

Small area estimation of violent crime victim rates in the Netherlands

Bart Buelens¹, Thijs Benschop²

¹Statistics Netherlands, e-mail: b.buelens@cbs.nl

²University of Maastricht, e-mail: thijsbenschop@gmx.net

Abstract

The National Safety Monitor (NSM), a sample survey held annually in the Netherlands, provides statistics on crime, public safety and satisfaction with the police. Figures are published at the national level as well as at the regional level of Police Zones (PZ), of which there are 25 in the country. In the regular production the generalized regression (GREG) estimator is used. With approximately 750 respondents per PZ, the variance of the regional estimates is too high to enable comparisons between PZs. This paper presents the results of research into the applicability of model-based small area estimation (SAE) techniques to the NSM, with the aim of reducing the variance of the estimates at the regional level. The target variable under consideration is the violent crime victim rate. Auxiliary data is obtained from registers including the police register of reported offences. Model selection criteria are utilized to determine the optimal model from a number of potential models. The selected model is used to produce regional estimates. In most PZs the 95% confidence intervals of the model-based estimates are up to 40% smaller than the GREG confidence intervals, which is a significant gain in precision.

Keywords: small area estimation, crime rates, official statistics

1. Introduction

Small area estimation (SAE) has been a topic of research at Statistics Netherlands for several years. The main application has been the Labour Force Survey (LFS), in particular small area estimation of unemployment rates at low regional levels such as municipalities, and for short time periods such as months (Boonstra et al., 2008). Recently these research results have been applied for the first time to the National Safety Monitor (NSM). The NSM is a survey held annually in the Netherlands. The NSM provides statistics about issues relating to criminal offences, public safety and satisfaction with the police. The target population consists of all individuals aged 15 and over living in the Netherlands. The NSM is a mixed-mode survey: persons with a known fixed or mobile phone number are approached by phone, while the remaining group is approached face-to-face. The design of the NSM is aimed at obtaining 750 responses in each of the 25 Police Zones (PZs). Within each PZ all municipalities are sampled with a sample size proportional to the population size. Consequently, people within the same PZ have equal inclusion probabilities. Publication of the resulting

statistics is at the national level as well as at the regional level of PZs. A generalized regression (GREG) estimator (Särndal et al., 1992) is routinely used in production.

In this paper it is investigated whether the estimates at the level of the PZs can be made more precise by using model-based small area estimation techniques. Section 2 describes the NSM data and the estimation methodology. Results are presented in section 3. Some final conclusions and future work are discussed in section 4.

2. Data and methodology

2.1. NSM data

The target variable under consideration is the violent crime victim rate, which is defined as the percentage of the population aged 15 and over who have been a victim of sexual offences, intentional threat or physical assault at least once in the preceding period of 12 months. This is one of several important variables resulting from the NSM. Data from the 2007 survey is used. The sample consists of approximately 750 respondents in each of the 25 PZs (Table 1). Publication tables with national and regional statistics are obtained using a GREG estimator with a weighting scheme including age, gender, ethnicity, marital status, income, household size, degree of urbanization, and province. Estimates and confidence margins for the violent crime victim rate are listed in Table 1. The confidence intervals are too wide to differentiate accurately between PZs, to compare PZs to the national average, or to detect change through time. The GREG estimates are referred to as direct estimates since an estimate for a PZ only use data from that PZ.

2.2. Model-based SAE

The use of model-based methods allows for the borrowing of strength between PZs. In this paper, the classic Fay-Herriot area level model is used (Fay and Herriot, 1979):

$$\hat{\theta}_i = \beta' Z_i + v_i + \varepsilon_i \quad (1)$$

with $\hat{\theta}_i$ direct estimates of the target variable θ_i for areas $i = 1, \dots, m$, Z_i a vector of known covariates, β the regression coefficients or fixed effects, $v_i \sim N(0, \sigma_v^2)$ the random effects with variance σ_v^2 , and $\varepsilon_i \sim N(0, \psi_i)$ the sampling errors with design variance ψ_i . Estimation of this model proceeds using the method of Empirical Best Linear Unbiased Prediction or EBLUP (Rao, 2003). EBLUP estimates based on model (1) are given by:

$$\hat{\theta}_i^{FH} = \hat{\beta}' Z_i + \hat{v}_i = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \hat{\beta}' Z_i, \quad (2)$$

$$\hat{\beta} = \left(\sum_i \hat{\gamma}_i Z_i Z_i' \right)^{-1} \sum_i \hat{\gamma}_i Z_i \hat{\theta}_i, \quad (3)$$

Table 1. NSM of 2007, with for each PZ the population and sample size, and GREG-estimate of violent crime victim rate and associated margin (1/2 of the 95% confidence interval). Registered crime rate is obtained from Police Registers.

