one covariate with equal regression coefficients for both surveys (FE_eq) is obtained in combination with a full covariance matrix (RE_f) with large random domain effects with a strong positive correlation of 0.98 for offtot and 0.81 for nuisance. Also for unsafe a more parsimonious model, with one covariate and equal regression coefficients (FE_eq) is obtained with the step-WAIC-se procedure. In this case a diagonal covariance matrix with equal variances (RE_s) is selected. For propvict and satispol the step-WAIC-se procedure avoids the selection of large amounts of covariates, found with the step-WAIC approach. The selected model for propvict has a full covariance matrix with a weak positive correlation of 0.1 (RE_f), and one covariate with unequal regression coefficients (FE_uq). The model for satispol has a diagonal covariance matrix with equal variances (RE_s) and one covariate with unequal regression coefficients (FE_uq). Since parsimonious models are preferred in this application, the models obtained with the step-WAIC-se approach are finally selected. See van den Brakel and Boonstra (2018) for a more detailed discussion of the model selection resulting in the finally selected models.

The models selected with the univeriate FH model for the direct estimates of the discontinuities, developed in Subsection 3.3, are summarized in Table 4.2. The models are selected with the step-WAIC-se procedure. For unsafe the step-WAIC results in a model with four covariates. For the other variables the same models are selected as with the step-WAIC-se procedure. The univariate FH models developed in van den Brakel et al. (2016) for the alternative survey approach are summarized in Table 4.3.

Standard model diagnostics test the underlying assumptions that the random domain effects and the residuals are normally and independently distributed. Since the number of domains in this application is

small, the power of the tests for normality are weak and do not indicate deviations from normality. Therefore the posterior predictive tests as summarised in Subsection 3.5 are used to evaluate the model adequacy. In addition, the domain predictions aggregated to the national level are compared with the direct estimates at the national level to evaluate the bias introduced with the small area estimation procedures in Subsection 4.2. The posterior predictive p-values for the domain estimates of the target variables and the discontinuities are summarized in Table 4.4 for the bivariate FH model and Table 4.5 for the univariate FH model for the discontinuities. The general measure for goodness-of-fit (T_1) indicates that the fit for the discontinuities of `offtot` is of reduced quality (other models considered had similar high values). The values for the bivariate FH model are slightly better compared to the univariate FH model for the discontinuities. The posterior predictive p-values for maximum (T_2) and minimum (T_3) values do not indicate problems with the tails of the distributions. For these posterior predictive p-values there are no systematic differences between bivariate and univariate FH model. The values for T_4, T_5, and T_6 for the discontinuities of the bivariate model are comparable with the values for the univariate model. The posterior predictive values for the mean (T_4) and asymmetry of the distribution (T_6) indicate that the distributions are symmetrically concentrated around their mean. The posterior predictive p-values for the variance (T_5) indicate some undershrinkage for the discontinuities of `nuisance`, `propvict`, and `offtot` under both the bivariate and univariate FH model.

Table 4.1
Final models bivariate FH model selected with `step-WAIC-se`. All models contain an intercept. ρ: correlation between the random effects

Variable	Model	Covariance structure random effects			
		type	σ_r^2	σ_a^2	ρ
offtot	FE_eq: δ_i + PR_weapon	RE_f	8.77	5.32	0.98
unsafe	FE_eq: δ_i + PR_propcrim	RE_s	1.17	1.17	-
nuisance	FE_eq: δ_i + MBA_immigrnw	RE_f	0.20	0.14	0.81
satispol	FE_uq: MBA_immigr	RE_s	0.78	0.78	-
propvict	FE_uq: PR_propcrim	RE_f	0.79	0.39	0.2

Table 4.2
Final models univariate FH model for direct estimates of the discontinuities. All models contain an intercept

Variable	Model	Variance random effects (σ_v^2)
offtot	PR_propcrim	0.80
unsafe	MBA_benefit	1.03
nuisance	PR_threat	0.038
satispol	MBA_benefit	0.928
propvict	PR_assault	0.485

Table 4.3
Final models univariate FH model for alternative CVS from van den Brakel et al. (2016). All models contain an intercept

Variable	Model	Variance random effects (σ_v^2)
offtot	CVSR_victim	0.003
unsafe	CVSR_nuisance + MBA_benefit + PR_propcrim + PR_drugs	2.997
nuisance	CVSR_nuisance + MBA_old	0.805
satispol	CVSR_funcpol	4.995
propvict	PR_propcrim + MBA_old	7.725

Table 4.4
Posterior predictive *p*-values for the final multivariate FH models from Table 4.1

Variable	T_1	T_2	T_3	T_4	T_5	T_6
	Discontinuities					
offtot	0.980	0.797	0.069	0.337	0.968	0.416
unsafe	0.343	0.841	0.833	0.454	0.437	0.912
nuisance	0.927	0.940	0.034	0.345	0.988	0.465
satispol	0.772	0.595	0.392	0.610	0.762	0.484
propvict	0.925	0.261	0.029	0.258	0.970	0.070
	Target variables					
offtot	0.859	0.249	0.024	0.317	0.524	0.089
unsafe	0.308	0.779	0.492	0.420	0.474	0.708
nuisance	0.766	0.317	0.108	0.433	0.504	0.156
satispol	0.742	0.929	0.584	0.457	0.875	0.797
propvict	0.695	0.339	0.168	0.379	0.655	0.194

Table 4.5
Posterior predictive *p*-values for the final univeriate FH models for direct estimates of the discontinuities from Table 4.2

Variable	T_1	T_2	T_3	T_4	T_5	T_6
	Discontinuities					
offtot	0.985	0.828	0.071	0.390	0.972	0.438
unsafe	0.382	0.885	0.816	0.464	0.523	0.920
nuisance	0.970	0.941	0.072	0.434	0.978	0.554
satispol	0.814	0.607	0.378	0.607	0.783	0.488
propvict	0.946	0.272	0.052	0.383	0.963	0.092

4.2 Estimation results

In this Subsection estimation results for the three different modelling approaches are discussed. In Subsection 4.2.1 the HB predictions for the target variables under the regular and alternative survey obtained with the bivariate FH model are compared with the direct estimates and with the domain predictions obtained with the univariate FH model where the direct estimates of the regular approach are

potential auxiliary variables in the model selection. Subsequently results for the domain discontinuities are discussed in Subsection 4.2.2. Here the results obtained with the univariate FH model for the discontinuities are also discussed.

With model-based small area estimation, the design variance of the direct estimators is reduced at the cost of accepting some amount of design bias. To evaluate differences in the direct point estimates and the small domain predictions, the following two measures are defined. The first one is the Mean Relative Difference (MRD), which summarizes the differences between the direct estimates and the domain predictions:

$$\text{MRD} = \frac{100\%}{m} \sum_{i=1}^{m} \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q},\ q = r,\ a, \tag{4.1}$$

and $\tilde{y}_i^q$ is the domain prediction based on the bivariate FH model or the univariate FH model. The second measure is the Absolute Mean Relative Difference (AMRD) between the direct estimate and the domain prediction, which is defined as:

$$\text{AMRD} = \frac{100\%}{m} \sum_{i=1}^{m} \left| \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q} \right|,\ q = r,\ a, \tag{4.2}$$

the increased precision of the small domain predictions is measured with Mean Relative Difference of the Standard Errors (MRDSE) between the direct estimates and the domain predictions and is defined as

$$\text{MRDSE} = \frac{100\%}{m} \sum_{i=1}^{m} \frac{\text{SE}\left(\hat{y}_i^q\right) - \text{SE}\left(\tilde{y}_i^q\right)}{\text{SE}\left(\hat{y}_i^q\right)},\ q = r,\ a, \tag{4.3}$$

these measures are defined in a similar way for the estimates and predictions of the domain discontinuities $\hat{\Delta}_i$ and $\tilde{\Delta}_i$.

