Statistics Netherlands

Discussion Paper

Model selection for small area estimation in repeated surveys

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

2014 | 23

Bart Buelens
Jan van den Brakel
16 October 2014

Many surveys are repeated at regular intervals to monitor temporal change in quantities of social or economic importance. When small area estimation methods are applied in such settings, the question arises of which models to use. Model based small area estimation is typically applied when the sample size is too small to deliver sufficiently precise estimates for sub populations. Known population characteristics play the role of covariates in models in which the dependent variables are the survey variables. Model selection techniques are used to identify an optimal set of covariates from a larger set of candidates. In the context of repeated surveys the question arises how to select a model for small area predictions that are comparable over time. To this end, this article presents and compares four approaches to model selection for repeated surveys: establishing a model once and applying it unaltered thereafter, selecting an optimal model for each survey edition independently, selecting a single model for survey data pooled over time, or conducting a model selection procedure averaging the optimization criterion over multiple survey editions. Consecutive editions of the Dutch crime victimization survey known as the Integrated Safety Monitor are used as a case study. Municipal small area estimates of three important survey variables are obtained using area level models with a set of covariates selected from a list of 21 candidates. Advantages and disadvantages of the four approaches are considered, leading to the conclusion that the fourth approach is preferred from an official statistics perspective. It results in a robust model for all editions, producing estimates with a small bias compared to optimal estimates for each edition, and only a slightly higher variance. The described methodology is used to produce official statistics about crime victimization and safety feelings at the municipal level.

Keywords: area level models, cAIC, cross validation, Hierachical Bayesian predictors

Contents

1 Introduction

Estimation procedures for surveys based on probability samples are at Statistics Netherlands, like many other national statistical institutes, traditionally based on design-based or model-assisted inference procedures. This refers to a class of estimators which are predominantly based on the probability structure of the sample design that is used to select a sample from a finite population while statistical models only play a minor role. Well known examples are the π-estimator (Narain, 1951; Horvitz and Thompson, 1952) and the general regression estimator (Särndal et al., 1992). In the case of large sample sizes, these estimators have nice statistical properties that make them very attractive to produce timely official statistics. In the case of small sample sizes, however, design-based and model-assisted estimators have unacceptably large variances. These problems occur if estimates are required for very 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 (Rao, 2003; Pfeffermann, 2013).

Since the relevance of official statistics increases with the level of detail of the publication domains, a small area estimation program was set up at Statistics Netherlands (Boonstra et al., 2008). Since June 2010, Statistics Netherlands uses a multivariate structural time series model to produce official monthly statistics about the labor force (Van den Brakel and Krieg, 2009, 2014). At the same time research was initiated into multilevel modeling to produce annual municipal statistics about the labor force, crime victimization and safety feelings and health, using respectively the Labor Force Survey, the Integrated Safety Monitor (ISM) and the Health Survey. It is foreseen that Statistics Netherlands will publish official statistics on these three themes at municipal level in the near future.

This paper focuses on the implementation of small area estimation procedures in repeated cross-sectional surveys using the ISM as an example, and in particular on the selection of robust models. In the literature, model selection procedures focus on the selection of optimal models for one particular data set. If in each edition of a repeated survey a separate and different model is selected, the question arises to which extent the small area predictions are comparable over time. This paper contributes to the existing literature by addressing the question how to select optimal models for the production of small area predictions that are comparable over time. Four approaches are proposed and compared: establishing a model once and applying it unaltered thereafter, selecting an optimal model for each survey edition independently, selecting a single model for pooled survey data, or conducting a model selection procedure averaging the optimization criterion over multiple survey editions. These approaches are applied to the ISM editions of 2008 through 2011, leading to conclusions that are useful for repeated cross-sectional surveys in general.

The paper starts with a brief description of the Integrated Safety Monitor. The small area estimation approach followed in this application is described in Section 3. Section

4 describes the four alternative model selection procedures followed to select robust models for this survey. Results are presented in Section 5. The paper concludes with a discussion in Section 6.

2 The Integrated Safety Monitor

The purpose of the ISM is to publish information on crime victimization, public safety and satisfaction with police performance, among others. In its current form, the ISM has been conducted since 2008. Preceding versions of the ISM are the National Safety Monitor conducted from 2005 until 2008 and the Justice and Security Module of the Permanent Survey on Living Conditions conducted from 1997 until 2004, (Van den Brakel et al., 2008). The ISM is based on a national sample conducted by Statistics Netherlands. In addition local authorities can draw additional samples in their own region. The national sample is based on stratified simple random sampling of persons aged 15 years or older residing in the Netherlands. 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 equally divided over the strata. As already mentioned, municipalities and police districts can draw additional samples in their own region on a voluntary basis with the purpose to provide precise local estimates. These 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, 2009, 2010 and 2011.

Data collection is based on a sequential mixed mode design. All persons included in the sample receive an advance letter where they are asked to complete a questionnaire via internet (WI). Persons can receive a paper version of the questionnaire on their request (PAPI). After two reminders, nonrespondents are contacted by telephone if a telephone number is available to complete the questionnaire (CATI). The remaining persons are visited at home by an interviewer to complete the questionnaire face to face (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 generalized regression (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, 2014). Variance estimates are obtained with the standard Taylor series approximation of the GREG estimator, see Särndal et al. (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, a small area estimation procedure is developed to produce sufficiently reliable official statistics on crime victimization at the municipal level. See Table 2 for an overview of the target variables for which small area estimates are developed.

In 2012 the survey design of the ISM was changed, resulting in discontinuities in the outcomes of the key parameters of this survey. This redesign results in discontinuities in the target variables, making small area predictions under the old and new design incomparable. Therefore this paper discusses the small area estimation procedures developed for the period from 2008 until 2011.

3 Small Area Estimation

In small area estimation, two types of models are commonly used. The first is the basic area level model, also known as the Fay-Herriot model (Fay and Herriot, 1979), where the input data for the model are the direct estimates for the domains. The second is the nested error regression model of Battese et al. (1988), which is often referred to as the basic unit level model. In this model the input data are the observations obtained from the sampling units. Through these models, other relevant information can be used to improve the estimation of small domain parameters. An important source of auxiliary data in this study is the Police Register of Reported Offences (PRRO). This information is available at an aggregated level per municipality. In addition, demographic information available or derived from the municipal administrations is used. Therefore the basic area level model (Fay and Herriot, 1979) is considered in this application. Another advantage of the area level model is that the complexity of the sample design is taken into account, since 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 GREG estimator.

Let $\hat{\theta}_i$ denote the direct estimates of the target variables θ_i for the domains $i = 1, \dots, m$. In this application $\hat{\theta}_i$ is the GREG estimator. In the case of 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. Furthermore, 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

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

Furthermore, 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 paper, 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. The HB estimates are computed with R (R Development Core Team, 2009) using package `hbsae` (Boonstra, 2012).

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.

4 Model selection

4.1 Auxiliary information

The success of increasing the precision of the domain estimates with the area level model from section 3 critically depends on the availability of correlated auxiliary

information. For example, it can be expected that crime victimization figures from the ISM correlate with the number of reported offences available from the Police Register of Reported Offences (PRRO). In addition, correlation is expected between crime and safety related survey variables and various socio-demographic variables, which are in the Netherlands available from municipal administrations. An overview of 21 potential auxiliary variables used for model building is given in Appendix A.

Some additional explanation is required for the auxiliary variables `mode2` and `oversampled`. The ISM is based on a sequential mixed-mode design. In areas where local authorities draw additional samples, the fraction non-interviewer administered modes is larger while the fraction of interviewer administered modes is smaller, compared to areas where no additional samples are drawn. There are also clear indications that there are systematic differences in measurement error between responses obtained with interviewer and non-interviewer administered modes, (Buelens and Van den Brakel, 2014; Schouten et al., 2013). As explained in section 2, 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, 2014). 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, as these covariates may be correlated with potential mode-dependent measurement error in municipal estimates.

4.2 Models for repeated surveys

The starting point is the area level model (1). This model will be applied to produce municipal estimates in the ISM, which is an annually repeated survey. For this research four subsequent years are available; 2008 through 2011. The question that is addressed is how to find an optimal but robust cross-sectional model on the basis of the information in these four available years and to illustrate the kind of problems that are encountered by implementing an SAE approach in a repeated cross-sectional survey.

4.2.1 Repeating one model unaltered

The most straightforward approach is to select a model once, using the first edition of the survey, and continuing to use that model in subsequent editions. The model is written as

$$\hat{\theta}_{i,t} = z_{i,t}^{[rf08]'} \beta_t + v_{i,t} + e_{i,t}. \tag{3}$$

for $i = 1, \ldots, m$, and $t = 2008, \ldots, 2011$ and with model assumptions

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

The vector $z_{i,t}^{[rf08]}$ consists of the values for area i at time t of the covariates selected using the 2008 edition of the survey, which serves as the reference (rf) survey. Clearly, the covariates selected using 2008 data may not be optimal in subsequent years.

4.2.2 Separate models for each year

The most obvious alternative is to select for each separate year an optimal model, independently from the other years, i.e.

$$\hat{\theta}_{i,t} = z_{i,t}^{[annu]'}\beta_t + v_{i,t} + e_{i,t}, \tag{5}$$

with model assumptions (4) and $z_{i,t}^{[annu]}$ a vector of covariates obtained through model selection conducted using survey data collected in time period t ($annu$ stands for annual). This approach can result in different models for the subsequent years. If small area predictions and their MSE estimates are not robust for the selected covariates in the model, then large fluctuations in the selected covariates raise questions about the stability of the small area predictions, their MSE estimates and the interpretation of the results over time.

