STATISTICS IN TRANSITION new series and SURVEY METHODOLOGY
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 523–540

COVARIATE SELECTION FOR SMALL AREA ESTIMATION IN REPEATED SAMPLE SURVEYS

Jan A. van den Brakel[1], **Bart Buelens**[2]

ABSTRACT

If the implementation of small area estimation methods to multiple editions of a repeated sample survey is considered, then the question arises which covariates to use in the models. Applying standard model selection procedures independently to the different editions of the survey may identify different sets of covariates for each edition. If the small area predictions are sensitive to the different models, this is undesirable in official statistics since monitoring change over time of statistical quantities is of utmost importance. Therefore, potential confounding of true change and methodological alterations should be avoided. An approach to model selection is proposed resulting in a single set of covariates for multiple survey editions. This is achieved through conducting covariate selection simultaneously for all editions, minimizing the average of the edition-specific conditional Akaike Information Criteria. Consecutive editions of the Dutch crime victimization survey are used as a case study. Municipal estimates of three survey variables are obtained using area level models. The proposed averaging strategy is compared to the standard method of considering each edition separately, and to an elementary approach using covariates selected in the first edition. Resulting models, point estimates and MSE estimates are analyzed, indicating no substantial adverse effects of the conceptually attractive averaging strategy.

Key words: area level models, cAIC, Hierachical Bayesian predictors.

1. Introduction

At national statistical institutes, estimation procedures for surveys based on probability samples are traditionally based on design-based or model-assisted inference procedures. Well-known examples are the π-estimator (Narain, 1951; Horvitz and Thompson, 1952) and the generalized regression estimator (Särndal, Swensson and Wretman, 1992). These approaches are particularly appropriate in the case of large sample sizes. In the case of small sample sizes, however, design-based and model-assisted estimators have unacceptably large variances. This occurs when estimates

[1]Statistics Netherlands, Department of Statistical Methods and Maastricht University, Department of Quantitative Economics. E-mail: jbrl@cbs.nl
[2]Statistics Netherlands, Department of Statistical Methods. E-mail: bbus@cbs.nl

are required for detailed breakdowns of the population in subpopulations or domains according to various socio-demographic or geographic classification variables. In such cases, model-based estimation procedures are required to increase the effective sample size of the separate domains with sample information observed in other domains or preceding periods. This class of estimation procedures is known in the literature as small area estimation (SAE) (Rao, 2003; Pfeffermann, 2013) and offers promising opportunities for official statistics (Boonstra et al., 2008).

A common approach to introducing SAE in an existing survey is to apply SAE methods to historic editions of the survey, producing small area estimates for multiple past editions at the same time. This article focuses on the selection of covariates to be used in the SAE models in this setting. In the literature, model selection procedures mostly focus on the selection of optimal models for one particular survey data set (Claeskens and Hjort, 2008). If in each edition of a repeated survey a separate and different model is selected, the question arises to what extent the small area predictions are comparable over time. In official statistics potential confounding of estimates of change over time of some statistic with variations in the inference procedures must be avoided. This article contributes to the existing literature by addressing the question how to select a single optimal model for the production of SAE predictions for independent, repeated editions of a sample survey.

An approach is proposed in which the model selection criterion is averaged over all available editions, leading to a single set of covariates to be used in each edition. This novel approach is compared to the standard approach of selecting a set of covariates for each edition independently using four past editions of the Dutch crime victimization survey. In addition, a simple scenario is included whereby covariates are selected using only the first of a series of survey editions. In this paper models are considered that only use cross-sectional correlation. Alternative approaches that combine cross-sectional and temporal data are proposed by Rao and Yu (1994), Datta et al. (1999) and Pfeffermann and Tiller (2006). These approaches might also be considered to select one single optimal model for subsequent survey editions. These approaches are not considered for implementation in the Dutch crime victimization survey since they are considerably more complex and computationally intensive.

The article continues in Section 2 with a presentation of the SAE methods used and details covariate selection procedures. Section 3 introduces the crime victimization survey and potential covariates. Results are presented and discussed in Section 4. Conclusions are drawn in Section 5.