4.2.1 Results for variables under the regular and alternative survey

In Table 4.6 the domain predictions and their standard errors averaged over the domains as well as the MRD, AMRD and MRDSE are given for the alternative survey under the univariate FH model with the models presented in Table 4.3. Results under the bivariate FH model, based on the final models of Table 4.1, are presented in Table 4.7 for the variables under the alternative survey and in Table 4.8 for the variables under the regular survey. Comparing the standard errors (SE) and the MRDSE in Table 4.6 and Table 4.7 shows that the bivariate FH model results in stronger reductions of the standard errors for all variables with the exception of `nuisance`. This comes at the cost of an increased bias. Comparing MRD and AMRD in both tables shows that the deviations between the direct estimates and the small area predictions are larger under the bivariate FH model. Comparing the SE and MRDSE in Tables 4.7 and 4.8 shows that the improvement in precision with the bivariate FH model for the regular survey is smaller, as expected since the sample size of the regular survey is larger. The bias in the bivariate FH model

predictions for the regular survey are also smaller, which follows from a comparison of MRD and AMRD in Tables 4.7 and 4.8.

The domain predictions under the univariate and bivariate FH model are plotted against the GREG estimates in Figures 4.1 through 4.5. The graphs also contain the GREG estimate at the national level versus the domain predictions aggregated to the national level according to (4.4). Figures 4.1 and 4.3 show that there is only a small amount of shrinkage for `offtot` and `nuisance`. Figure 4.2 shows for `unsafe` that the bivariate FH model shrinks the domain predictions for the alternative survey while the amount of shrinkage for the univariate FH model for the alternative CVS and the bivariate FH model for the regular survey is smaller. For `propvict`, see Figure 4.4, there is a small difference between the amount of shrinkage of the alternative CVS under the bivariate and univariate model. From Figure 4.5 it follows that the bivariate FH model for `satispol` cannot adequately model the observations under the alternative survey with the auxiliary information from the two registers (MBA and PRRO). In this case the domain predictions of `satispol` under the alternative approach display extreme overshrinkage. The univariate FH model indeed selects the same auxiliary variable from the regular survey only, see Table 4.3 and results in more realistic domain predictions.

For variables related to opinions and views such as `unsafe` and `satispol`, the reduction in the standard errors is accompanied by a relatively strong increase in the bias. This is especially the case with the small area prediction of the bivariate FH model for the alternative survey. For these variables, there are no strongly correlated covariates in the MBA and PRRO. In these cases the univariate FH model performs better since related covariates from the regular survey are selected (see Table 4.3), while the bivariate model doesn't detect correlation between the random effects (see Table 4.1).

Table 4.6
Average of domain predictions alternative survey with univariate FH model from van den Brakel et al. (2016)

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	33.21	2.90	-0.44	7.03	47.74
unsafe	19.83	1.64	-0.96	7.58	41.16
nuisance	1.29	0.08	-0.74	5.02	37.96
satispol	55.09	2.54	-0.11	6.43	61.98
propvict	9.85	0.84	-3.17	11.86	60.69

Table 4.7
Average of domain predictions alternative survey with bivariate FH model

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	33.26	2.82	-0.99	6.93	49.36
unsafe	19.82	1.21	-2.54	11.97	56.47
nuisance	1.28	0.08	-0.98	4.28	35.72
satispol	55.08	1.97	-0.49	8.97	70.06
propvict	9.91	0.73	-4.81	14.70	65.35

Table 4.8
Average of domain predictions regular survey with bivariate FH model

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	41.34	3.76	0.96	4.56	17.95
unsafe	24.22	1.07	-0.01	6.02	46.78
nuisance	1.60	0.09	0.38	2.62	15.93
satispol	60.82	1.47	-0.77	5.38	64.83
propvict	12.18	0.88	1.58	7.84	43.70

The direct estimates at the national level are accurate estimates since they are based on sufficiently large sample sizes. Therefore the bias in model-based domain predictions is often assessed by comparing the direct estimates at the national level with the domain predictions aggregated to the national level. The target variables in this application are all defined as population means. Therefore the aggregated domain predictions are obtained as the average over the domains weighted with the relative domain sizes,

$$\tilde{y}^q = \sum_{i=1}^{m} \frac{N_i}{N} \tilde{y}_i^q \tag{4.4}$$

with N_i the population size of domain i and N the size of the total population.

Table 4.9 compares the weighted average of the domain predictions according to (4.4) with the national GREG estimates. For the univariate FH model for the alternative CVS, the aggregated domain predictions are almost exactly equal to the GREG estimates at the national level. For the bivariate FH model the differences are slightly larger but the aggregated domain predictions are still very close to the GREG estimates at the national level. The largest relative difference amounts to 3% and is observed for offtot under the regular survey.

Table 4.9
GREG estimates national level and aggregated HB predictions regular and alternative survey approach (4.4)

Variable	Regular		Alternative			Discontinuity			
	GREG	biv. FH	GREG	biv. FH	uni. FH	GREG	biv. FH	uni. FH	Δ FH *)
offtot	43.79	42.47	34.09	34.02	34.09	9.7	8.45	9.7	9.04
unsafe	25.07	24.89	20.48	20.49	20.48	4.59	4.40	4.59	4.69
nuisance	1.67	1.66	1.34	1.34	1.34	0.33	0.32	0.34	0.33
satispol	59.88	60.36	55.10	55.06	55.12	4.78	5.29	5.04	5.07
propvict	13.02	12.76	10.32	10.33	10.32	2.70	2.43	2.70	2.63

*) : Δ FH are the HB predictions with univariate FH model for the direct estimates of the discontinuities, weighted similarly to (4.4).

Figure 4.1 Domain estimates GREG versus HB predictions offtot. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

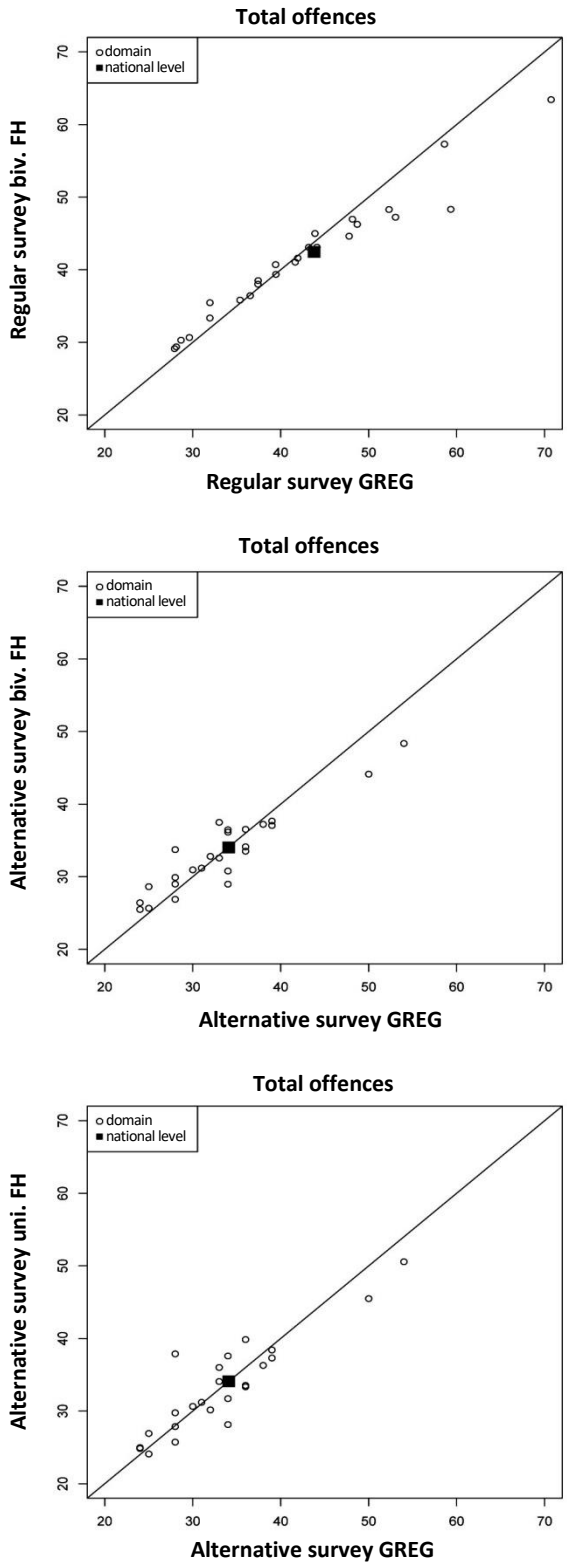


Figure 4.2 Domain estimates GREG versus HB predictions unsafe. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