4.2.3 Combining the available years

One possibility to avoid differences between the selected models over the years, is to pool the survey data of the different editions and to model the small area estimates of all available years simultaneously. This implies that the units in the model are the cross classification of small areas and years. The model is defined as

$$\hat{\theta}_{i,t} = \alpha_t + z_{i,t}^{[pool]'}\beta + v_{i,t} + e_{i,t}. \tag{6}$$

with model assumptions

$$v_{i,t} \overset{iid}{\sim} N(0, \sigma_v^2) \;\; and \;\; e_{i,t} \overset{ind}{\sim} N(0, \psi_{i,t}). \tag{7}$$

In this case $z_{i,t}^{[pool]}$ contains the same set of covariates for each year and the regression coefficients are equal over time ($pool$ stands for pooled). Systematic differences between the years are modeled with a fixed effect α_t. A drawback of this approach is the implementation in a production system for official statistics. Each year a new set of covariates becomes available, which might result in an adjustment of the model as well as a revision of the small area predictions of the preceding periods. Pooling the cross classifications of domains and years implies that the model does not discriminate between temporal effects and cross-sectional effects. Models with a time series component, mentioned in Subsection 4.2.5, allow for different structures between both effects.

4.2.4 Each year the same model

Instead of selecting each year an optimal model, it is also possible to apply the same model to each year separately. The model is defined as

$$\hat{\theta}_{i,t} = z_{i,t}^{[avrg]'} \beta_t + v_{i,t} + e_{i,t}. \tag{8}$$

with model assumptions (4). The model selection is based on the data of the available years. A model that is on average optimal over the available years is selected by minimizing the average value of some selection criterion (see next section for details). The vector $z_{i,t}^{[avrg]}$ refers to the resulting set of covariates ($avrg$ stands for average). Once a model is selected, it is used in production to produce small area prediction without adjusting the model when new data become available.

4.2.5 Models with a time series component

The aforementioned models use cross sectional information observed in other domains to increase the precision of the small are estimates. In the approach of section 4.2.3 data from other time periods is used through the pooled estimation of the model parameters. More advanced time series models could be used to take advantage of sample information observed in preceding periods. Rao and Yu (1994) and Datta et al. (1999) extended the area level model with a time series component to combine cross-sectional data with information observed in preceding periods. A different approach is to use a state-space model to combine time series data with cross sectional data, see e.g. Pfeffermann and Burck (1990) and Pfeffermann and Bleuer (1993). These more advances approaches are currently not considered for implementation in the production of official small area figures in repeated surveys.

4.3 Selection criteria

The covariates for the models are selected from a set of suitable auxiliary variables through a step forward variable selection procedure. Two criteria for variable selection are used as a comparison measure to select the most suitable models.

The first criterion is the conditional Akaike Information Criterion (cAIC). The cAIC is proposed by Vaida and Blanchard (2005) for mixed models where the focus is on prediction at the level of clusters or areas. It is defined as $cAIC = -2L + 2p$, where 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 cAIC, p is the effective degrees 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).

The second criterion is based on cross validation (CRV). This measure is an indication of the predictive power of the model for the data that are not used to fit the model. In this

application the "leave one out" method is used, which implies that the model is applied m times to the data, leaving out one domain each time. The prediction for the domain that is left out, is compared to the observed value of that domain. The CRV is defined as

$$CRV = \left(\sum_{i=1}^{m} w_i\right)^{-1} \sum_{i=1}^{m} w_i(\hat{\theta}_{i,t} - \tilde{\theta}_{i,t}^{-i})^2, \tag{9}$$

with $\tilde{\theta}_{i,t}^{-i}$ the HB prediction for domain i estimated from the data excluding domain i and w_i an adjustable weight, Hastie et al. (2003). In the present study the weights are defined as $w_i = \left(\psi_{i,t} + \hat{\sigma}_{v,t}^2\right)^{-1}$.

An additional criterion is the mean reduction in the coefficient of variation (MRCV) over the domains with respect to the GREG estimator. This criterion is evaluated in the ISM case study but it is not used as a model selection criterion. The MRCV measures the increased precision of the small area predictions and is defined as:

$$MRCV = \frac{1}{m} \sum_{i=1}^{m} \frac{CV(\hat{\theta}_{i,t}) - CV(\tilde{\theta}_{i,t})}{CV(\tilde{\theta}_{i,t})}, \tag{10}$$

with $\tilde{\theta}_{i,t}$ 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).

The cAIC and CRV are used in a step forward selection procedure. The procedure starts with a model that only contains an intercept. Subsequently covariates are added one by one, each time selecting the covariate that results in the largest decrease of the cAIC or CRV, until no further improvement is possible or all potential covariates are selected.

5 Results

5.1 Covariate selection results

The covariate selection approaches proposed in sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4 are applied to the ISM editions of 2008 through 2011, for the three survey variables listed in Table 2. Table 3 contains the results for the variable `victim` for the four different approaches. The available covariates are shown in the columns. When a covariate is selected in a particular year using a particular selection approach, the corresponding cell in the table is colored orange and contains a number indicating the order in which the covariate was selected; i.e. covariate 1 got selected first, covariate 2 second, etc. Consequently, the highest number in a given row corresponds to the number of covariates included in the model for that instance. All models include an intercept.

The top section of Table 3 contains the results of selecting a model in 2008 and applying it in the years thereafter. Both results based on CRV and cAIC are given. Selecting

covariates based on the cAIC criterion results in different covariates than with the CRV criterion. The second section of the Table contains results of selecting models each year, in an independent fashion (see section 4.2.2). While there is considerable overlap among the sets of covariates selected in the four years, they are all different. There are differences between the years as well as between the selection criteria. Overall, the models chosen using cAIC are more parsimonious than those obtained using CRV. The criteria CRV and MRCV can be compared between the years. The CRV gives better scores in the years with more oversampling, 2009 and 2011. This is not surprising since in these years more data are available with lower design variance. The results of this approach are also better than the results obtained through the first approach, as expected. The variance reduction measured through the MRCV is larger in the years with less oversampling, 2008 and 2010, since small area predictions and GREG estimates become more and more equally efficient if the sample size increases.

The third section of the table contains the result of pooling data for all four years and fitting a single model (see section 4.2.3). As there are four times as many data points compared to the single year approach, the resulting models are larger. The cAIC and CRV approaches lead to different models, although the first eight covariates selected are identical and occur in the same order.

Results of joint model selection for all four years by averaging the model selection criterion over the years (see section 4.2.4) are shown in the bottom section of Table 3. The cAIC based model is smaller than the CRV based one, while the differences in cAIC, CRV and MRCV are almost negligible.

From an official statistics perspective, where outcomes of surveys are monitored through time and where temporal change of quantities of interest is of primary importance, it is desirable to produce statistics using a process that does not change between survey editions. Otherwise, changes in the process could confound with true changes in the quantity to be estimated. If small area predictions are sensitive for the selected covariates, then using different models every year is to be avoided. The fourth approach using averages is attractive because it results in a model specification that can be used in future years not included in this analysis. Doing so using the pooled approach would alter all estimates of earlier years, which would involve publishing revised results, an undesirable situation in official statistics.

Results for the variables `degen` and `contpol` are given in Tables 4 and 5 respectively. The conclusions drawn based on the results for `victim` largely apply to these two other variables, although some differences are seen. For `degen`, the averaging approach resulting in the same model for every year (bottom section in Table 4) provides models that are not smaller than the annual models (second section), and the CRV and cAIC models contain the same covariates except for one. All models selected for `contpolg` -- see Table 5 -- are smaller than the models for the other two variables. The MRCV values indicate that these models are exceptionally powerful in the sense that they dramatically reduce the variance of the direct estimates.

A tentative conclusion from this section is that the method using averages of the cAIC model selection criteria is preferred. This method is best suited to the requirements of official statistics and results in relatively parsimonious models, performing well on the

data used to establish them. These models are applicable to future editions of the survey as well without requiring a revision of earlier publications. The question is to which extent this approach results in less optimal models compared to the annually selected optimal models.

5.2 Model comparisons

In this section the point and MSE estimates resulting from the different models discussed above are compared. For simplicity, the four approaches to model selection are identified and abbreviated by *rf08*, *annu*, *pool* and *avrg* indicating the approaches of sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4 respectively. The different alternatives considered in this section all use cAIC as the selection criterion, as the resulting models are preferred to those obtained through CRV, as concluded in the previous section.

Table 6 lists the estimates of the between area variance parameter ($\hat{\sigma}_v^2$) for the different approaches. The estimates in the *annu* approach are smaller than those obtained through the *rf08* and *avrg* approaches, in line with the fact that the *annu* estimates are optimal each year. In the *pool* approach, only one variance parameter is estimated combining spatial and temporal effects, with a value often but not always lower than in the *annu* approach.

Scatterplots of the point estimates resulting from models obtained through the approaches *rf08*, *pool* and *avrg* are compared to the *annu* estimates in Figures 1, 2 and 3 respectively for the variables `victim`, `degen` and `contpol`. The *rf08*-estimates and *pool*-estimates generally deviate more from the *annu*-estimates than the *avrg*-estimates do. Note that for the year 2008 the *annu*-estimates are equal to the *rf08*-estimates by definition. For `degen`, the 2008 model appears to work reasonably well. For the other two variables the *avrg* approach is better. For `contpol` the annual selected models are larger in the years of large oversampling, 2009 and 2011. This explains the larger differences between annual and average selected models in these two years (bottom panel of Figure 3).

The differences between two sets of estimates are quantified through defining the Mean Relative Absolute Difference (MRD) as

$$MRD(\hat{\theta}^{mod_1}, \hat{\theta}^{mod_2}) = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{\theta}_i^{mod_2} - \hat{\theta}_i^{mod_1}}{\hat{\theta}_i^{mod_1}}\right| \tag{11}$$

for two sets of estimates for $\hat{\theta}$ using two different models mod_1 and mod_2 referring to models resulting from one of the scenarios under consideration, *rf08*, *annu*, *pool* or *avrg*. The MRDs between the *rf08*, *pool* and *avrg* estimates compared to the *annu* estimates are given in Table 7. The estimates obtained through the *pool* approach deviate more from the *annu* estimates than the *avrg* estimates, except for the variable `contpol` in 2011. The *rf08* approach shows mixed results. For the variable `degen` these estimates always differ more from the *annu* estimates than the *avrg* estimates. For `victim` and `contpol`, the differences are comparable or even smaller in some years, while they are

larger in other years. Overall, the estimates obtained through the *avrg* approach are closest to the *annu* estimates and emerge as the preferred set of estimates.