2. Methods

2.1. Small Area Estimation

In small area estimation multilevel models are used to improve the estimation of small domain parameters. These models use relevant auxiliary information as covariates. In this article the area level model is used (Fay and Herriot, 1979), where the input data for the model are the direct estimates for the domains. Approaches to covariate selection discussed below can be applied to unit level models (Battese, Harter and Fuller, 1988) as well. The area level model is considered, since it takes the complexity of the sample design into account as the dependent variables of the model are the design-based estimates derived from the probability sample and available auxiliary information used in the weighting model of the generalized regression (GREG) estimator. Let $\hat{\theta}_i$ denote the GREG estimates of the target variables θ_i for the domains $i = 1, \ldots, m$. In the area level model, the direct domain estimates are modeled with a measurement error model, i.e. $\hat{\theta}_i = \theta_i + e_i$, where e_i denotes the sampling error with design variance ψ_i. The unknown domain parameter is modeled with available covariates for the $i-$th domain, i.e. $\theta_i = z_i^{'}\beta + v_i$, with z_i a K-vector with the covariates $z_{i,k}$ for domain i, β the corresponding K-vector with fixed effects and v_i the random area effects with variance σ_v^2. For each variable a separate univariate model is assumed. Combining both components gives rise to the basic area level model, originally proposed by Fay and Herriot (1979):

$$\hat{\theta}_i = z_i^{'}\beta + v_i + e_i, \tag{1}$$

with model assumptions

$$v_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2) \text{ and } e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_i). \tag{2}$$

It is assumed that v_i and e_i are independent and that ψ_i is known.

Model(1) is a linear mixed model and estimation often proceeds using Empirical Best Linear Unbiased Prediction (EBLUP), where the between domain variance σ_v^2 is estimated with the Fay-Herriot moment estimator, maximum likelihood or restricted maximum likelihood, see Rao (2003), ch. 6 for details. A weakness of these methods is that in some situations the estimated model variance tends to zero, see e.g. Bell (1999) and Rao (2003). To avoid these problems, the Hierarchical Bayesian (HB) approach is followed in this article, Rao (2003), section 10.3. Therefore, the basic area level model is expressed as an HB model by (1) and (2) and a flat prior on β and σ_v^2. The HB estimates for θ_i and its MSE are obtained as the posterior mean and variance of θ_i. To account for the uncertainty in the between

domain variance, integration over the posterior density for σ_v^2 is conducted.

Estimates for the design variances ψ_i are available from the GREG estimator but are used as if the true design variances are known, which is a standard assumption in small area estimation. Therefore, it is important to provide reliable estimates for ψ_i. The stability of the estimates for ψ_i is improved using the following ANOVA-type pooled variance estimator

$$\begin{aligned} \psi_i &= \frac{1-f_i}{n_i} S_p^2, \\ S_p^2 &= \frac{1}{n-m} \sum_{i=1}^{m} (n_i - 1) S_{i;GREG}^2, \end{aligned}$$

with f_i the sample fraction in domain i, n_i the sample size in domain i, $n = \sum_{i=1}^{m} n_i$ and $S_{i;GREG}^2$ the estimated population variance of the GREG residuals.

2.2. Conditional AIC

The model selection procedures discussed here are optimization routines, minimizing the conditional Akaike Information Criterion (cAIC) proposed by Vaida and Blanchard (2005).

The cAIC is applicable to mixed models where the focus is on prediction at the level of clusters or areas (Vaida and Blanchard, 2005). It is defined as cAIC $= -2\mathscr{L} + 2p$, where $\mathscr{L}$ is the conditional log-likelihood and p a penalty based on a measure for the model complexity. In the case of a fixed effects model, p is the number of model parameters. The random part of a mixed model also contributes to the number of model degrees of freedom p with a value between 0 in the case of no domain effects (i.e. $\hat{\sigma}_v^2 = 0$) and the total number of domains m in the case of fixed domain effects (i.e. $\hat{\sigma}_v^2 \to \infty$). In the expression of the cAIC, p is the effective degree of freedom of the mixed model and is defined as the trace of the hat matrix H, which maps the observed data to the fitted values, i.e. $\hat{y} = Hy$, see Hodges and Sargent (2001).

When comparing models, the one with the lowest cAIC value is preferred.

2.3. Covariate selection procedures

Covariate selection procedures are aimed at establishing a set of covariates – in the present setting the fixed effects – to use in models specified by equation (1). This boils down to finding an optimal subset from a larger set of available candidates. All three methods detailed below proceed along the same lines: they follow a step-forward covariate selection strategy which starts from an intercept-only model

adding covariates one-by-one until there is no improvement in terms of the selection criterion. This may result in sub-optimal models as the procedure converges to a local minimum of the selection criterion but not necessarily to the global minimum (Claeskens and Hjort, 2008). The focus here, however, is on establishing a single set of covariates for use in repeated survey editions. Alternative search routines converging to the global minimum of the selection criterion can be applied analogously to the step-forward routine used here.