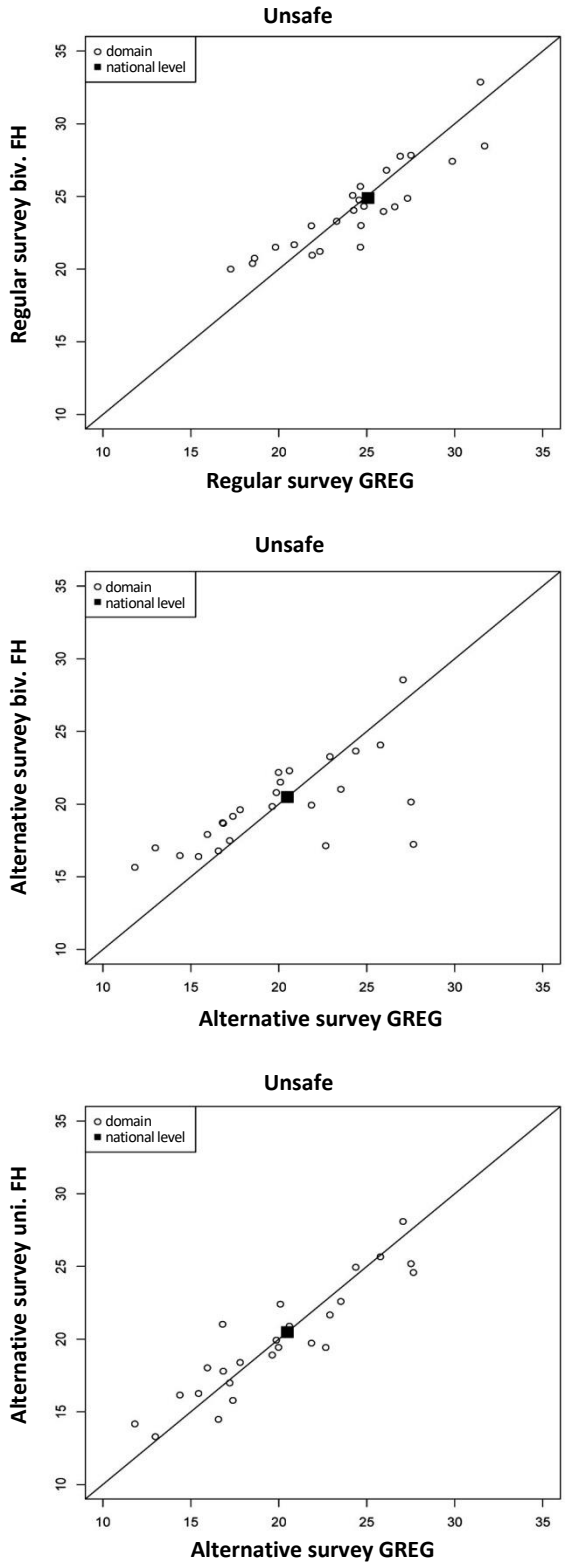


Figure 4.3 Domain estimates GREG versus HB predictions nuisance. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

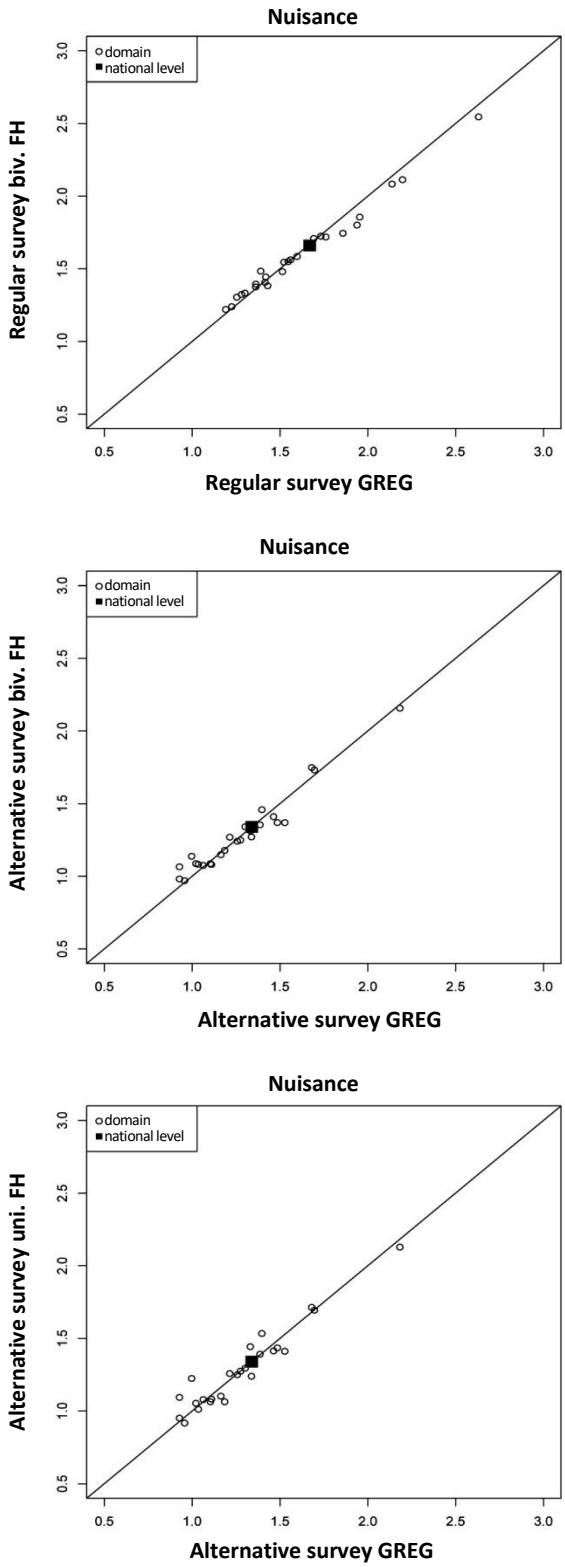


Figure 4.4 Domain estimates GREG versus HB predictions propvict. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