Table 8 contains the standard errors (SEs) of all estimates averaged over the domains. The SEs resulting from the *annu* and *pool* approaches are comparable, with the *pool* SEs being smaller in more cases. Clearly there is benefit in pooling data from a variance perspective. The SEs obtained through the *avrg* approach are generally somewhat larger, which is explained by the fact that these models are suboptimal when considered on a year-by-year basis. Using the model resulting from the *rf08* approach may result in good SEs, such as for the variable `contpol` where the *rf08* SEs are smaller than the *avrg* SEs. Note that only the years 2009-2011 are relevant to consider in this case. For the other two variables, the *avrg* SEs are generally smaller than or almost equal to the *rf08* SEs.

From the analysis in this section it is concluded that the *avrg* approach is preferred from a bias perspective, because it results in estimates that differ less from the *annu* estimates than the *pool* and *rf08* estimates. From a variance perspective, the *pool* estimates are preferred. Using the *rf08* estimates may work well in some years for some variables. In combination with the findings in the previous section, the *avrg* approach emerges as the preferred approach to follow in the present study.

A question arising naturally is whether all four years are needed in this approach, or whether using fewer years would result in the same model. This question is addressed in Table 9. For each of the three survey variables, models are selected using the averaging approach. First the survey of 2008 is used, then data of 2009, 2010 and 2011 are added one by one in a cumulative fashion. For `victim`, the 2008 model is large, containing 10 covariates. Adding more data reduces the number of covariates in the model while those covariates that are retained stabilize, at least to some extent. For `degen` the models become larger, with the set of selected variables the same for the 2008-2010 and 2008-2011 runs, although they are selected in a different order. The 2008 model for `contpol` contains a single covariate, `logdens`. Adding more data causes the model to increase, always containing `logdens` and adding different variables in different runs. The overall image that emerges is that there is some convergence towards a stable set of covariates as data of more years are added, but some variability remains. In addition, it is not always the case that the models resulting from averaging more editions are more parsimonious. Care is needed in scenarios where a model selected using data from several survey editions is applied in subsequent editions. Given some apparent instability, it is advisable to carefully check models when they are used with data that was not used to establish the model.

5.3 Small area estimates

While the focus of the present research is on model selection, the ultimate goal of the models is to provide small area estimates for the survey variables of interest. In this section results obtained using the models selected through the *avrg* approach applied to all four editions are presented. These models are given in Table 9, on the rows referring to the 2008-2011 runs.

Figures 4, 5 and 6 show the resulting point estimates and estimated standard errors (SEs) for the variables `victim`, `degen` and `contpol` respectively. The top panels show a scatter plot of the direct GREG estimates on the horizontal axis and the corresponding SAE estimates on the vertical axis, with the diagonal in black. Areas in which oversampling was conducted are plotted in blue, the others in red. All three variables exhibit a regression-to-the-mean effect: the dispersion of the SAE estimates is smaller than the dispersion of the GREG estimates. This effect is strongest for `contpolg` and weakest for `degen`. Areas without oversampling -- in which the sample sizes are smaller -- suffer more from this effect since the small area estimator attaches more weight to the synthetic component of the predictor if the design variance of a domain increases. The bottom panels show the estimated SEs, with areas ordered along the horizontal axis according to increasing sample size. SEs of the SAE estimates are comparable for all areas while the SEs of the GREG estimates decrease with increasing sample size. The biggest reduction in SEs is achieved for areas that are not oversampled, shown in red. For `degen`, oversampled areas do not benefit from the SAE approach as SEs of both sets of estimates are almost equal. For `contpol`, the SEs of SAE estimates are smaller than these of the GREG estimates both in oversampled and non-oversampled areas.

Time series of the estimates for the three variables are shown in 7. Nine municipalities are selected, with varying sample sizes. The number at the top of each panel in the plot refers to the rank of the municipality when ordered according to increasing sample size. GREG estimates are shown using circles connected by solid lines. SAE estimates obtained through the averaging approach are represented by triangles connected by dashed lines. Color codes indicating oversampling are as before. Generally, the series of SAE estimates are smoother than the GREG series, in particular for areas without oversampling. In Amsterdam, the largest municipality with the largest sample, the GREG and SAE estimates almost coincide. In the smaller municipalities, the differences are largest when there is no oversampling. When studying these results it is important to be aware of the fact that the GREG estimates for areas that are not oversampled are not published, and will not be. The model based SAE estimates obtained in this research are specifically aimed at those areas.

SAE estimates obtained through the averaging and annually optimal approaches are compared with GREG estimates in a similar manner as the previous plot, in Figure 8. This plot is simplified for clarity: color codes for oversampling and different symbols for different estimates are omitted. The colors now represent the three types of estimates that are compared. The differences between the two types of SAE estimates are very small, almost negligible in most cases. This suggests that even though the averaging approach may be preferred from an official statistics viewpoint, using the annually optimal models instead would not result in substantially different results in this case study.

5.4 Model evaluation

The model specification for the four different approaches is evaluated through two diagnostic analyzes. The benchmark diagnostic compares aggregated small area estimates with the GREG estimate at national level. The distributions diagnostic considers the distributions of the residues and of the random effects.

Table 10 presents the benchmark diagnostic. Relative differences are shown between aggregated small area estimates and GREG estimates of country level totals. Since GREG estimates at this level are based on the full sample they are regarded as sufficiently precise. It is desirable for the small area estimates to produce estimates without systematic bias. Large deviations from the aggregated GREG estimate are indicative of model failure (Brown et al., 2001). In the table, the least biased approach for each variable for each year is indicated with a #-symbol. The *rf08* approach results in aggregates that are more biased than the other approaches, apart for `victim` in 2008. The *annu* models are the least biased in more instances than all other approaches. This is in line with these models being optimal on a year-by-year basis. The *pool* and *avrg* approaches score well in some cases. Generally the relative bias of the *avrg* approach is smaller than that of the *pool* approach, another indication that the *pool* models are not to be preferred, possibly due to the pooled modeling of cross sectional and temporal variability.

The distributions of the residues and random effects are diagnosed by means of quantile-quantile (QQ) plots in Figure 9. Only an evaluation for 2011 is shown; the other survey years exhibit very similar characteristics. The top panel shows QQ plots for the residues for all four approaches for each of the variables. The patterns for `degen` are more linear than for the two other variables, although all plots deviate from linear in a similar manner, indicative of a left-skewed distribution. QQ plots of the random effects are shown in the bottom panel exhibiting again the same pattern. There are almost no differences between results for the different variables and the different approaches. As a deviation from linear is clearly seen, all considered models appear to be misspecified to some extent. Further research into non-normally distributed random effects and residues is warranted.

6 Conclusion

The issue considered in this paper is the choice of covariates in models when applying small area estimation repeatedly in consecutive editions of a survey -- assume annual surveys for simplicity and without loss of generality. 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.

Four alternatives are proposed and compared. The first and easiest method is to choose an optimal set in the first edition of the survey and repeatedly use this set unchanged in subsequent editions. While practical, this approach is liable to the risk of choosing a set of covariates that happens to perform well in that first edition but that is suboptimal in future editions. In the ISM case study, this is indeed the case. When data of multiple survey editions are available, there is benefit in using all the data to select an optimal model.

The second approach is to establish an optimal set of covariates every year. This approach is not guaranteed to result in the same set of covariates every year. In the

application of the ISM, there is considerable variation in sets of covariates that are selected. From an official statistics perspective, this is not desirable since comparability of published statistics through time is essential. Variations in covariates could entail variations in outcomes which could be confounded with true changes of the quantity of interest -- a situation to be avoided.

The third alternative is a pooled approach, considering survey data of multiple survey editions simultaneously. The model specification includes the time period specifically as a covariate and allows for fitting a single model for all editions combined. This approach has the benefit of resulting in a single model with a unique set of covariates. Drawbacks are that the complexity increases with increasing numbers of surveys, the model not allows for different structures for the spatial and temporal effects, and that all past estimates need updating when new survey editions become available. The ISM case study further shows that the resulting estimates deviate a lot from the optimal approach where an optimal model is chosen every year.

The fourth and final alternative is an averaging approach resulting in a single set of covariates to be used every year. Data of multiple editions are used independently -- in contrast to the pooled approach. The model selection procedure considers all available editions and proceeds by selecting covariates that perform best on average for all editions. The same model is used for all editions, independently. This approach has the advantage of resulting in a single model that can be used for all editions, both those that are available already and future ones. Future editions do not necessitate a revision of the published statistics. The ISM case study suggests that the set of selected covariates seems to settle from the third editions and appears to be sustainable into the future. The models selected in this way are suboptimal when considered on a year-by-year basis, but perform better than the pooled approach.

While selecting an optimal model every year may be preferable from a modeling perspective, in official statistics it is important to avoid all potentially confounding elements in temporal change in published time series. Using the same model every year is considered essential. Pooling several available editions of a survey may necessitate revisions of past publications, an undesirable situation. The averaging method is the best of the two approaches that use the same model in all years, since it results in more stable models than the approach in which only one edition is used to select covariates. 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 differences among the small area estimates obtained with the four different approaches are not very large in the case of the ISM. The choice for the averaging approach is motivated more by the characteristics and principles of the different approaches than by substantial differences in statistical results. In other applications, actual differences between the approaches might be larger.

All but the first approach use survey data of multiple years. It remains a challenge to use data of only the first edition to establish a model and set of covariates, and to be able to continue to use that in subsequent editions. The idea was briefly explored to use a bootstrap approach to achieve this, by conducting model selection procedures for bootstrap samples drawn from the first edition of the survey. Methodological problems

related to bootstrapping multilevel data have prevented further development of this line of research so far. In particular, naive bootstrap estimates of the between area variance are biased upward when the between area variance is small compared to the within area variance (Davison and Hinkley, 1997). As a result, the model selection criteria used in the covariate selection procedures too are biased, rendering these procedures unreliable. Further research is needed into bootstrap estimation of model selection criteria.

Other topics for further research are the use of time series multilevel models as an alternative method to allow for temporal and cross-sectional effects simultaneously (Rao and Yu, 1994; Datta et al., 1999), and the extension of the models to include spatial effects (You and Zhou, 2011).