Some general notation is introduced. When C candidate covariates are available for inclusion as a fixed effect in a model specified by equation (1), the set of selected covariates is denoted by s and the set of remaining covariates by r. For ease of use the candidate covariates are assumed to be ordered in a fixed but arbitrary order, so that they can be referred to by their index. For example, a model containing the jth covariate – with $1 \leq j \leq C$ – as a fixed effect, can be identified by $s = \{j\}$. Consequently, in such case $r = \{i\}_{i \neq j}$. Evidently, the equality $s \cup r = \{1, \ldots, C\}$ always holds. Sets of selected and remaining covariates that are specific to a survey edition t are denoted by s_t and r_t respectively.

2.3.1 Selecting an optimal set for each edition separately

For a series of independent cross sectional surveys repeated at times $t = 1, \ldots, T$, a standard covariate selection routine consists in selecting covariates for each edition indepentently.

Covariate selection procedure *'stnd'*. Repeat for all editions $t \in \{1, \ldots, T\}$:

Initialization Set $r_t = \{1, \ldots, C\}$ and $s_t = \{\}$, obtain the corresponding cAIC value and call this $cAIC_0$. Set $i = 0$.

Repeat Attempt extending the model with one covariate:

a / Set $i = i + 1$.

b / Calculate cAIC for all models $s_t \cup \{j\}$, $\forall j \in r_t$, and call these $cAIC_j$.

c / If $min(cAIC_j) < cAIC_{i-1}$ then set $cAIC_i = min(cAIC_j)$, extend s_t to include the corresponding covariate j, remove that covariate from r_t.

Until The model is not extended or all candidate covariates are included in the model.

The result are sets s_t of selected covariates for each edition t. In general, s_t and $s_{t'}$ can be different for $t \neq t'$.

The generic model specification given by equation (1) is adapted to reflect the repeated nature of the survey.

$$\hat{\theta}_{i,t} = z_{i,t}^{[stnd]'} \beta_t + v_{i,t} + e_{i,t}, \tag{3}$$

for $i = 1,\ldots,m$ and $t = 1,\ldots,T$, with model assumptions

$$v_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_{v,t}^2) \text{ and } e_{i,t} \overset{\text{ind}}{\sim} \mathcal{N}(0,\psi_{i,t}). \tag{4}$$

The vectors $z_{i,t}^{[stnd]}$ consist of covariates contained in s_t at the level of the domains i, with s_t established through the *stnd* covariate selection procedure.

2.3.2 Selecting one optimal set for all editions simultaneously

Since the standard method may result in different sets of covariates for different survey editions, an alternative is proposed here, resulting in a single set of covariates for all editions. Formally, the following procedure enforces that $s_t = s_{t'}$ for all $t,t' \in \{1,\ldots,T\}$.

Covariate selection procedure *'avrg'*. Consider all survey editions $t = 1\ldots T$ simultaneously.

Initialization Let $r = \{1,\ldots,C\}$ and $s = \{\}$. Use r and s for all t, obtain the corresponding cAIC values, and call these $cAIC_{0,t}$. Define $cAIC_0 = \frac{1}{T}\sum_t cAIC_{0,t}$. Set $i = 0$.

Repeat Attempt extending the model with one covariate:

a / Set $i = i+1$.

b / For all editions $t \in \{1,\ldots,T\}$, calculate cAIC for all models $s \cup \{j\}$, $\forall j \in r$, and call these $cAIC_{j,t}$.

c / Define $cAIC_j = \frac{1}{T}\sum_t cAIC_{j,t}$.

d / If $min(cAIC_j) < cAIC_{i-1}$ then set $cAIC_i = min(cAIC_j)$, extend s to include the corresponding covariate j, remove that covariate from r.

Until The model is not extended or all candidate covariates are included in the model.

This strategy is based on averaging the model selection criterion cAIC and results in a single set s of covariates to be used in all editions t. The corresponding model specification, with a fixed set of covariates for repeated surveys, is written as (3)

and (4) where the vectors $z_{i,t}^{[stnd]}$ are replaced by vectors, say $z_{i,t}^{[avrg]}$, that consist of covariates contained in s at the level of the domains i at the time periods t, with s established through the *avrg* covariate selection procedure.

2.3.3 Selecting an optimal set based on the first edition only

An elementary approach also resulting in a single set of covariates is to use the first edition of a series of repeated surveys to establish the set of covariates and to use these in all subsequent editions.

Covariate selection procedure '*frst*'.

Apply procedure *stnd* for $t = 1$ to obtain s_1.