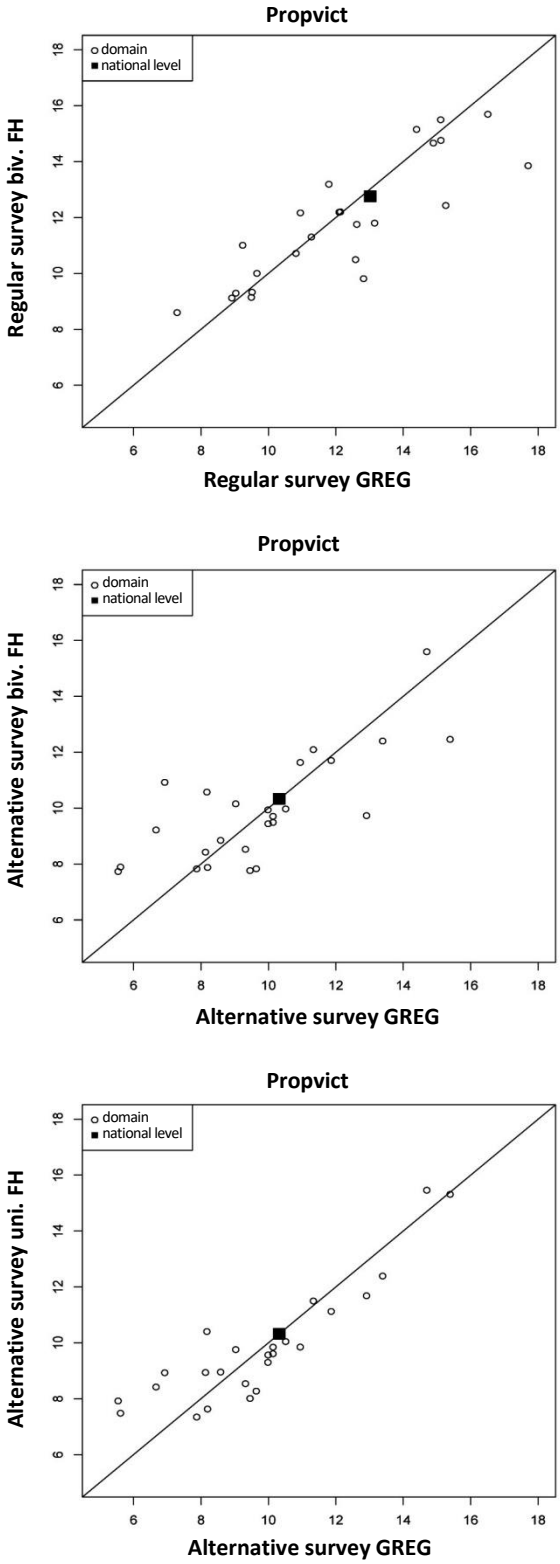
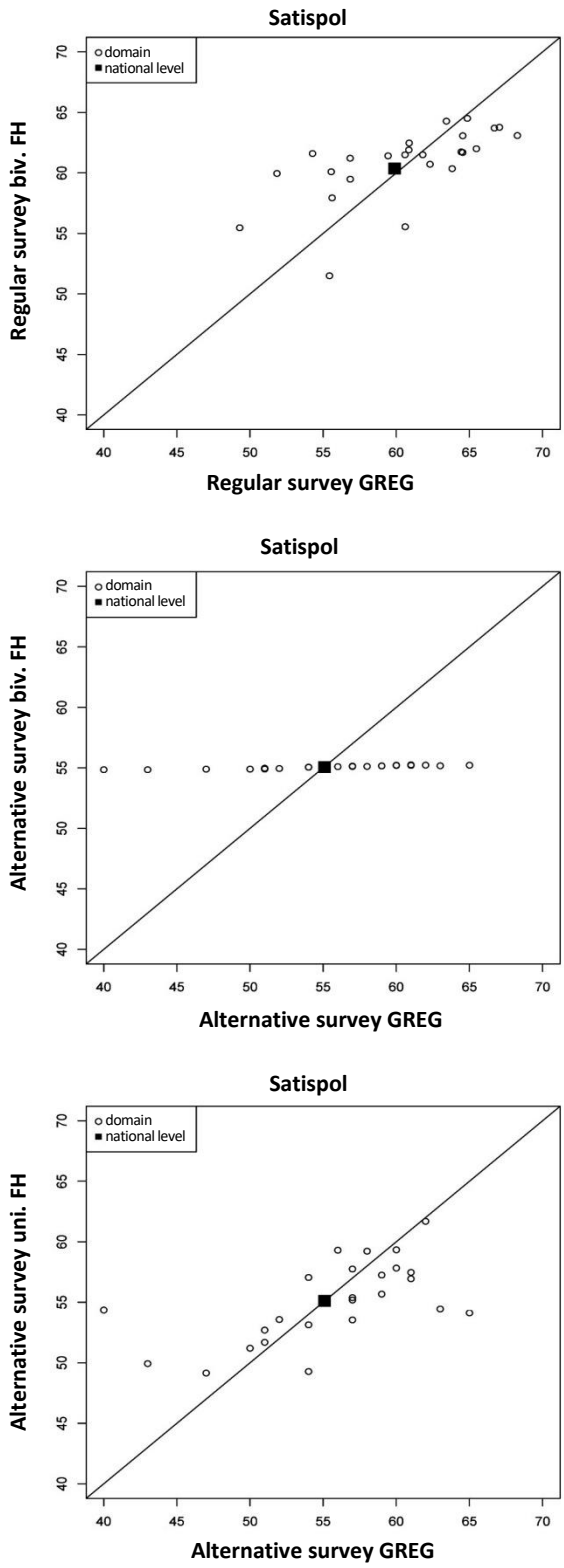


Figure 4.5 Domain estimates GREG versus HB predictions satispol. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).



4.2.2 Results for discontinuity estimates

In the last four columns of Table 4.9, the GREG estimates for the discontinuities at the national level are compared with the domain predictions obtained with the univariate FH model for the alternative CVS, the bivariate FH model and the univariate FH model for the discontinuities, aggregated to the national level using (4.4). The differences between the GREG estimates for the discontinuities at the national level and the aggregated domain predictions are the largest for the bivariate model and the smallest for the univariate FH model for the alternative CVS. This can be expected since the bivariate FH model shrinks both the domain estimates for the regular and alternative survey. With the univariate FH model for the alternative CVS, only the estimates for the alternative survey are replaced by domain predictions, while the estimates for the regular survey are not adjusted. In addition the domain predictions for the alternative survey have larger MRD's and AMRD's under the bivariate FH model compared to the univariate FH model (compare Table 4.6 and 4.7). The differences for the univariate FH model for the discontinuities are smaller compared to the bivariate FH model but larger compared to the univariate FH model for the alternative CVS.

In Tables 4.10, 4.11, and 4.12 the domain predictions and their standard errors for the discontinuities averaged over the domains as well as the MRD and MRDSE are summarized for the univariate FH model for the alternative CVS, bivariate FH model and the univariate FH model for the discontinuities respectively. The MRD's are large because the GREG estimates for the discontinuities in the denominator of (4.11) frequently take values close to zero, which make these indicators unstable. Therefor the AMRD is replaced by the median of the absolute relative differences, $\left|\left(\hat{y}_i^q - \tilde{y}_i^q\right)/\hat{y}_i^q\right|$, and is abbreviated as MARD. The latter are indeed more stable indicators for bias. The MARD is the smallest for the univariate FH model for the alternative CVS, since this approach only adjusts the domain predictions of the alternative CVS. The MARD values for the bivariate FH model on their turn are smaller than those for the univariate FH model for the discontinuities.

With the exception of `nuisance` the standard errors for the domain predictions under the bivariate FH model are smaller compared to the univariate FH model for the alternative CVS. In the case of `propvict` and `offtot`, this is the result of slightly more precise domain predictions for the alternative survey with respect to the univariate FH model (compare Table 4.6 with 4.7), a clear improvement in precision of the domain predictions of the regular survey compared to the GREG estimators (Table 4.8 and 2.2) and the positive correlation between the random effects. In the case of `satispol` and `unsafe` this is mainly the result of a clear improvement of precision of the domain predictions with the bivariate FH model for the regular compared to the GREG estimators (Table 4.8 and 2.2) and also a clear improvement of the precision of the domain predictions with the bivariate FH model for the alternative survey compared to the univariate model (compare Table 4.6 with Table 4.7).

For all five variables, the smallest standard errors are obtained with the univariate FH model for the discontinuities. This comes at the cost of a larger bias, as illustrated with the MARD values. An exception is `satispol`, for which the bias in terms of MARD for the bivariate FH model is clearly larger than the univariate FH model for the discontinuities. For this variable the bias is the lowest with the univariate FH model for the alternative CVS, but the reduction of the standard errors is also smaller.

The last columns of Tables 4.11 and 4.12 contain the shrinkage factors for the domain discontinuities averaged over the domains. For the univariate FH model for the discontinuities the shrinkage factors for the predictions of the domain discontinuities, i.e., the weights attached to the direct estimator for the discontinuities, are defined as $\gamma_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \left(\psi_i^r + \psi_i^a\right)\right)$. For the bivariate model the shrinkage factors for the predictions of the domain discontinuities are defined as $\gamma_i = \boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t \big/ \left(\boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t + \left(\psi_i^r + \psi_i^a\right)\right)$, with $\boldsymbol{\iota} = (1, -1)^t$. The average shrinkage factor is defined as $\bar{\gamma} = 1 \big/ M \sum_{i=1}^{M} \gamma_i$. Note that this statistic is not available for the discontinuities obtained with the univariate FH model for the alternative CVS, since under this approach domain discontinuities are obtained as the contrast between the GREG estimate for the regular survey and the domain prediction for the alternative approach. With the exception of `satispol`, the shrinkage factors under the univariate FH model for discontinuities are a factor 10 smaller compared to those of the bivariate FH model. The question rises whether the extremely small shrinkage factors of the univariate FH model for the discontinuities overshrink the direct estimates of the discontinuities.