References

Battese, G. E., Harter, R. M., and 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*. Statistics Netherlands.

Boonstra, H. J., Van den Brakel, J., Buelens, B., Krieg, S., and Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *METRON International Journal of Statistics* LXVI (1), 21--49.

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of Small Area Estimation Methods - An Application to Unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium 2001*.

Buelens, B. and Van den Brakel, J. (2014). Measurement error calibration in mixed mode surveys. *Sociological Methods and Research* online, 1--36.

Datta, G., Lahiri, P., Maiti, T., and Lu, K. (1999). Hierarchical Bayes estimation of unemployemtn rates for the states of the U.S. *Journal of the Royal Statisitcal Society* 94, 1074--1082.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press.

Fay, R. E. and 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.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The elements of statistical learning*. Springer-Verlag, New York.

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

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663--685.

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. and Bleuer, S. R. (1993). Robust Joint Modelling of Labour Force Series of Small Areas. *Survey Methodology* 19, 149--163.

Pfeffermann, D. and Burck, L. (1990). Robust Small Area Estimation combining Time Series and Cross-sectional Data. *Survey Methodology* 16, 217--237.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

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.

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

Schouten, B., Van den Brakel, J., Buelens, B., Van der Laan, J., and Klausch, T. (2013). Disentangling mode-specific selection bias and measurement bias in social surveys. *Social Science Research* 42, 1555--1570.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* 92, 351--370.

Van den Brakel, J. and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modeling in a rotating panel design. *Survey Methodology* 35, 177--190.

Van den Brakel, J. and Krieg, S. (2014). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel. *Survey Methodology* under review, --.

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

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

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.

Tables

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

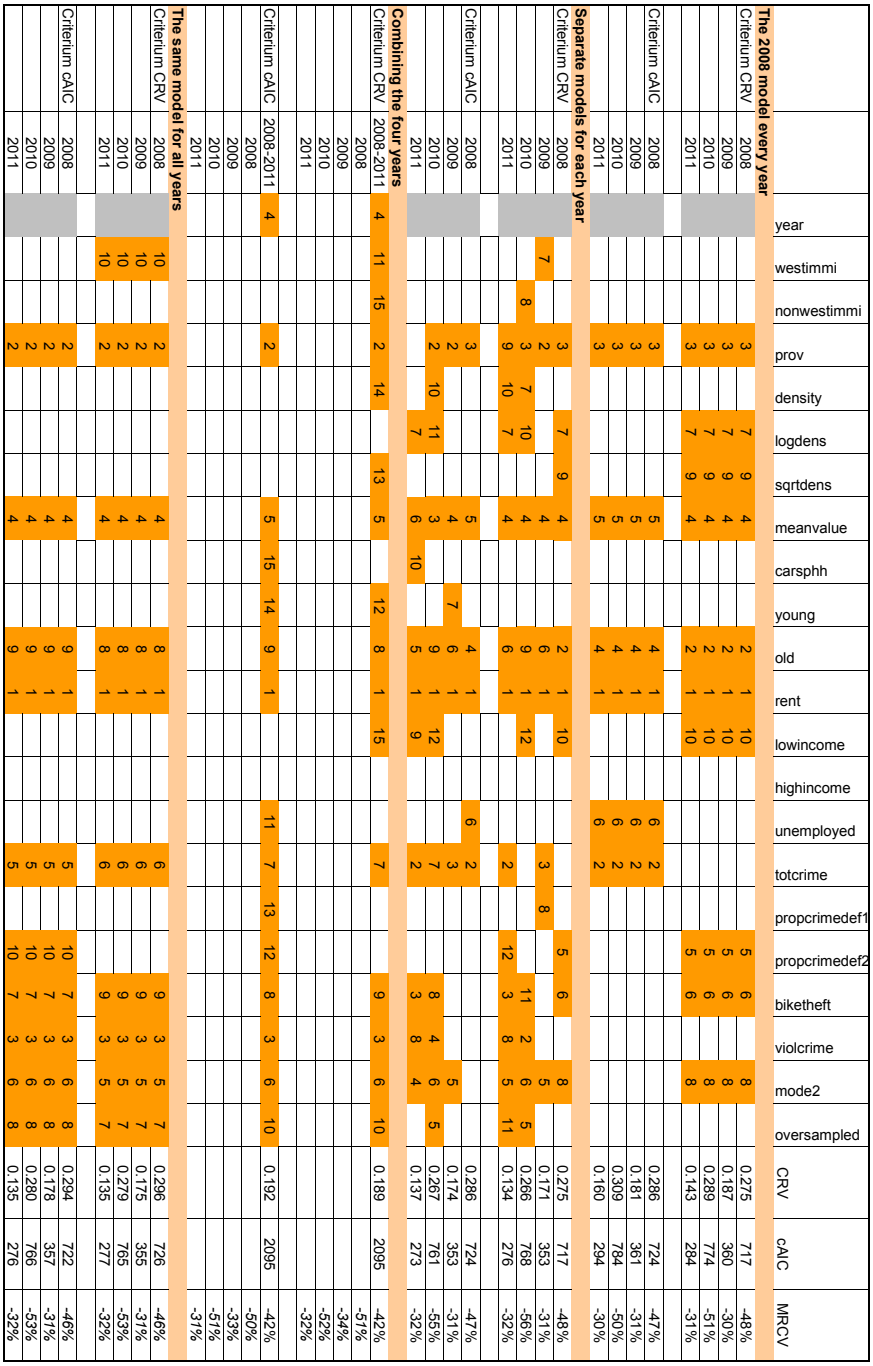
Table 1 Overview of oversampling in ISM surveys 2008 - 2011.

Variable	Description of statistic
victim	Percentage of people indicating to have been a victim of crime in the last 12 months
degen	Degeneration of the neighborhood (on a scale 1-5)
contpol	Percentage of people who have had contact with the Police in the last 12 months

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

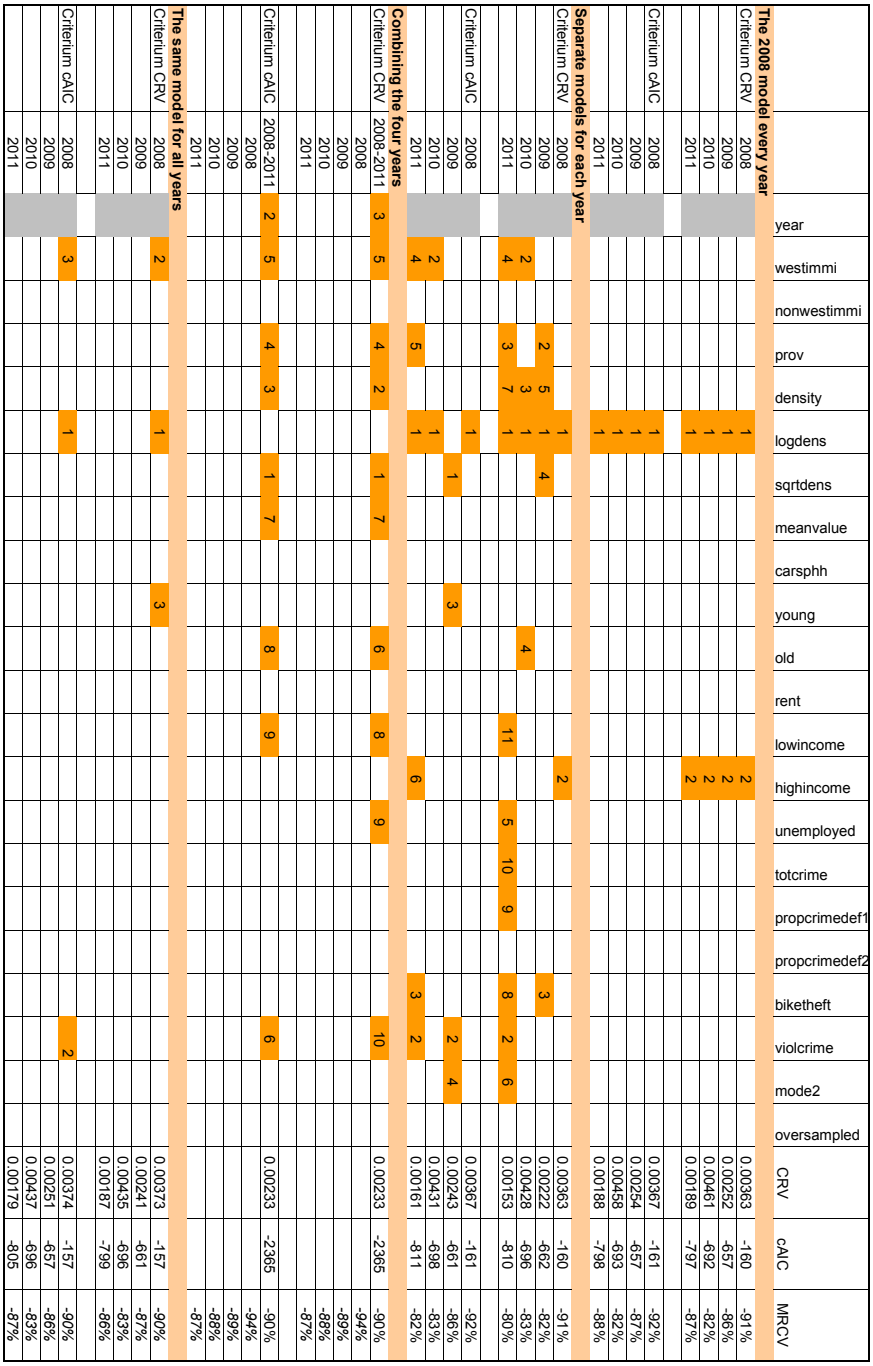
		year	westimmi	nonwestimmi	prov	density	logdens	sqrtdens	meanvalue	carsphh	young	old	rent	lowincome	highincome	unemployed	totcrime	propcrimedef1	propcrimedef2	biketheft	violcrime	mode2	oversampled	CRV	cAIC	MRCV
The 2008 model every year																										
Criterium CRV	2008		5				1			6		2		8	7			9	4			10	3	0.00258	-948	-77%
	2009		5				1			6		2		8	7			9	4			10	3	0.00143	-1306	-56%
	2010		5				1			6		2		8	7			9	4			10	3	0.00245	-906	-83%
	2011		5				1			6		2		8	7			9	4			10	3	0.00103	-1385	-60%
Criterium cAIC	2008		6	4		9	1			10		5		8	7			2				3		0.00279	-949	-76%
	2009		6	4		9	1			10		5		8	7			2				3		0.00152	-1304	-55%
	2010		6	4		9	1			10		5		8	7			2				3		0.00250	-905	-83%
	2011		6	4		9	1			10		5		8	7			2				3		0.00103	-1393	-60%
Separate models for each year																										
Criterium CRV	2008		5				1			6		2		8	7			9	4			10	3	0.00258	-948	-77%
	2009		6		3	8		1			4	9		10	11			2	2		7	5	0	0.00134	-1310	-57%
	2010			6	4			2	8		3			5			7	1					0	0.00231	-910	-80%
	2011																							0.00092	-1395	-64%
Criterium cAIC	2008		6	4		9	1			10		5		8	7			2				3		0.00279	-949	-76%
	2009		6					1		3		5		8	7			2	2			4		0.00142	-1308	-57%
	2010								8		3						4	1						0.00241	-914	-88%
	2011							1	2		1	3	6	9				2				5	10	0.00093	-1395	-63%
Combining the four years																										
Criterium CRV	2008-2011	3	7		4	8		1			5	9		11	12			15	2	14	10	13	6	0.00154	-4565	-69%
	2008																									-77%
	2009																									-62%
	2010																									-79%
	2011																									-59%
Criterium cAIC	2008-2011	3	7	9	4	8		1			5	13	11	14	15				2	10	12	16	6	0.00155	-4566	-69%
	2008																									-76%
	2009																									-61%
	2010																									-79%
	2011																									-78%
The same model for all years																										
Criterium CRV	2008		6		3		1				4	7						2		5	8		4	0.00279	-934	-72%
	2009		6		3		1				4	7						2		5	8		4	0.00137	-1308	-57%
	2010		6		3		1				4	7						2		5	8		4	0.00241	-902	-79%
	2011		6		3		1				4	7						2		5	8		4	0.00102	-1381	-59%
	2008																									-59%
Criterium cAIC	2008		6				1				3							5	2				4	0.00309	-938	-72%
	2009		6				1				3							5	2				4	0.00146	-1304	-56%
	2010		6				1				3							5	2				4	0.00247	-910	-86%
	2011		6				1				3							5	2				4	0.00101	-1386	-60%