The set of covariates s_1 obtained based on the first edition is used at all times. The model takes the form of (3) and (4) where the vectors $z_{i,t}^{[stnd]}$ are replaced by vectors, say $z_{i,t}^{[frst]}$, that consist of covariates contained in s_1 at the level of the domains i at the time periods t. This strategy is included to assess and illustrate its performance. In other settings than the one discussed in the present article, statisticians may be in a situation where a survey is foreseen to be repeated in the future, but SAE estimates are required at the time of the first edition. The only option then is to use that edition for covariate selection.

3. Data

3.1. Crime victimization survey

The Dutch crime victimization survey underwent several redesigns in the past, including in 2008 and 2012. In the period from 2008 through 2011 the survey is known as the Integrated Safety Monitor (ISM). These four editions of the ISM are used as a case study in the present article. The purpose of the ISM is to publish information on crime victimization, public safety and satisfaction with police performance, among others. Each annual ISM sample is obtained independently through stratified simple random sampling of persons aged 15 years or older residing in the Netherlands. The population register serves as the sampling frame. The country is divided into 25 police districts, which are used as the stratification variable in the sample design. The yearly sample size of about 19,000 respondents is divided equally over the strata. In addition to this national sample, local authorities such as municipalities and police districts can draw supplementary samples in their own regions on a voluntary basis, with the purpose to obtain precise local estimates.

These supplementary samples are also based on stratified simple random sampling, but now with a more detailed geographical stratification variable, usually neighborhood. Table 1 gives an overview of the oversampling and the number of respondents for the years 2008 through 2011. Participation in the oversampling scheme by local authorities was encouraged in the years 2009 and 2011 resulting in much larger samples in these editions.

Table 1: Overview of response and oversampling in ISM surveys 2008 - 2011.

	2008	2009	2010	2011
Number of oversampled municipalities	77	239	21	225
Size response national sample	16,964	19,202	19,238	20,325
Size response supplemental sample	45,839	182,012	19,982	203,621
Percentage of population in oversampled areas	29%	65%	16%	66%

Data collection is based on a sequential mixed mode design using internet (WI), paper (PAPI), telephone interviewing (CATI) or face-to-face interviewing (CAPI). For the data collection of the additional regional samples the WI, PAPI and CATI modes are mandatory. The use of the CAPI mode is recommended but not mandatory since this mode is very costly. Statistical inference for official publication purposes is based on the GREG estimator. The inclusion probabilities in the ISM are determined by the sampling design, accounting for stratification and oversampling at regional levels. The GREG estimator uses a complex weighting scheme that is based on the auxiliary variables age, gender, ethnicity, urbanization, household size, police district, and the strata used in the regional oversampling scheme. In addition, the weighting scheme contains a component that calibrates the response to a fixed distribution over the data collection modes with the purpose to stabilize the measurement error between the subsequent editions of the ISM, (Buelens and Van den Brakel, 2015). Variance estimates are obtained with the standard Taylor series approximation of the GREG estimator, see Särndal, Swensson and Wretman (1992), ch. 6.

The GREG estimator can be used to produce reliable official statistics for regions with relatively large sample sizes. With the aforementioned sample design this implies that the GREG estimator can be used to produce official statistics at the level of police districts and in the regions where additional samples are drawn also at the level of municipalities. For regions where no additional samples are drawn, sample sizes are too small to produce reliable estimates at the level of municipalities with the GREG estimator. Since there is a growing demand for such figures, SAE procedures are developed to produce reliable official statistics on crime victimization at the municipal level. Three important ISM variables under study are listed in

Table 2.

Table 2: Overview of key ISM variables and their associated statistics.

Variable	Description of statistic
victim	Percentage of people who indicated that they were a victim of crime in the last 12 months
degen	Degeneration of the neighborhood (on a scale 1-5)
contpol	Percentage of people who had contact with the Police in the last 12 months

3.2. Candidate covariates

The success of increasing the precision of the domain estimates using SAE methods critically depends on the availability of correlated auxiliary information. An overview of 21 potential covariates used for model building is given in Appendix A. These are obtained from the Police Register of Reported Offences (PRRO) and from the population register.

The auxiliary variables `mode2` and `oversampled` require some explanation. In areas where local authorities draw supplemental samples, the fraction of responses obtained through non-interviewer administered modes is larger compared to areas without such oversampling. This is caused by the fact that CAPI is not conducted for the supplemental samples. There are clear indications that there are systematic differences in measurement error between responses obtained through interviewer and non-interviewer administered modes (Buelens and Van den Brakel, 2015; Schouten et al., 2013). As mentioned in section 3.1, the GREG estimator calibrates the response to fixed mode distributions to level out large fluctuations in measurement error due to large fluctuations in the distribution of the response over the different modes (Buelens and Van den Brakel, 2015). Since the calibration occurs at the police district level and not at the municipal level, it can be expected that the fraction of non-interviewer administered modes or a dummy indicator to differentiate between municipalities where oversampling took place or not, has predictive power for at least some of the target variables, due to potential correlation between these covariates and mode-dependent measurement error present within the municipal estimates.