Table 4.10
Domain predictions for discontinuities univariate FH model for the alternative CVS

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)
`offtot`	9.08	3.92	-5.67	22.14	48.47
`unsafe`	4.55	2.46	42.98	23.44	29.45
`nuisance`	0.33	0.07	-5.80	11.49	57.49
`satispol`	5.52	4.72	99.26	47.72	40.86
`propvict`	2.70	1.83	-142.60	30.49	32.75

Table 4.11
Domain predictions for discontinuities bivariate FH model

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)	$\bar{\gamma}$
`offtot`	8.07	2.68	1.94	23.34	63.60	0.208
`unsafe`	4.40	1.56	55.49	41.24	55.09	0.317
`nuisance`	0.31	0.09	-4.01	18.43	44.47	0.327
`satispol`	5.74	2.46	228.50	73.43	68.75	0.019
`propvict`	2.27	1.10	-113.00	27.11	59.02	0.376

Table 4.12
Domain predictions for discontinuities univariate FH model for the direct estimates of the discontinuities

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)	$\bar{\gamma}$
`offtot`	8.35	2.25	-33.39	26.63	69.35	0.012
`unsafe`	4.43	1.48	45.45	42.14	59.99	0.082
`nuisance`	0.32	0.06	-13.85	20.97	63.04	0.049
`satispol`	5.67	2.39	186.07	65.33	69.61	0.014
`propvict`	2.54	0.91	-55.59	45.73	65.84	0.032

Plots of discontinuities estimated with the GREG estimator, the univariate FH model for the alternative CVS, the bivariate FH model and the univariate FH model for the discontinuities are provided in Figures 4.6 through 4.10. The predictions for the domain discontinuities obtained with the three models are more stable compared to the GREG estimates. This is e.g., clearly illustrated with `unsafe` (Figure 4.7), where the GREG estimates for the discontinuity are sometimes positive and sometimes negative. The predictions for the domain discontinuities under the bivariate FH model and the univariate FH model for the discontinuities are consistently positive, which appears more plausible since it is unlikely that the domain discontinuities have opposite signs. The predictions for the domain discontinuities under the univariate FH model for the alternative CVS are closer to the GREG estimates and consequently less stable. A similar pattern can be observed for the other variables.

These plots illustrate that for `propvict`, `offtot` and `unsafe` the bivariate FH model results in a clear improvement of the predictions for the domain discontinuities compared to the univariate FH model for the alternative CVS. For `nuisance` the standard errors for the discontinuities increase with the bivariate FH model compared to the univariate FH model for the alternative CVS. The bivariate FH model for `satispol` cannot adequately model the observations under the alternative survey with the auxiliary information from the two registers (MBA and PRRO). In this case the domain predictions of `satispol` under the alternative approach display overshrinkage. The univariate FH model indeed selects an auxiliary variable from the regular survey, see Table 4.3, and clearly performs better.

It was anticipated that it would be difficult to produce reasonable predictions for the domain discontinuities with the univariate FH model for the direct estimates of the discontinuities since it is hard to imagine that the available auxiliary variables from registers like the MBA and PRRO contain good predictors for systematic differences in survey errors. Nevertheless, reasonable results are obtained with this more pragmatic approach. A possible interpretation is that the discontinuities are to some extent proportional to the values of the target variable and therefore show some systematic pattern that can be explained partially with the selected covariates. A point of concern are the very small shrinkage factors under this model, which might be an indication that the model gives too much weight to the synthetic estimator.

Figure 4.6 Discontinuities offtot based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

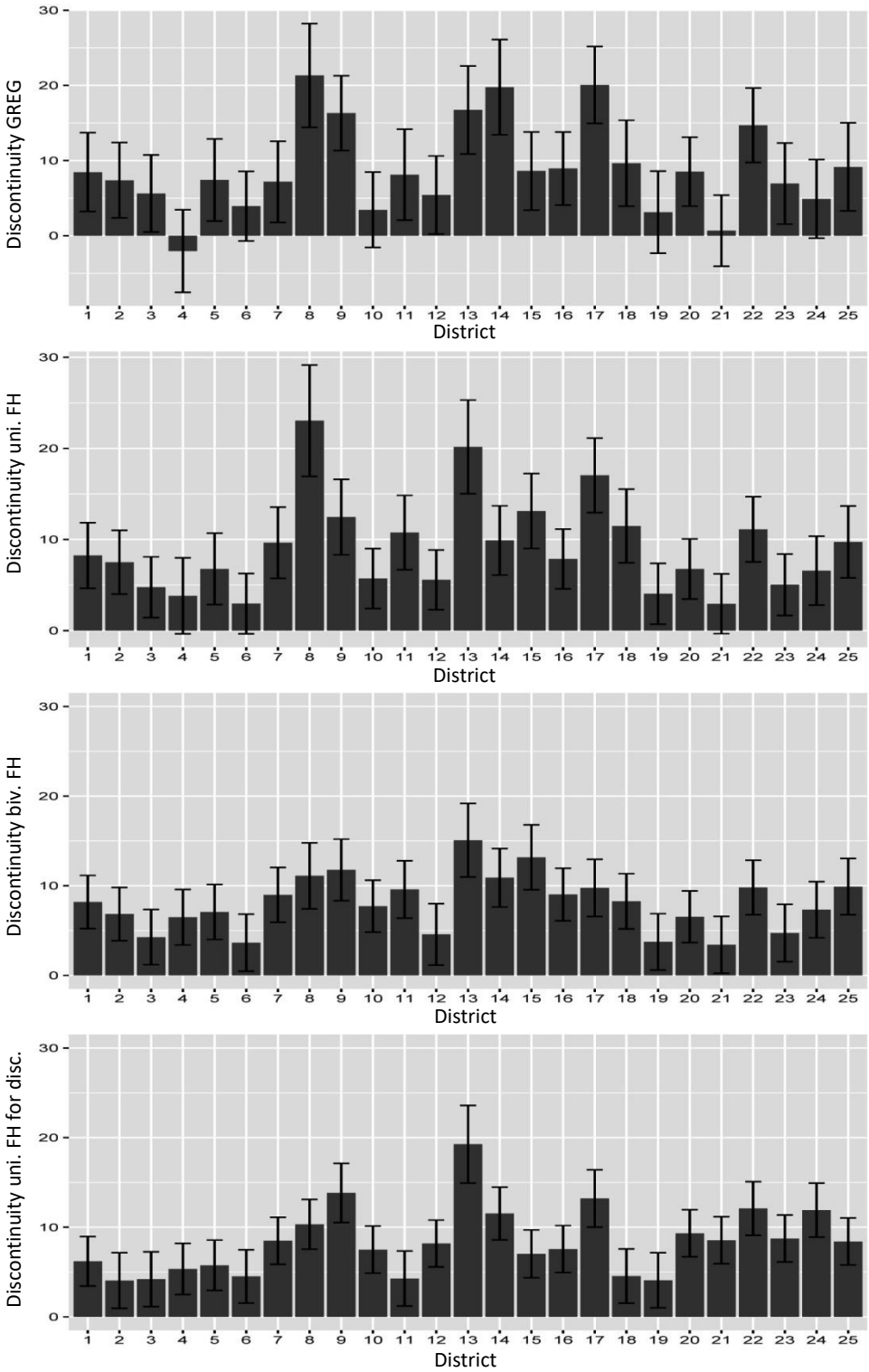


Figure 4.7 Discontinuities unsafe based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