Table 3 Selected models for the survey variable `victim`.



		year	westimmi	nonwestimmi	prov	density	logdens	sqrtdens	meanvalue	carsphh	young	old	rent	lowincome	highincome	unemployed	totcrime	propcrimedef1	propcrimedef2	biketheft	violcrime	mode2	oversampled	CRV	cAIC	MRCV
The 2008 model every year																										
Criterium CRV	2008				3		7	9	4			2	1	10					5	6		8		0.275	717	-48%
	2009				3		7	9	4			2	1	10					5	6		8		0.187	360	-30%
	2010				3		7	9	4			2	1	10					5	6		8		0.289	774	-51%
	2011				3		7	9	4			2	1	10					5	6		8		0.143	284	-31%
Criterium cAIC	2008				3		7	9	5			4	1	10		6	2							0.286	724	-47%
	2009				3		7	9	5			4	1	10		6	2							0.181	361	-31%
	2010				3		7	9	5			4	1	10		6	2							0.309	784	-50%
	2011				3		7	9	5			4	1	10		6	2							0.160	294	-30%
Separate models for each year																										
Criterium CRV	2008				3		7	9	4			2	1	10					5	6		8		0.275	717	-48%
	2009		7		2		4		4			6	1				3	8				5		0.171	353	-31%
	2010			8	3	7	10		4			9	1	12			2		12	11	2	6	5	0.266	768	-56%
	2011				9	10	7		4			6	1							3	8	5	11	0.134	276	-32%
Criterium cAIC	2008				3		7	9	5			4	1	10		6	2							0.286	724	-47%
	2009				2	10	11		4		7	6	1				3							0.174	353	-31%
	2010				2				3	10		9	1	12			7			8	4	6		0.267	761	-55%
	2011				2				6			5	1	9			2			3	8	4	5	0.137	273	-32%
Combining the four years																										
Criterium CRV	2008-2011	4	11	15	2	14		13	5		12	8	1	15			7			9	3	6	10	0.189	2095	-42%
	2008																									-51%
	2009																									-34%
	2010																									-52%
	2011																									-32%
Criterium cAIC	2008-2011	4			2				5	15	14	9	1			11	7	13	12	8	3	6	10	0.192	2095	-42%
	2008																									-50%
	2009																									-33%
	2010																									-51%
	2011																									-31%
The same model for all years																										
Criterium CRV	2008	10			2				4			8	1			6				9	3	5	7	0.296	726	-46%
	2009	10			2				4			8	1			6				9	3	5	7	0.175	355	-31%
	2010	10			2				4			8	1			6				9	3	5	7	0.279	765	-53%
	2011	10			2				4			8	1			6				9	3	5	7	0.135	277	-32%
Criterium cAIC	2008				2				4			9	1			5			10	7	3	6	8	0.294	722	-46%
	2009				2				4			9	1			5			10	7	3	6	8	0.178	357	-31%
	2010				2				4			9	1			5			10	7	3	6	8	0.280	766	-53%
	2011				2				4			9	1			5			10	7	3	6	8	0.135	276	-32%

Table 4 Selected models for the survey variable degen.



		year	westimmi	nonwestimmi	prov	density	logdens	sqrtdens	meanvalue	carsphh	young	old	rent	lowincome	highincome	unemployed	totcrime	propcrimedef1	propcrimedef2	biketheft	violcrime	mode2	oversampled	CRV	cAIC	MRCV
The 2008 model every year																										
Criterium CRV	2008						1								2									0.00363	-160	-91%
	2009						1								2									0.00252	-657	-86%
	2010						1								2									0.00461	-692	-82%
	2011						1								2									0.00189	-797	-87%
Criterium cAIC	2008						1																	0.00367	-161	-92%
	2009						1																	0.00254	-657	-87%
	2010						1																	0.00458	-693	-82%
	2011						1																	0.00188	-798	-88%
Separate models for each year																										
Criterium CRV	2008						1								2									0.00363	-160	-91%
	2009		2		2	5	1	4												3				0.00222	-662	-82%
	2010					3	1					4												0.00428	-696	-83%
	2011		4		3	7	1						11			5	10	9		8	2	6		0.00153	-810	-80%
Criterium cAIC	2008						1																	0.00367	-161	-92%
	2009						1	1			3										2	4		0.00243	-661	-86%
	2010		2		5		1								6					3	2			0.00431	-698	-83%
	2011		4				1									9						10		0.00161	-811	-82%
Combining the four years																										
Criterium CRV	2008-2011	3	5		4	2		1	7			6	8			9								0.00233	-2365	-90%
	2008																									-94%
	2009																									-89%
	2010																									-88%
	2011																									-87%
Criterium cAIC	2008-2011	2	5		4	3		1	7			8	9									6		0.00233	-2365	-90%
	2008																									-94%
	2009																									-89%
	2010																									-88%
	2011																									-87%
The same model for all years																										
Criterium CRV	2008		2				1				3													0.00373	-157	-90%
	2009																							0.00241	-661	-87%
	2010																							0.00435	-696	-83%
	2011																							0.00187	-799	-86%
Criterium cAIC	2008		3				1															2		0.00374	-157	-90%
	2009																							0.00251	-657	-86%
	2010																							0.00437	-696	-83%
	2011																							0.00179	-805	-87%

Table 5 Selected models for the survey variable `contpol`.

variable	year	rf08	annu	pool	avrg
victim	2008	2.17E-04	2.17E-04	2.55E-04	3.72E-04
	2009	4.18E-04	3.88E-04	2.55E-04	4.18E-04
	2010	6.30E-05	5.60E-05	2.55E-04	6.06E-05
	2011	2.37E-04	1.74E-04	2.55E-04	2.39E-04
degen	2008	1.00E-01	1.00E-01	9.04E-02	1.00E-01
	2009	1.03E-01	1.01E-01	9.04E-02	1.02E-01
	2010	9.06E-02	5.67E-02	9.04E-02	6.72E-02
	2011	9.92E-02	8.60E-02	9.04E-02	7.97E-02
contpol	2008	2.08E-04	2.08E-04	3.76E-05	2.26E-04
	2009	1.73E-04	1.39E-04	3.76E-05	1.72E-04
	2010	3.51E-04	2.65E-04	3.76E-05	2.76E-04
	2011	1.01E-04	5.63E-05	3.76E-05	9.66E-05

Table 6 Estimates of the variance parameter of the area effects ($\hat{\sigma}_v^2$).

variable	year	rf08	pool	avrg
victim	2008	-	5.6	5.3
	2009	2.4	3.8	2.3
	2010	3.8	5.2	1.3
	2011	2.5	4.2	2.7
degen	2008	-	2.9	1.7
	2009	1.6	1.7	0.6
	2010	4.4	3.3	1.8
	2011	2.8	2.4	1.6
contpol	2008	-	6.3	0.8
	2009	4.6	6.3	4.8
	2010	3.2	5.4	0.2
	2011	7.6	4.8	6.4

Table 7 MRD (in %) for three alternatives compared to the annual models.

variable	year	rf08	annu	pool	avrg
victim	2008	.0170	.0170	.0167	.0197
	2009	.0186	.0177	.0154	.0184
	2010	.0129	.0095	.0169	.0109
	2011	.0150	.0134	.0152	.0148
degen	2008	.2735	.2735	.2539	.2769
	2009	.2147	.2140	.2044	.2161
	2010	.2721	.2383	.2610	.2495
	2011	.2059	.1946	.1972	.1957
contpol	2008	.0198	.0198	.0144	.0235
	2009	.0144	.0153	.0123	.0154
	2010	.0195	.0179	.0124	.0187
	2011	.0112	.0190	.0120	.0124

Table 8 Mean standard error (SE) for all alternatives.