4. Results

The different covariate selection strategies are applied to the four ISM editions for selecting covariates for SAE models for the three study variables. The sections be-

low discuss the sets of selected covariates and compare performance of the resulting models. The HB estimates are computed using the statistical software environment R (R Development Core Team, 2009) and package `hbsae` (Boonstra, 2012).

4.1. Covariate selection results

The covariate selection results are given in Tables 3 and 4. When using the *stnd* approach, different sets of covariates are selected in different survey editions. Not only do the covariates differ, also their number can vary between years. The variables selected through the *avrg* strategy often appear in at least one of the *stnd* models. The *frst* approach is not listed in Table 3 as it uses the set of covariates selected through the *stnd* approach in 2008.

Naturally, the *stnd* models result in lower cAIC values than the other strategies, see Table 4. By definition, the *avrg* and *frst* procedures result in the same sets of covariates to be used for all editions. For 2008, the *frst* and *stnd* approaches are identical and can therefore be expected to perform better than the *avrg* approach in that edition. For the subsequent years, 2009-2011, the cAIC values associated with the *avrg* approach are mostly smaller than or equal to the cAIC values obtained with the *frst* approach. In some cases the covariates selected for 2008 perform well in other years too, this is the case for example with `victim` in 2010.

4.2. Small area estimates

The purpose of applying SAE techniques in official statistics is to increase precision of area estimates. When considering the use of the *avrg* or *frst* approaches it is of interest to compare the reductions in variance achieved with these strategies compared to the *stnd* approach. An appropriate quantity to study in this context is the mean reduction in the coefficient of variation (MRCV),

$$MRCV = \frac{1}{m}\sum_{i=1}^{m}\frac{CV(\hat{\theta}_{i,t}) - CV(\tilde{\theta}_{i,t})}{CV(\tilde{\theta}_{i,t})}, \tag{5}$$

with $\hat{\theta}_{i,t}$ the GREG estimator and $\tilde{\theta}_{i,t}$ the HB prediction for domain i, and $CV(x)$ the coefficient of variation of estimator x (the estimated standard error divided by the point estimate). Note that MRCV would not be a suitable model selection criterion as it is susceptible to over fitting.

Table 3: Covariates selected by the different methods. Strategy *frst* uses the covariates selected for *stnd* in 2008.

Variable	Method	Covariates (listed in order in which they were selected)
victim	stnd 2008	logdens, propcrimedef2, oversampled, nonwestimmi, old, westimmi, highincome, lowincome, density, carsphh
	stnd 2009	sqrtdens, propcrimedef2, carsphh, mode2, old, westimmi
	stnd 2010	sqrtdens, propcrimedef1, young
	stnd 2011	propcrimedef1, sqrtdens, old, totcrime, mode2, rent, nonwestimmi, meanvalue, lowincome, oversampled
	avrg	sqrtdens, propcrimedef1, young, oversampled, totcrime, westimmi
degen	stnd 2008	rent, totcrime, prov, old, meanvalue, unemployed
	stnd 2009	rent, prov, totcrime, meanvalue, mode2, old, young
	stnd 2010	rent, prov, meanvalue, violcrime, oversampled, mode2, totcrime, biketheft, old, density, logdensity, lowincome
	stnd 2011	rent, totcrime, biketheft, mode2, old, meanvalue, logdens, violcrime, lowincome, carsphh
	avrg	rent, prov, violcrime, meanvalue, totcrime, mode2, biketheft, oversampled, old, propcrimedef2
contpol	stnd 2008	logdens
	stnd 2009	sqrtdens, violcrime, young, mode2
	stnd 2010	logdens, westimmi
	stnd 2011	logdens, violcrime, biketheft, westimmi, prov, highincome
	avrg	logdens, violcrime, westimmi

The MRCV values obtained in this study are listed in Table 5. While the largest reductions are naturally achieved with the *stnd* models, the suboptimality of the *avrg* and *frst* models is mild. Overall, the reductions achieved with the latter methods are only a few percentage points smaller than those achieved with the optimal models. Comparing the *avrg* and *frst* approaches, the former mostly result in greater reductions, although not always.