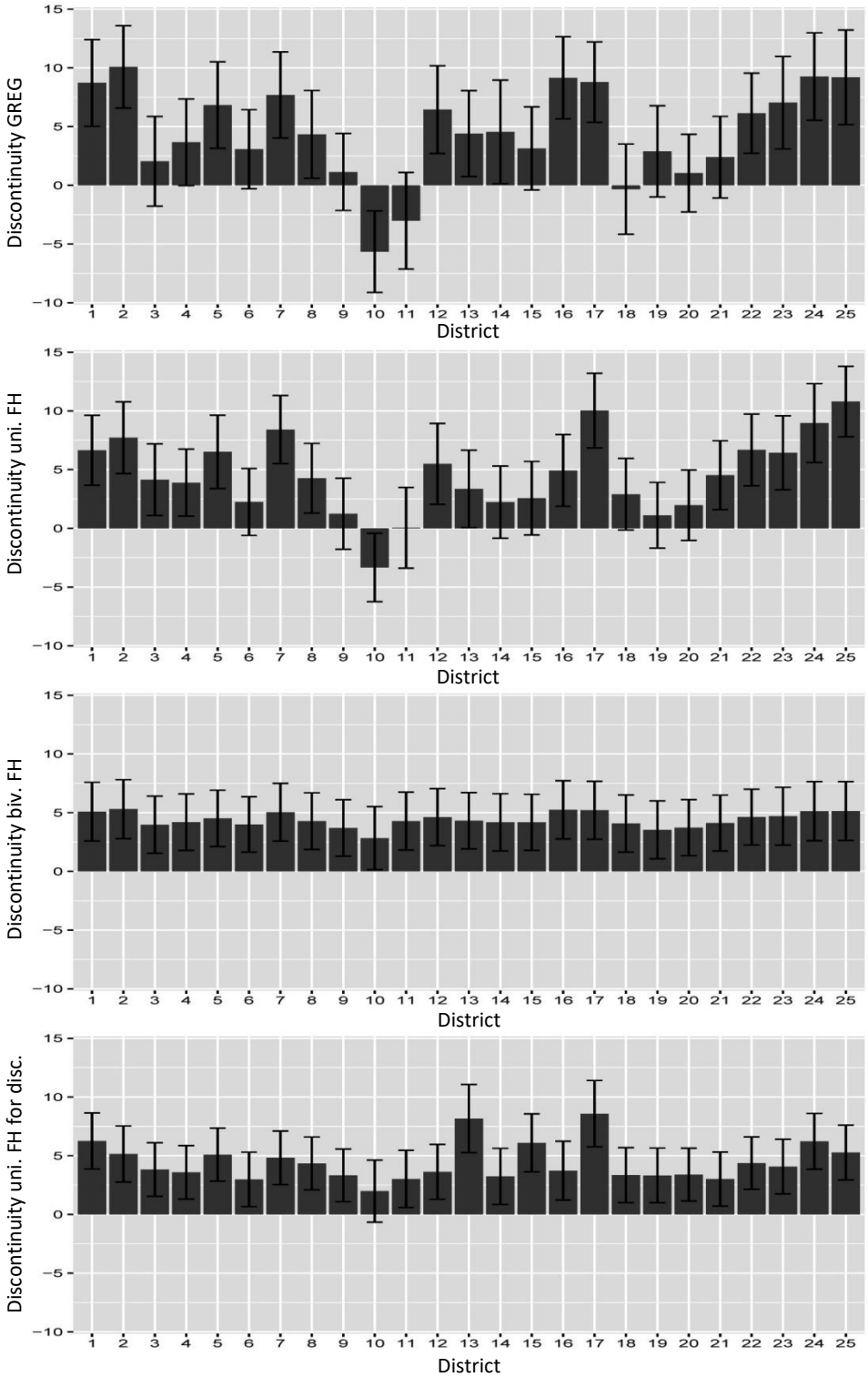


Figure 4.8 Discontinuities nuisance based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

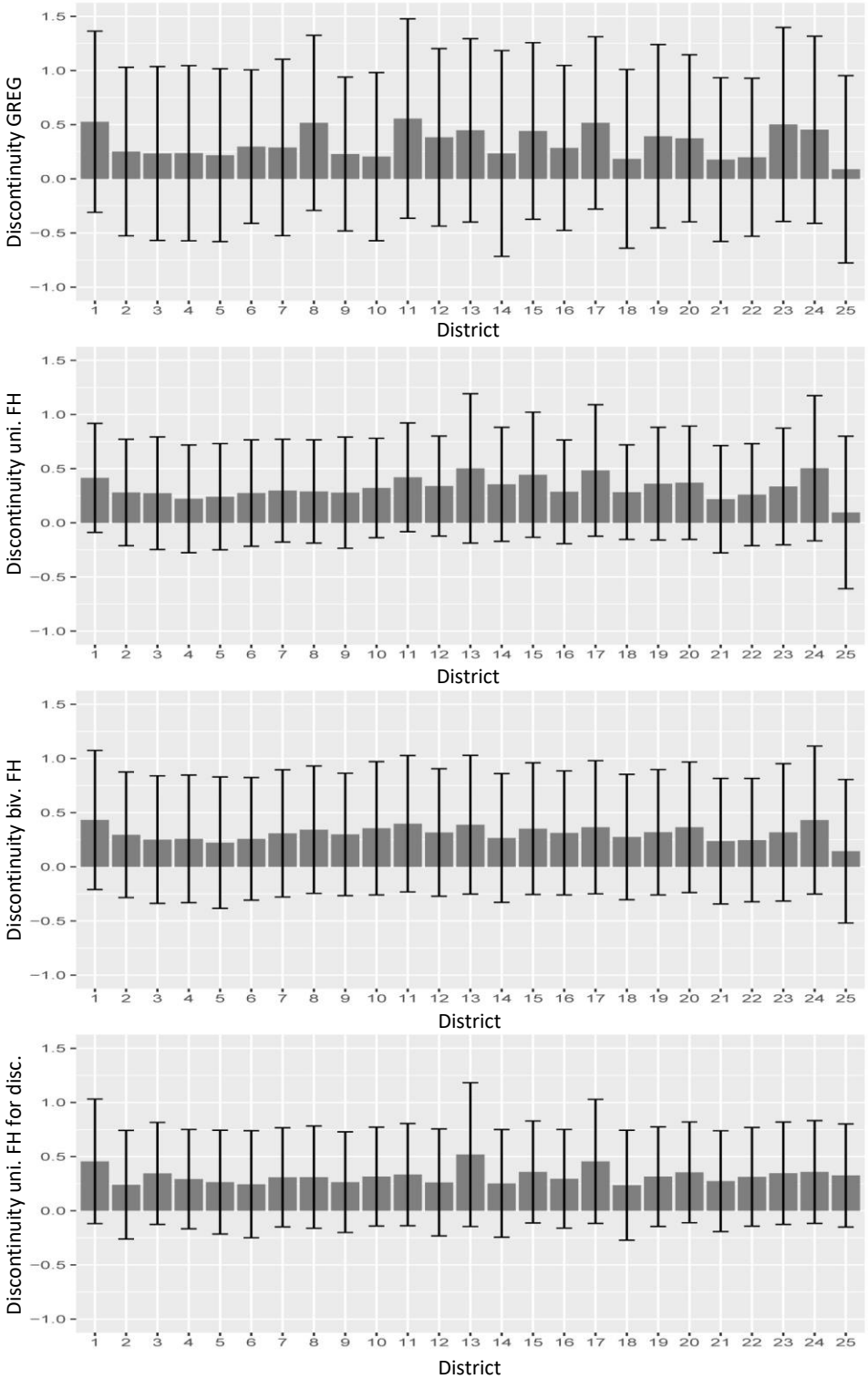


Figure 4.9 Discontinuities propvict based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