Variable	Editions	Covariates in the model
victim	2008	logdens, propcrimedef2, oversampled, nonwestimmi, old, westimmi, highincome, lowincome, density, carsphh
	2008-2009	sqrtdens, propcrimedef2, oversampled, nonwestimmi, biketheft, density
	2008-2010	sqrtdens, propdrimedef1, oversampled, young, westimmi
	2008-2011	sqrtdens, propcrimedef1, young, oversampled, totcrime, westimmi
degen	2008	rent, totcrime, prov, old, meanvalue, unemployed
	2008-2009	rent, prov, totcrime, meanvalue, mode2, old, westimmi
	2008-2010	rent, prov, violcrime, meanvalue, totcrime, mode2, oversampled, old, biketheft, propcrimedef2
	2008-2011	rent, prov, violcrime, meanvalue, totcrime, mode2, biketheft, oversampled, old, propcrimedef2
contpol	2008	logdens
	2008-2009	logdens, young
	2008-2010	logdens, westimmi, young
	2008-2011	logdens, violcrime, westimmi

Table 9 Results of the averaging approach applied to data of increasing numbers of survey editions.

variable	year	rf08 (%)	annu (%)	pool (%)	avrg (%)
victim	2008	-0.01 #	-0.01 #	2.33	0.30
	2009	0.51	0.64	-0.26 #	0.52
	2010	-0.19	-0.09 #	-1.35	-0.20
	2011	-0.06	-0.03 #	-0.60	-0.12
degen	2008	-0.24	-0.24	0.14	-0.13 #
	2009	0.33	-0.10	0.28	0.04 #
	2010	-0.65	-0.17 #	-0.78	-0.19
	2011	0.29	-0.17	-0.03 #	-0.09
contpol	2008	0.83	0.83	0.89	0.69 #
	2009	0.32	-0.06 #	0.07	0.68
	2010	1.50	1.13	0.92 #	1.13
	2011	-0.31	0.07 #	-0.17	0.10

Table 10 Relative bias at national level of the four SAE approaches compared to the GREG estimates expressed as percentages. The least biased estimates are indicated by #.

Figures

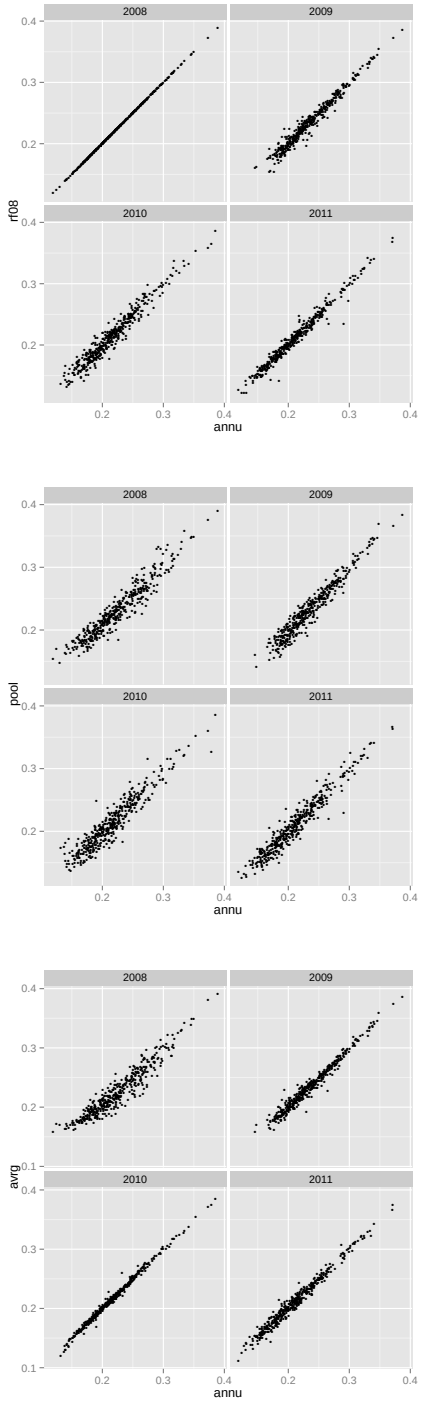


Figure 1 Comparing estimates for `victim` using different models.

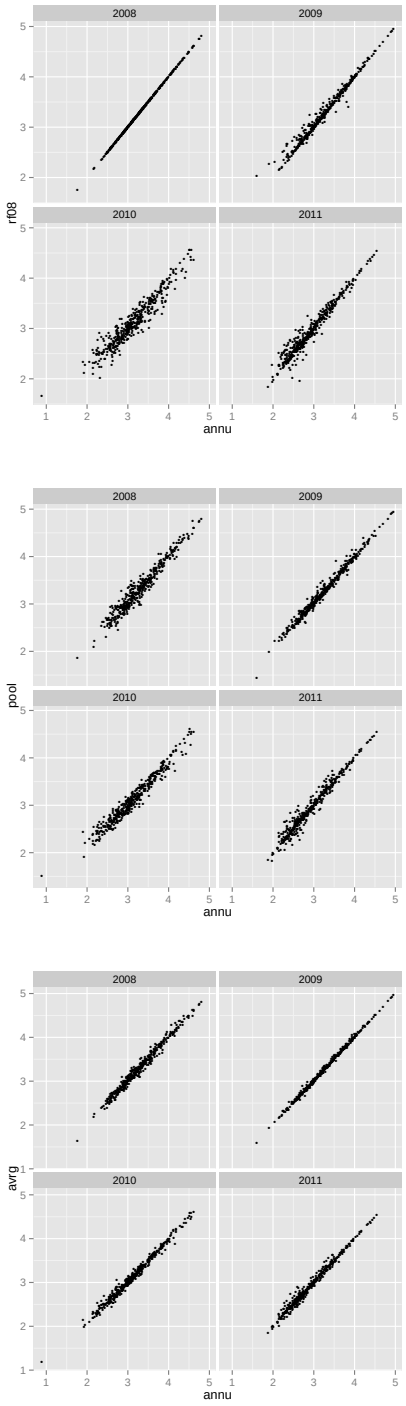


Figure 2 Comparing estimates for degen using different models.

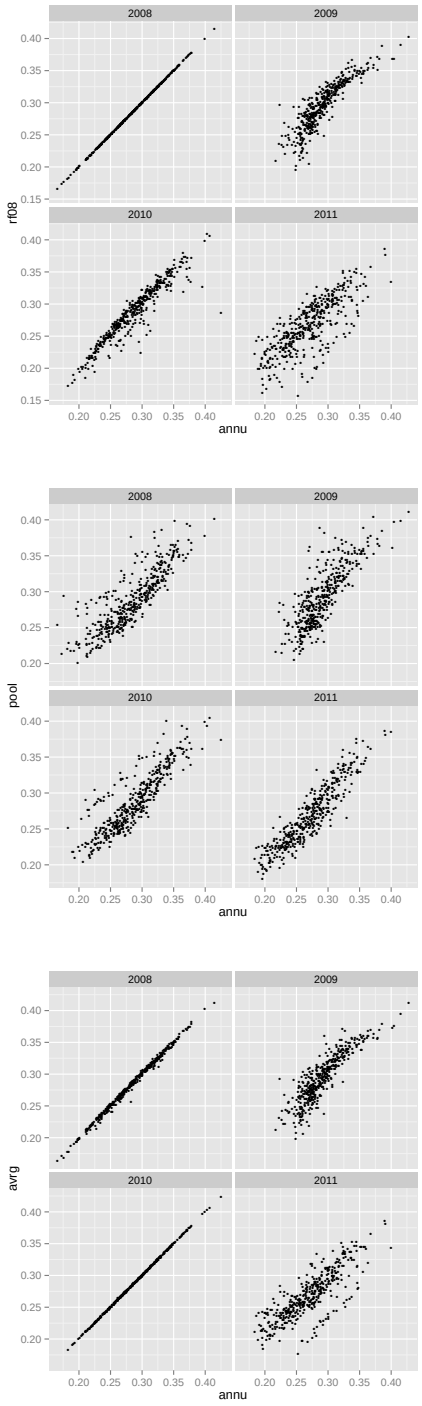


Figure 3 Comparing estimates for `contpol` using different models.