Police Zone	popu- lation size	sam- ple size	victim rate %	95% conf. margin	registered crime rate %
Amsterdam-Amstelland	762153	759	5,71	1,53	1,50
Brabant-Noord	500838	753	4,56	1,50	0,68
Brabant-Zuid-Oost	590472	786	5,61	1,52	0,84
Drenthe	388864	761	4,52	1,41	0,72
Flevoland	284469	768	5,29	1,59	0,87
Friesland	515094	786	3,98	1,38	0,39
Gelderland-Midden	513770	779	5,74	1,59	0,75
Gelderland-Zuid	421640	751	5,70	1,72	0,72
Gooi-en Vechtstreek	195706	762	4,08	1,34	0,65
Groningen	473864	779	6,71	1,78	0,82
Haaglanden	809888	756	7,32	1,79	1,03
Hollands-Midden	604132	773	5,65	1,60	0,71
Ijsselland	391558	761	4,88	1,43	0,72
Kennemerland	406776	762	5,47	1,60	0,70
Limburg-Noord	414608	765	5,46	1,54	0,73
Limburg-Zuid	518263	755	5,70	1,56	0,82
Midden- en West-Brabant	860334	786	5,02	1,50	0,87
Noord- en Oost-Gelderland	644273	762	2,94	1,19	0,65
Noord-Holland-Noord	508953	755	4,69	1,52	0,77
Rotterdam-Rijnmond	1008924	765	5,75	1,56	0,82
Twente	495155	779	6,41	1,71	0,56
Utrecht	942889	750	5,23	1,61	0,71
Zaanstreek-Waterland	252714	758	5,02	1,55	0,72
Zeeland	307032	753	3,66	1,40	0,65
Zuid-Holland-Zuid	380351	764	5,11	1,54	0,64
Total	13192720	19128	5,32	0,34	0,82

with $\hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\psi_i + \hat{\sigma}_v^2}$ the estimated ratios of model variance to total variance. An MSE estimator for the EBLUP is given in Rao (2003). Various estimation methods for σ_v^2 are given in Rao (2003) and discussed in Datta et al. (2005). Both publications recognize that the recommended method, the Fay-Herriot moments estimator, can become unstable in situations with few areas, as is the case with the NSM. Furthermore, ML and REML estimates may be too low in such situations. Earlier research at Statistics Netherlands suggests that estimating σ_v^2 using a Bayesian approach yields better results (Buelens et al., 2009). In this case the estimate $\hat{\sigma}_v^2$ is taken to be the mean of the posterior distribution $p(\sigma_v^2 | \hat{\theta})$, which is obtained from the marginal likelihood $p(\hat{\theta} | \sigma_v^2)$, and a non-informative prior $p(\beta, \sigma_v^2)$ through Bayes' rule: $p(\sigma_v^2 | \hat{\theta}) \sim p(\sigma_v^2)p(\hat{\theta} | \sigma_v^2)$. In this paper the Bayesian estimates of σ_v^2 are used.

Model (1) is appealing in a survey setting since it allows to take the design into account through the input estimates $\hat{\theta}_i$ and ψ_i . Furthermore, the estimator of the form (2) can be regarded as a weighted combination of direct estimates and model predictions. Asymptotically, $\hat{\theta}_i^{FH}$ goes to $\hat{\theta}_i$ as $\hat{\gamma}_i$ goes to 1.

Finally, the model-based estimates are benchmarked to the GREG estimate at the national level, by making adjustments proportional to the estimated variances.

2.3. Covariates and models

Data from registers are used as covariates. The main source of covariates is the police register of reported offences. Four types of crimes and offences are distinguished in the present research: violent crimes, property crimes, incidences of vandalism, and traffic offences. Additional covariates are sourced from other registers. In particular the address density is used, the proportion of the population aged 30 or over, and the proportion of non-westerners. These seven covariates are used to define 17 different models. All models are of the form (1), with the covariates Z_i for each model as listed in Table 2. Of the four registered crimes and offences, reported violent crimes and incidences of vandalism correlate respectively best and second best with the target variable. Therefore they are included in the proposed models more often than the other variables from the police register. Since there are only 25 PZs, models cannot be very large. The extreme situation with all seven covariates included would lead to a model with too many degrees of freedom.