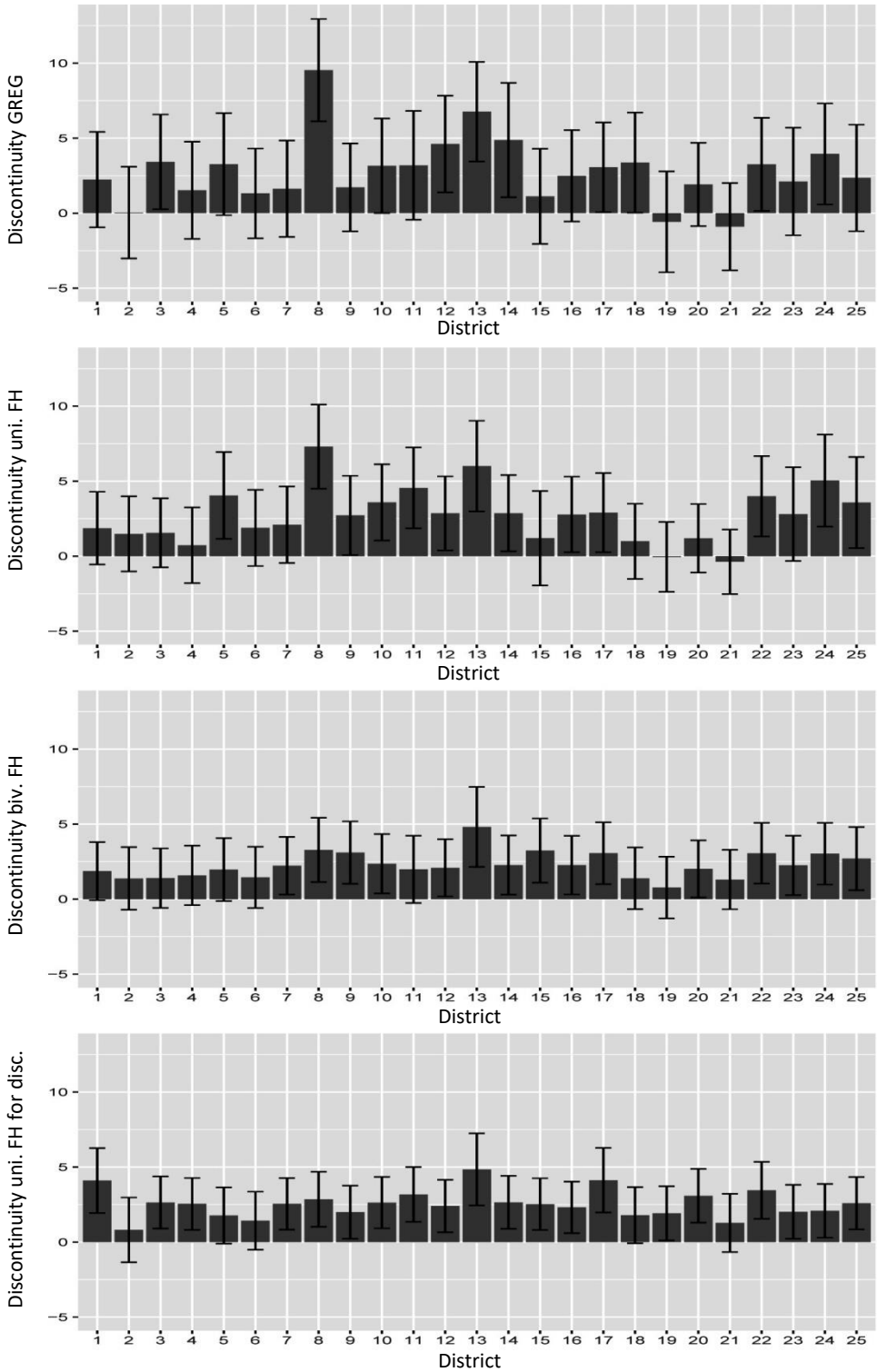
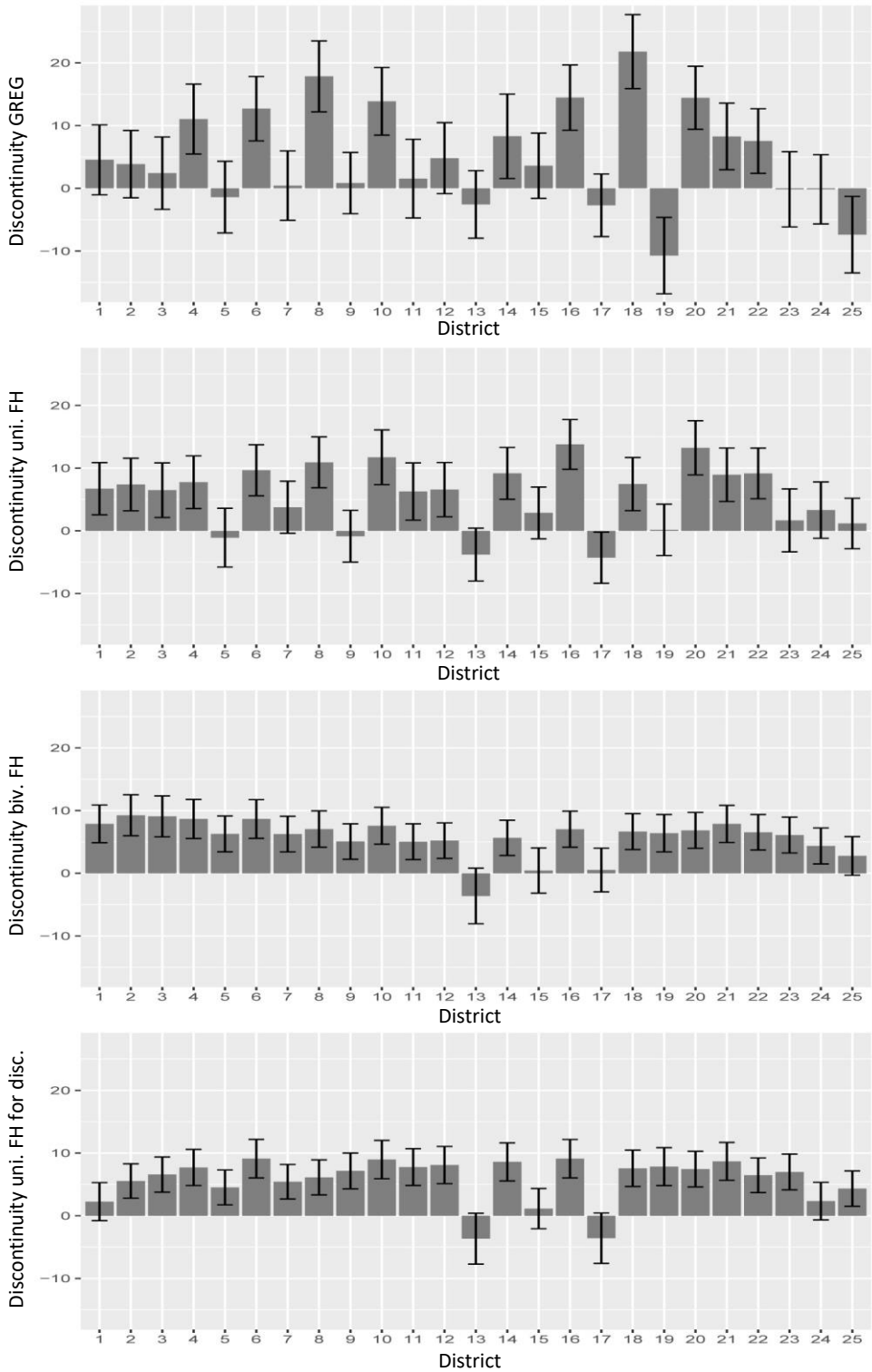


Figure 4.10 Discontinuities satispol based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.



5. Discussion

Survey process redesigns often result in discontinuities that disturb the comparability of the outcomes over time obtained with a repeated survey. To avoid confounding real period-to-period change with differences in measurement bias, it is important that such discontinuities are quantified during the implementation of a new survey process. A straightforward approach is to collect data under the old and new design in parallel to each other for some period of time. Available budgets for parallel data collection often do not meet the minimum required sample sizes that come from power calculations to detect minimum prespecified differences at certain significance and power levels. This might be sufficient for quantifying discontinuities at the national level but not at the domain level, even for the planned domains of the regular survey. To obtain more precise predictions for the domain discontinuities a small area estimation approaches based on hierarchical Bayesian Fay-Herriot (FH) models is proposed.

In an earlier paper (van den Brakel et al., 2016) a univariate FH model is proposed, where reliable direct domain estimates of the regular survey are considered as potential auxiliary variables in a step-forward model selection procedure to build adequate models for small domain prediction of the small sample assigned to the alternative survey. In this paper a bivariate FH model for the direct estimates obtained under both the regular and alternative survey is proposed as an alternative to obtain adequate predictions for domain discontinuities. In addition a univariate FH model applied to the direct estimates of the discontinuities is considered as a simple alternative. The methods are applied to a small scale parallel run conducted to quantify discontinuities in a survey process redesign of the Dutch Crime Victimization Survey (CVS).

Using direct estimates from the regular survey as auxiliary variables in models for small domains under the alternative approach results in a substantial improvement of precision, compared to univariate models that only use auxiliary variables from available registers. This can be expected since both surveys attempt to measure the same variables with a different survey approach. A drawback of the univariate approach is that the variance estimation procedure for the discontinuities is complex, since a non-negligible covariance between the direct estimates from the regular design and the model based predictions for the alternative design arises. The method is complex since a model-based MSE is combined with a design-based variance of a direct estimator. This might even result in negative variance estimates for the discontinuities. These complications are partially circumvented by developing a design-based estimator for the MSE of the small domain predictions and the covariance component (van den Brakel et al., 2016).