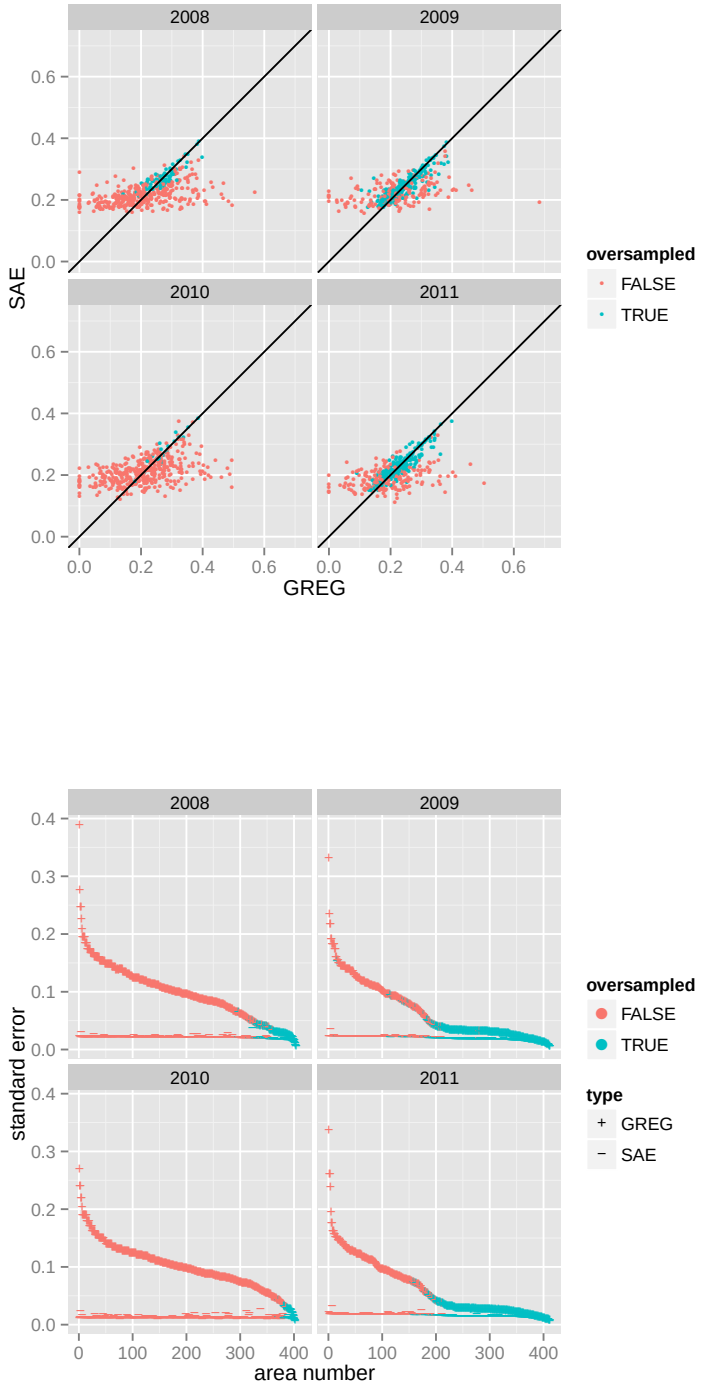


Figure 4 SAE estimates versus GREG estimates (top) and estimated standard errors (bottom) for the variable `victim`. Areas are ordered by increasing sample size.

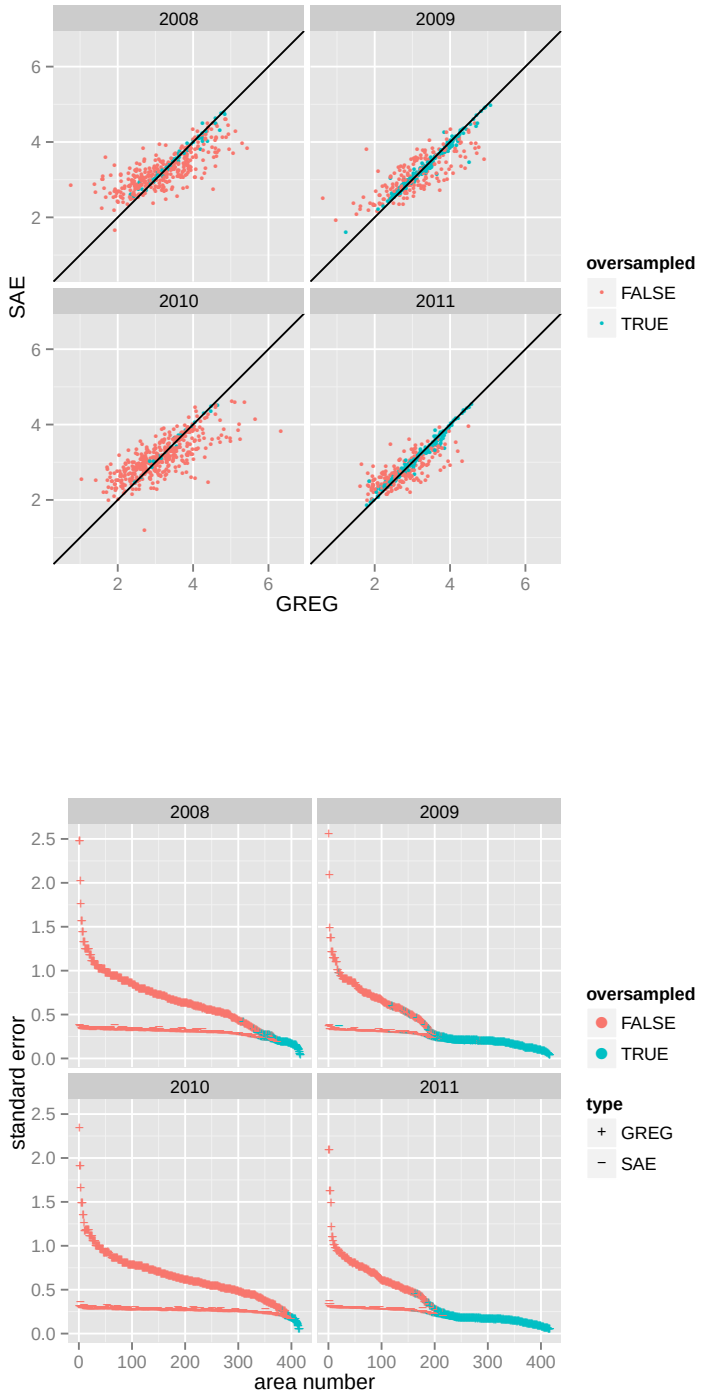


Figure 5 SAE estimates versus GREG estimates (top) and estimated standard errors (bottom) for the variable degen. Areas are ordered by increasing sample size.

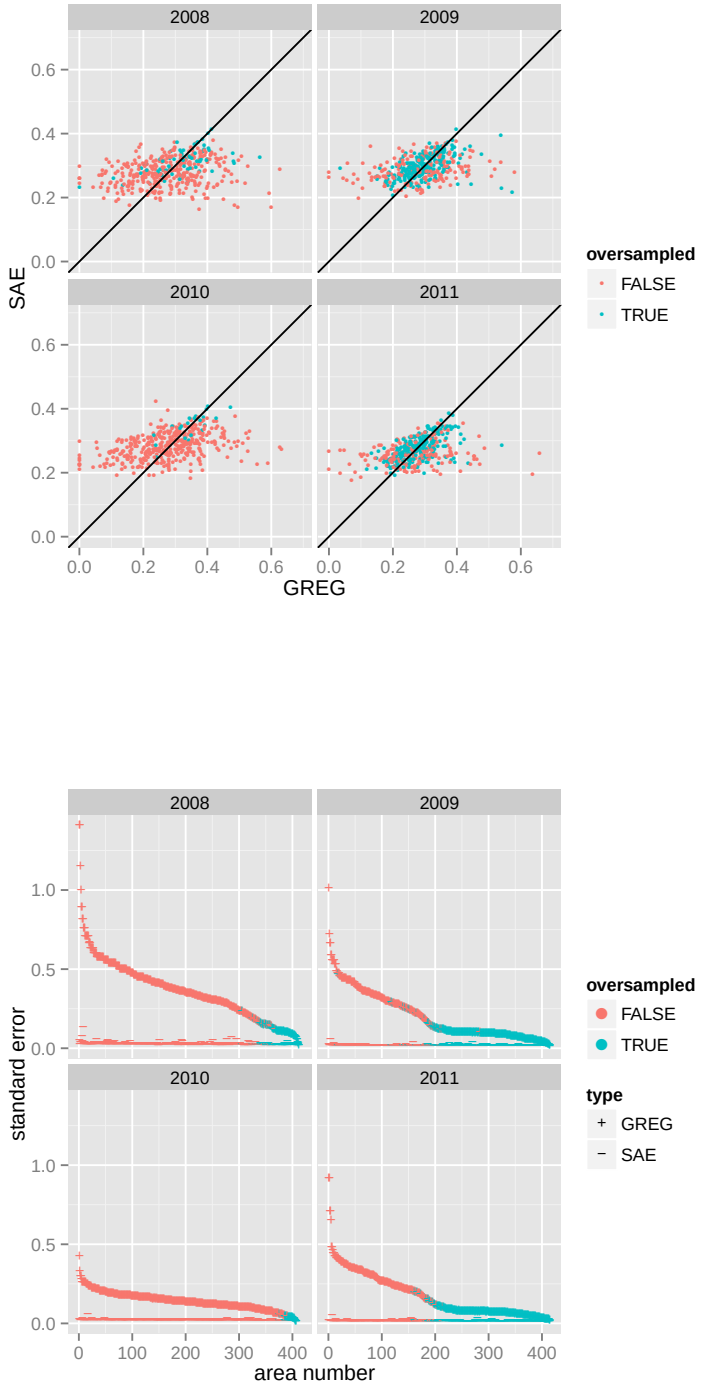


Figure 6 SAE estimates versus GREG estimates (top) and estimated standard errors (bottom) for the variable `contpol`. Areas are ordered by increasing sample size.