Table 4: Covariate selection results.

		number of covariates				cAIC			
Variable	Method	2008	2009	2010	2011	2008	2009	2010	2011
victim	stnd	10	6	3	10	-949	-1308	-914	-1395
	avrg	6	6	6	6	-934	-1308	-902	-1381
	frst	10	10	10	10	-949	-1304	-905	-1393
degen	stnd	6	7	12	10	724	353	761	273
	avrg	10	10	10	10	722	357	766	276
	frst	6	6	6	6	724	361	784	294
contpol	stnd	1	4	2	6	-161	-661	-698	-811
	avrg	3	3	3	3	-157	-657	-696	-805
	frst	1	1	1	1	-161	-657	-693	-798

Table 5: Mean reduction in coefficient of variation (in %).

Variable	Method	2008	2009	2010	2011
victim	stnd	-76	-57	-88	-63
	avrg	-72	-56	-86	-60
	frst	-76	-55	-83	-60
degen	stnd	-47	-31	-55	-32
	avrg	-46	-31	-53	-32
	frst	-47	-31	-50	-30
contpol	stnd	-92	-86	-83	-82
	avrg	-90	-86	-83	-87
	frst	-92	-87	-82	-88

Comparing Tables 4 and 5, it is observed that cAIC and MRCV values do not always exhibit the same pattern. For example the cAIC values for the variable `contpol` in 2011 indicate that the *stnd* approach is best, followed by the *avrg* and *frst* approaches. The corresponding MRCV values on the other hand reverse this pattern, with the *frst* approach resulting in greatest reduction and the *stnd* in lowest.

The values in Table 5 indicate that the SAE method in this case is most beneficial for the variable `contpol` with reductions in the coefficient of variation of up to almost 90%. The gains in precision for `victim` are smaller and for `degen` the smallest at around 30% in 2009 and 2011.

In line with the observation in section 3.1 that the oversampling in the ISM was much more intense in 2009 and 2011, it is seen in Table 4 that the cAIC values for these years are smaller than for 2008 and 2010 for each variable and method,

indicating better model fits in editions with larger samples. The gains to be had from SAE, however, are larger in the editions with smaller sample sizes, see Table 5.

Of practical relevance is the effect of the covariate selection strategy on the HB point estimates. SAE estimates obtained through the *stnd*, *avrg* and *frst* approaches are compared to GREG estimates in Figure 1. Four municipalities with varying sample sizes are chosen as an example. The number at the top of each panel in the plot refers to the rank of the municipality when ordered according to sample size (0001 being the smallest, and 0418 the largest). The four types of estimates are compared. The differences between the three types of SAE estimates are much smaller than the difference between the SAE and GREG estimates, apart from the larger municipalities where all estimates almost coincide, such as in Amsterdam. In the smaller municipalities, where the sample sizes are generally smaller, the differences are larger and the advantage of using SAE methods becomes apparent. While the *avrg* and *frst* approaches lead to point estimates close to those obtained through the *stnd* approach, sometimes there are differences, in particular in the smaller municipalities. An example is the variable `degen` in municipality '0008' (top left in middle panel of Fig. 1). There, the *avrg* estimates are closer to the *stnd* estimates than the *frst* estimates for the years 2009-2011. This is an indication that situations can arise where the covariates selected in 2008 are suboptimal in later editions, while those selected through the averaging strategy perform better overall.

More detailed results are available online in a Statistics Netherlands research report (Buelens and Van den Brakel, 2014). In this document the point estimates and variance estimates under the different model selection procedures are compared. Additional information on model evaluation is also included in this paper.

5. Conclusion

The issue considered in this article is the choice of model covariates when applying small area estimation repeatedly in consecutive, independent editions of a survey. The model under consideration is the area level model known as the Fay-Herriot model in combination with an Hierarchical Bayesian prediction approach. Model selection in this setting boils down to selecting an optimal set of covariates from a set of possible candidates.

While selecting an optimal set of covariates for each edition separately may be preferable from a modeling perspective, in official statistics it is important to avoid all potentially confounding elements in estimation of temporal change of published statistical results. Using the same set of covariates in SAE models every year is deemed essential. A strategy is proposed in which all editions of a survey are considered simultaneously, and a single set of covariates is selected. This approach uses

the cAIC criterion and operates by minimizing the cAIC averaged over all survey editions. A simple additional approach is included in the analyses, consisting of selecting covariates based on the first edition of a survey and using this set in all subsequent editions.

In the four editions of the crime victimization survey, it is shown that the models obtained through the averaging approach are only mildly suboptimal. The resulting coefficients of variation are marginally larger than those obtained for estimates based on specific optimal models for each edition. Models based on the first edition only are somewhat worse than the models obtained through averaging, but not substantially. In this application, point estimates are found to be very similar under all three SAE approaches, with the estimates obtained through the averaging models closer to the optimal models than the estimates obtained by using only the first edition. The models obtained through the averaging approach are used to produce official statistics about crime victimization and public safety at the municipal level, for twelve ISM survey variables in addition to the three discussed in this article.