Under a bivariate FH model in a fully Bayesian framework negative variance estimates are avoided since the variances for disontinuities are derived from positive-definite covariance matrices of the bivariate model. The bivariate FH model improves the predictions for the domain discontinuities since the model improves the precision of the estimates of both the regular and alternative approach, and the strong positive correlation between the random domain effects further reduces the variance of the contrasts. For four out of five variables of the Dutch CVS the bivariate FH model indeed resulted in more precise

predictions for domain discontinuities compared to the univariate FH model. Another advantage of the bivariate model is that it improves the domain predictions of both the regular and alternative model while the univariate model assumes that the sample size of the regular survey is sufficiently large to make reliable precise direct domain estimates. The bivariate model is therefore also appropriate in parallel runs where e.g., the sample size of the regular survey is reduced in order to increase the sample size for the alternative survey. Finally the bivariate FH model avoids the complications to account for sampling error in the covariates, which is often required if the direct estimates of the regular survey are used as covariates in a univariate FH model.

For one variable (satisfaction with police performance) no adequate model could be constructed with the available auxiliary variables from the registers only. For this variable the multivariate model seems to result in overshrinkage of the predictions for the domain discontinuities. The results of the univariate model are clearly better in this case since the direct estimates from the regular survey are the only auxiliary variables that result in an adequate model for small domain predictions.

The univariate FH model for the direct estimates of the domain discontinuities turns out to be a reasonable alternative. It avoids the complications of the univariate FH model for the alternative CVS and the method is considerably simpler compared to the bivariate FH model. A point of concern are the extremely small shrinkage factors, which are an indication that the model puts too much weight on the synthetic part of the domain predictions. The bias of these domain predictions is indeed larger compared to that of the bivariate FH model.

A general problem in this application with the step-forward model selection procedure where covariates are included in the model as long as the WAIC value is reduced, is that this results in models with relatively large sets of covariates. With the limited number of domains in this application there is a real risk of overfitting the data. For some variables the covariates appear to be strong predictors for the domain variables, resulting in small random effects. Fitting a model without these covariates results in models with large random effects and strong positive correlations between the regular and alternative survey estimates. For other variables a model with a full covariance structure automatically results in parsimonious models for the fixed effect part, probably because the set of available covariates are less strong predictors for these target variables.

The aformentioned issue of selecting models with too many covariates is circumvented with an alternative step-forward selection approach. Since the WAIC values are estimated from the Gibbs sampler output, these values are observed with some degree of uncertainty. This is an argument not to include covariates if they only result in a small reduction of the WAIC. In an alternative step-forward selection approach, covariates are only selected if the decrease in the WAIC value exceeds the estimated standard error of the WAIC. With this approach parsimoneous models are selected since it avoids the selection of one or more covariates that only marginally improve the WAIC. For variables where initially large sets of covariates were selected, this approach results in a reasonable compromise between model fit and model complexity. As an alternative, models with equal regression coefficients can be considered. Such models

are, however, less appropriate for predicting domain discontinuities if the random effects are small. In such situations the dummy indicator is the only model component that discriminates between the regular and alternative approach. This results in synthetic predictions for domain discontinuities that are almost equal over the domains, and approximately equal to the direct estimator for the discontinuity at the national level. Depending on the type of changes in the survey process, it might be correct to assume that domain discontinuities are equal. In that case the best estimate is obtained with the direct estimator at the national level.

For a better understanding of the properties and behaviour of the three different models for estimating domain discontinuities, including the proposed model selection approach, a comprehensive simulation is required. This will provide a better understanding under what conditions, which of the three different modeling approaches are preferred. Such a study is left for future research.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily agree with the policy of Statistics Netherlands. We would like to thank the Associate Editor and the two referees for careful reading of our manuscript and providing useful comments. They proved to be very helpful to improve the quality of our paper.

Appendix

Table A.1
Overview auxiliary data

Variable	Description	Source
MBA_benefit	Percentage of social benefit claimants	MBA
MBA_immigr	Percentage of immigrants in population	MBA
MBA_immigrnw	Percentage of non-western immigrants in population	MBA
MBA_old	Percentage of elderly people (aged over 65)	MBA
MBA_benefit	Percentage of social benefit claimants	MBA
PR_assault	Peported physical assaults	PRRO
PR_propcrim	Property crimes	PRRO
PR_threat	Reported threats	PRRO
PR_weapon	Weapon offences	PRRO
PR_drugs	Illicit drug offences	PRRO
CVSR_nuisance	Perceived nuisance in the neighbourhood	regular survey
CVSR_victim	Percentage of people saying that they have been victim to a crime	regular survey
CVSR_funcpol	Opinion on functioning of the police on a 10-point scale	regular survey

References

Arora, V., and Lahiri, P. (1997). On the superiority of the bayesian method over the blup in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bell, W.R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*, 52nd Session.

Benavent, R., and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372-390.

Bollineni-Balabay, O., van den Brakel, J.A., Palm, F. and Boonstra, H.J. (2017). Multilevel hierarchical Bayesian vs. state space approach in time series small area estimation: The Dutch Travel Survey. *Journal of the Royal Statisitcal Society*, forthcoming.

Boonstra, H.-J. (2020). `mcmcsae`: *Markov Chain Monte Carlo Small Area Estimation*. R package version 0.5.0.

Boonstra, H.J., and van den Brakel, J.A. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 3, 395-425. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00005-eng.pdf.

Datta, G., Ghosh, M., Nangia, N. and Natarjan, K. (1996). Estimation of median income in four-peron families: A Bayesian approach. In *Bayesian analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner and J.M. Geweke), New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94(448), 1074-1082.

Estaban, M.D., Morales, D., Perez, A. and Santamaria, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840-2855.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366), 269-277.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

Gelman, A., Van Dyk, D.A., Huang, Z. and Boscardin, W.J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1), 95-122.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2004). *Bayesian Data Analysis*. New York: Chapman and Hall.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

Gonzales-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and Santamaria, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, 52, 5242-5252.

Gonzalez, M.E., and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Technical report, paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, 18-25 August, 1973.

Hawala, S., and Lahiri, P. (2018). Variance modelling for domains. *Statisitcs and Applications*, 16, 399-409.

Hirose, M., and Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *Annals of Statistics*, 46, 1721-1741.

Hodges, J.S., and Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2), 367-379.

Lahiri, P., and Pramanik, S. (2019). Evaluation of synthetic small-area estimators using design-based methods. *Austrian Journal of Statistics*, 48, 43-57.

Li, H., and Lahiri, P. (2010). Adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 57, 308-325.

Marker, D. (1995). *Small Area Estimation: A Bayesian Perspective*. Ph.D. thesis, University of Michigan.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484), 1405-1418.

Pfeffermann, D., and Ben-Hur, D. (2018). Estimation of randomisation mean square error in small area estimation. *International Statistical Review*, 87, 31-49.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Polson, N.G., and Scott, J.G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887-902.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. Unpublished research paper.

Rivest, L.-P., and Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology*, 26, 1, 67-78. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000001/article/5179-eng.pdf.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64(4), 583-639.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44(1), 150-167.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statisitcal Society*, 171, 581-613.

van den Brakel, J.A., and Boonstra, H.J. (2018). Hierarchical Bayesian bivariate Fay-Herriot model for estimating domain discontinuities. Discussion paper, Statistics Netherlands.

van den Brakel, J.A., and Krieg, S. (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statisitcal Society*, 179, 763-791.

van den Brakel, J.A., Buelens, B. and Boonstra, H.J. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. *Journal of the Royal Statisitcal Society*, 179, 229-250.

van den Brakel, J.A., Smith, P.A. and Compton, S. (2008). Quality procedures for survey transitions-experiments, time series and discontinuities. *Survey Research Methods*, 2, 123-141.

Vehtari, A., Gelman, A. and Gabry, J. (2015). `loo` : *Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models*. R package version 0.1.3.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistical Computation*, 27, 1413-1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

You, Y. (2008). Small area estimation using area level models with model checking and applications. *Proceedings of the Survey Methods Section of the Statistical Society of Canada*.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 2, 173-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000002/article/5537-eng.pdf.