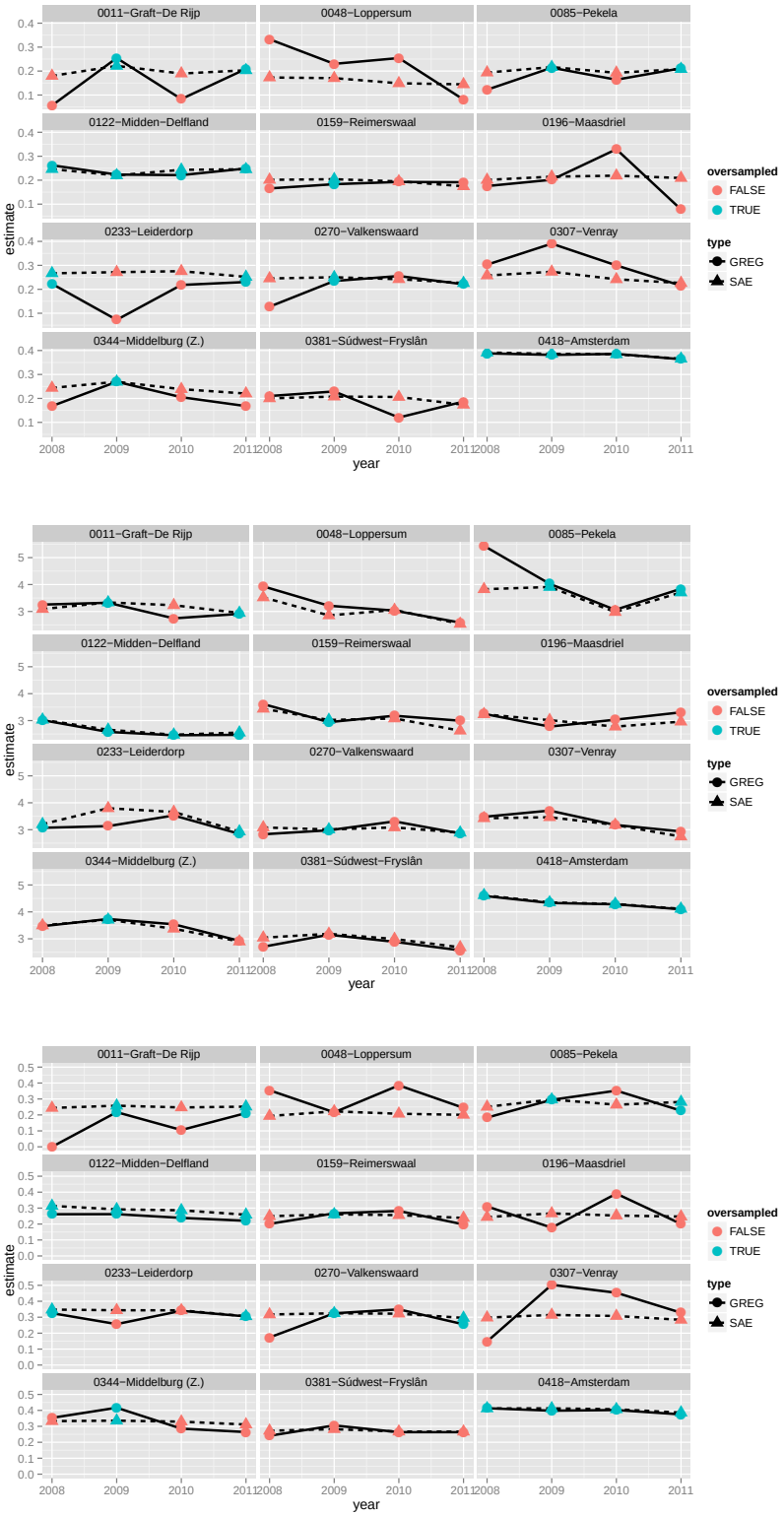


Figure 7 Time series of GREG and SAE estimates for nine municipalities for `victim` (top), `degen` (middle) and `contpol` (bottom).

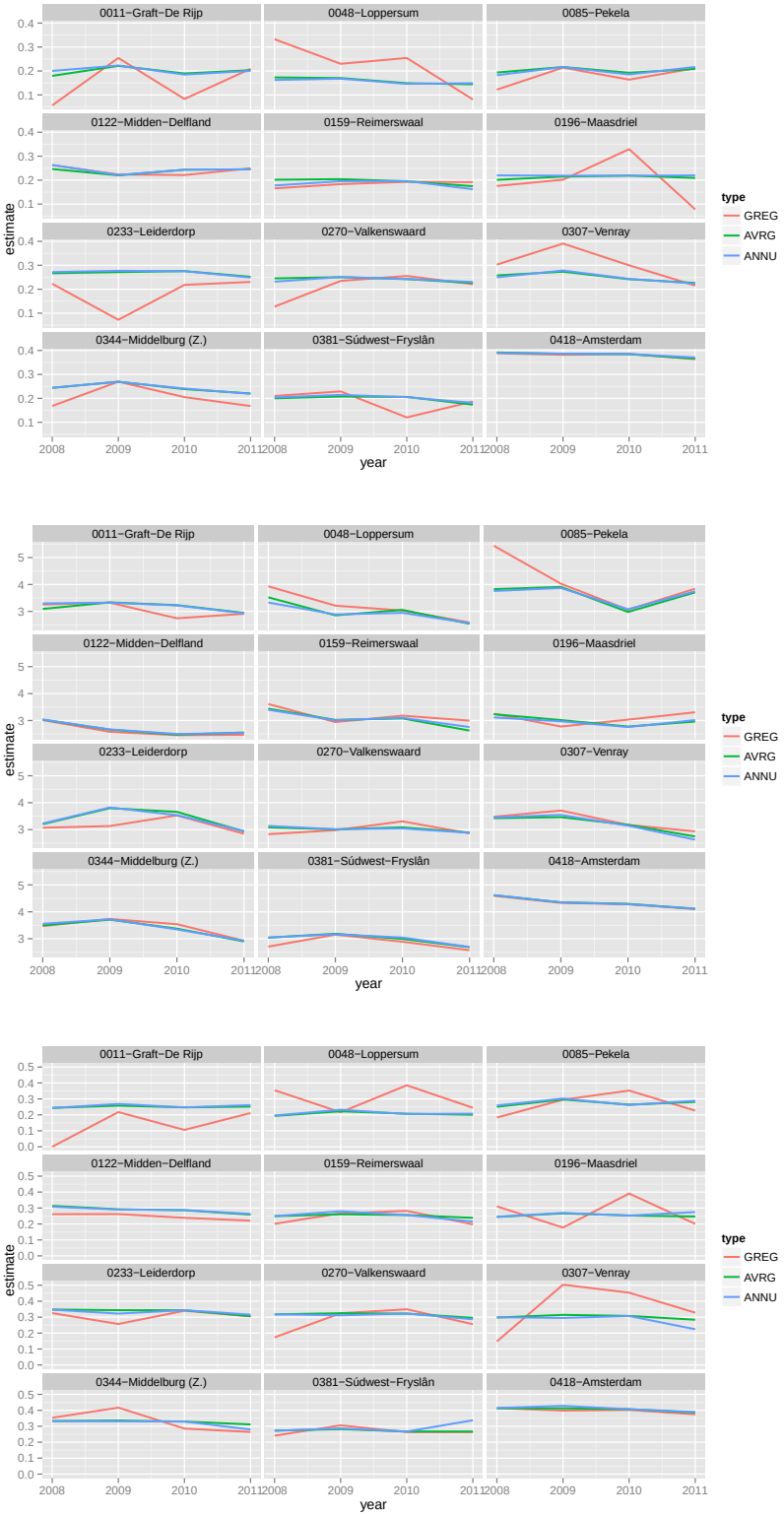


Figure 8 Time series of GREG and SAE estimates obtained through the *annu* and *avrg* approaches for nine municipalities for `victim` (top), `degen` (middle) and `contpol` (bottom).

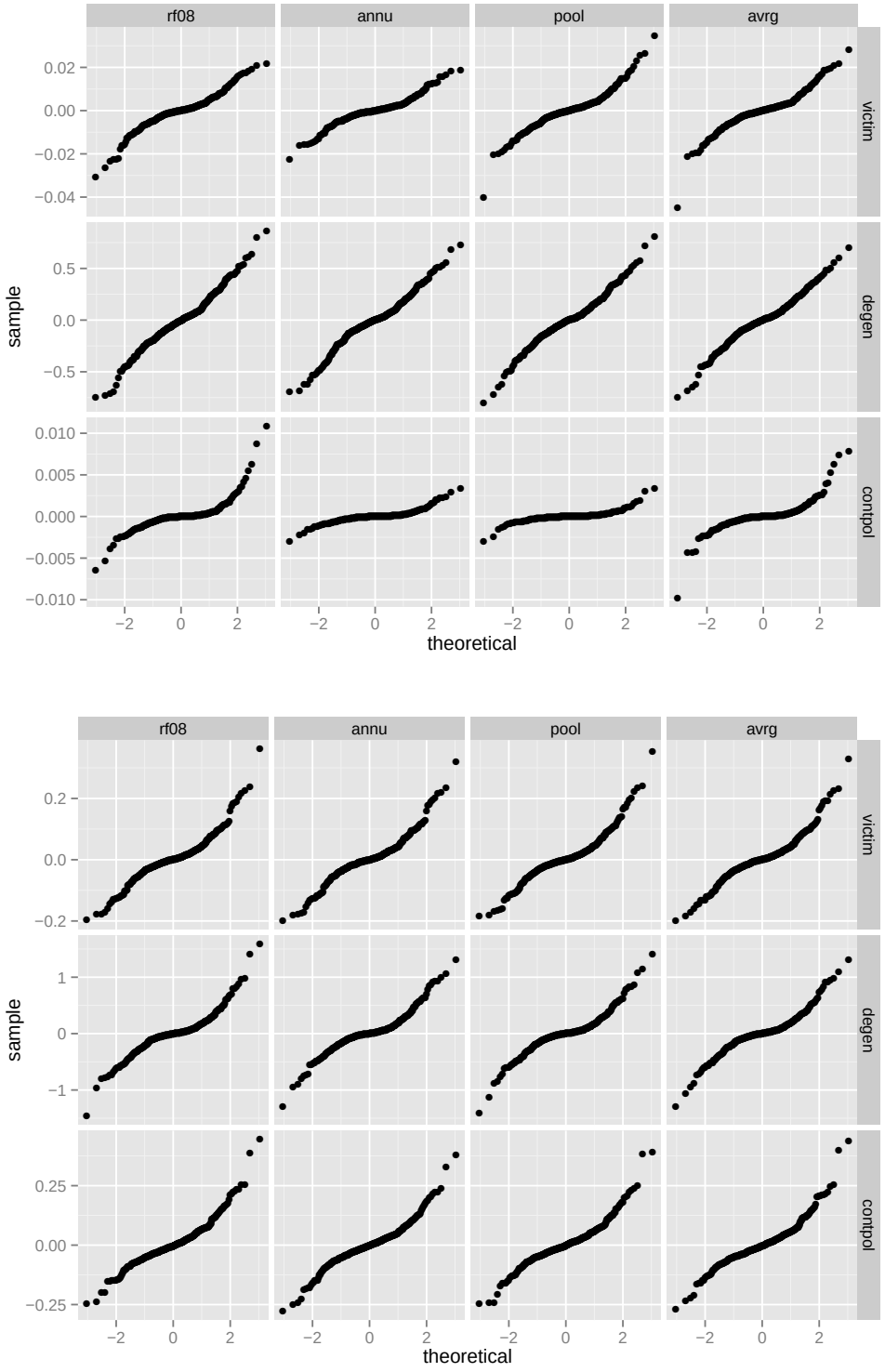


Figure 9 QQ plots for the 2011 edition of the ISM of the random effects (top) and the residues (bottom) for each of the four approaches for the three survey variables.