The fact that using the first edition of a repeated survey to establish models once and that using them unaltered thereafter provides reasonable results not dramatically different from using optimal models in each edition, is an empirical finding for this application. This is the approach that would ordinarily be taken when a new survey is introduced with the plan to repeat it at future points in time. When SAE statistics are required in the first edition there is no other option than to base model selection on that edition alone. When multiple editions are available, however, it is recommendable to conduct model selection on these editions simultaneously using the proposed averaging strategy. Even if a number of past editions are available, it remains necessary to evaluate the selected models if the data under new editions of the survey become available. Changing the model might require a revision strategy for figures already published in the past.

While the averaging method is developed for area level models in the present study, it is in principle applicable to situations with other models as well including the unit level model (Battese, Harter and Fuller, 1988) and models with spatial effects (You and Zhou, 2011). Similarly, other model selection criteria than cAIC could be used if desired. Buelens and Van den Brakel (2014) considered leave-one-out cross validation and found it to result in less parsimonious models than cAIC. Recently, Lahiri and Suntornchost (2015) proposed a new variable selection criterion specifically for Fay-Herriot models. Each of these alternatives can immediately be plugged into the selection strategies presented in this article.

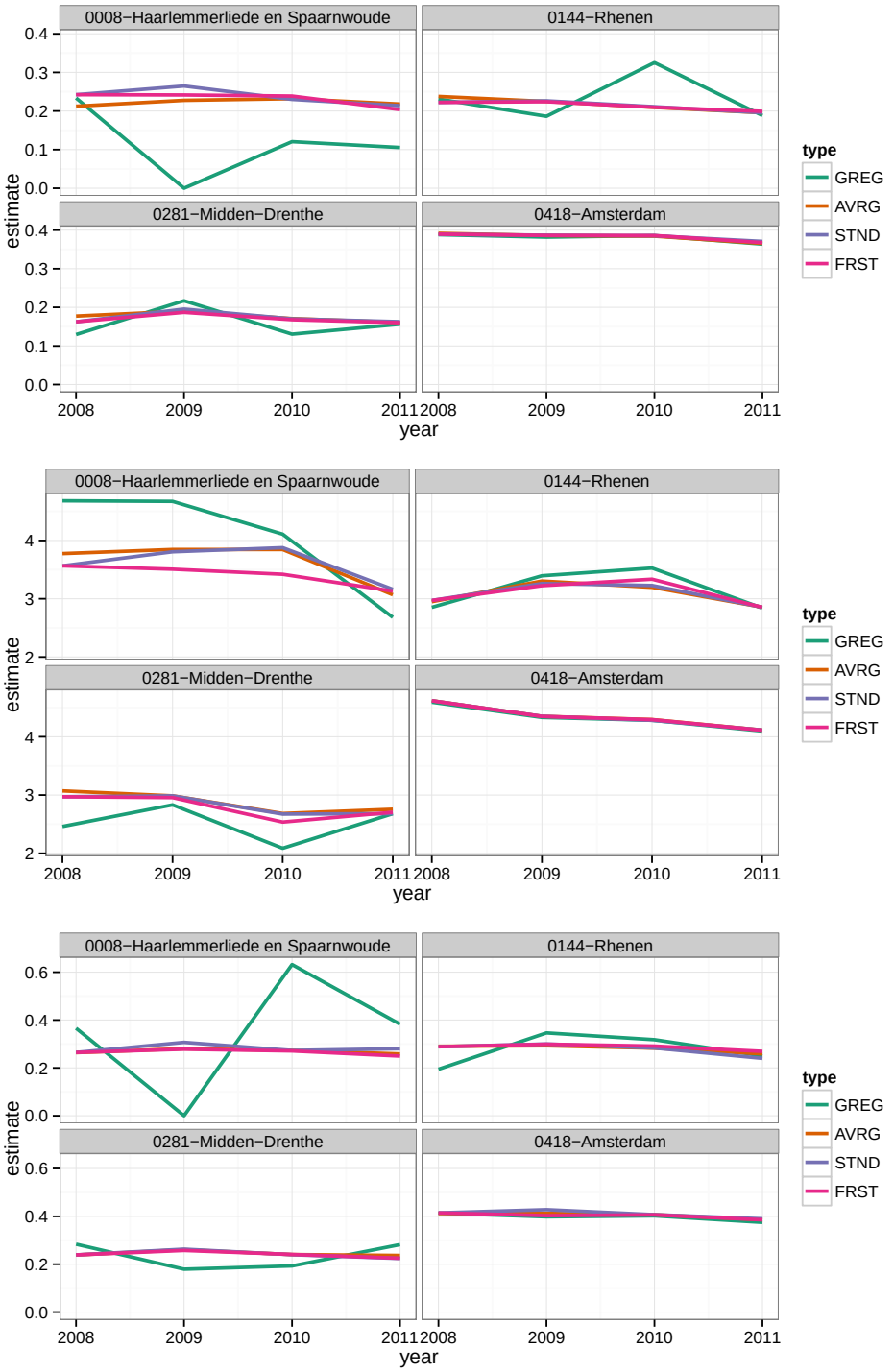

Figure 1: Time series of GREG and SAE estimates obtained through the *stnd*, *avrg* and *frst* approaches for four municipalities for `victim` (top), `degen` (middle) and `contpol` (bottom).

Appendix A: Auxiliary variables defined for municipalities

westimmi:	share of western immigrants in the population
nonwestimmi:	share of non-western immigrants in the population
prov:	province
density:	housing density (number of dwellings per square kilometer)
logdens:	natural logarithm of density
sqrtdens:	square root of density
meanvalue:	mean house value (available from housing register)
carsphh:	average number of cars owned by households
young:	share of population aged 15-30
old:	share of population aged 65+
rent:	share of houses that are rented (as opposed to owned)
lowincome:	share of households with a low income (nationwide in lowest quintile)
highincome:	share of households with a high income (nationwide in highest quintile)
unemployed:	share of population registered at the employment agency as looking for work
totcrime:	number of crimes registered by the Police per 1.000 inhabitants
propcrimedef1:	number of property crimes registered by the Police per 1.000 inhabitants (definition CBS)
propcrimedef2:	number of property crimes registered by the Police per 1.000 inhabitants (definition Bureau Veiligheid)
biketheft:	number of bicycle thefts registered by the Police per 1.000 inhabitants
violcrime:	number of violent crimes registered by the Police per 1.000 inhabitants
mode2:	share of non-interviewer administered modes (paper and web) in the ISM survey
oversampled:	binary variable indicating whether the municipality took part in the ISM oversampling scheme

REFERENCES

BATTESE, G.E., HARTER, R.M., FULLER, W.A., (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.

BELL, W.R., (1999). Accounting for uncertainty about variances in small area estimation. *Technical report*. Bulletin of the International Statistical Institute.

BOONSTRA, H.J., (2012). hbsae: Hierarchical Bayesian Small Area Estimation, *Manual R package version 1.0.*. Statistics Netherlands, Heerlen.

BOONSTRA, H.J., VAN DEN BRAKEL, J.A., BUELENS, B., KRIEG, S., SMEETS, M., (2008). Towards small area estimation at Statistics Netherlands. *METRON International Journal of Statistics*, LXVI, 21–49.

BUELENS, B., VAN DEN BRAKEL, J.A., (2014). Model selection for small area estimation in repeated surveys. *Discussion paper 201423*, Statistics Netherlands, Heerlen. http://www.cbs.nl/NR/rdonlyres/308ED398-714A-41A4-A57 C-9DCCC3F30D35/0/201423x10pub.pdf

BUELENS, B., VAN DEN BRAKEL, J.A., (2015). Measurement error calibration in mixed mode surveys. *Sociological Methods and Research*, 44, 391–426.

CLAESKENS, G., HJORT, N.L., (2008). *Model selection and model averaging*, Cambridge series on statistical and probabilistic mathematics, Cambridge University Press.

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

FAY, R.E., HERRIOT, R.A., (1979). Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268–277.

HODGES, J.S., SARGENT, D.J., (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, 88, 367–379.

HORVITZ, D.G., THOMPSON, D.J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, 1–9, doi = 10.1007/s13571-015-0096-0.

NARAIN, R., (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 581–613.

PFEFFERMANN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28, 40–68.

PFEFFERMANN, D., TILLER, R., (2006). Small Area Estimation with State-Space Models subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, 1387–1397.

R DEVELOPMENT CORE TEAM, (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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

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

SÄRNDAL, C-E., SWENSSON, B., WRETMAN, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.

SCHOUTEN, B., VAN DEN BRAKEL, J.A., BUELENS, B., VAN DER LAAN, J., KLAUSCH, T., (2013). Disentangling mode-specific selection bias and measurement bias in social surveys. *Social Science Research*, 42, 1555-1570.

VAIDA, F., BLANCHARD, S., (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92, 351–370.

YOU, Y., ZHOU, Q., (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37, 25–36.