It must be noted that there is a fundamental difference between the proportion of the population that has been a victim of violent crimes and the proportion that has registered such an offence with the police. Many crimes remain unnoticed because victims do not always register crimes, for a variety of reasons. For reference, the reported violent crime rate is included in Table 1 (see last column).

2.4. Model selection

Model selection is the process of selecting the optimal model from a set of available models. The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are used. Both combine a goodness-of-fit measure, the log-likelihood, with a penalty term for the complexity of the model. For mixed models such as model (1), the effective number of degrees of freedom p is commonly used as a measure of model complexity. It is defined as the trace of the hat matrix (Hastie et al., 2003). AIC and BIC are then given by

$$\text{AIC} = -2 L + 2 p, \quad \text{BIC} = -2 L + \log(m) p, \quad (4)$$

with L the log-likelihood and m the number of areas (PZs) as before.

Table 2. Inclusion of covariates for 17 models.

Model	Police-registered offences				Demographic register data		
	Violent crime	Property crime	Vandalism	Traffic offences	Address density	age 30+	non-westerners
1	×	×	×	×			
2	×						
3	×				×		
4	×					×	
5	×						×
6	×				×	×	
7	×				×		×
8	×					×	×
9	×				×	×	×
10	×		×				
11	×		×		×		
12	×		×			×	
13	×		×				×
14	×		×		×	×	
15	×		×		×		×
16	×		×			×	×
17	×		×		×	×	×

In addition, a measure of predictive power is used, leave-one-out cross-validation (CV). It is based on the repeated fitting of models using all-but-one observations, and using these models for prediction. The CV is the sum of the differences between the predictions and the data, using some loss function (Hastie et al., 2003). CV for model (1), with a quadratic loss function, is given by (Boonstra et al., 2009):

$$CV = \left(\sum_{i=1}^m w_i \right)^{-1} \sum_{i=1}^m w_i \left(\hat{\theta}_i - \hat{\theta}_i^{FH(-i)} \right)^2, \quad (5)$$

with $\hat{\theta}_i^{FH(-i)}$ the estimate of θ_i based on the model fit using all observations except the i th. The weights w_i are chosen as $w_i = \hat{\gamma}_i$.

3. Results

The models, all of form (1), with covariates listed in Table 2, are fitted and used to estimate regional violent crime victim rates. Results are listed in Table 3, with associated model selection measures AIC, BIC, CV and the estimated model variance $\hat{\sigma}_v^2$.

It is seen from this table that model 4 has the lowest AIC, BIC and CV values, as well as the lowest estimated model variance. Based on these results model 4 is chosen as the optimal model. This model includes only two covariates: registered violent crimes

Table 3. Model selection measures and estimated model variance for the 17 models.

Model	AIC	BIC	CV	$\hat{\sigma}_v^2$
1	92,82	110,46	1,67	0,45
2	86,36	100,14	1,11	0,34
3	87,54	102,36	1,35	0,36
4	79,57	92,25	0,95	0,25
5	86,85	101,43	1,15	0,35
6	81,58	95,55	1,15	0,27
7	89,34	105,31	1,40	0,38
8	82,06	96,13	1,00	0,28
9	83,79	99,09	1,30	0,30
10	87,87	102,84	1,37	0,37
11	89,45	105,54	1,62	0,39
12	80,44	94,21	1,19	0,26
13	88,24	103,97	1,43	0,37
14	82,79	97,90	1,40	0,29
15	90,71	107,77	1,61	0,41
16	82,93	98,08	1,26	0,29
17	85,28	101,74	1,51	0,32

Table 4. Estimates and 95% confidence margins of violent crime victim rates, using Model 4.

Police Zone	$\hat{\theta}_i$	marg.	$\hat{\theta}_i^{FH}$	marg.	marg. diff. %
Amsterdam-Amstelland	5,71	1,53	6,54	1,33	-13,32
Brabant-Noord	4,56	1,50	4,82	0,87	-42,37
Brabant-Zuid-Oost	5,61	1,52	5,39	0,87	-42,60
Drenthe	4,52	1,41	4,54	0,90	-35,89
Flevoland	5,29	1,59	5,87	1,00	-37,22
Friesland	3,98	1,38	4,48	0,97	-29,52
Gelderland-Midden	5,74	1,59	5,42	0,88	-44,57
Gelderland-Zuid	5,70	1,72	5,53	0,93	-45,61
Gooi-en Vechtstreek	4,08	1,34	4,15	0,94	-29,99
Groningen	6,71	1,78	6,13	1,01	-43,36
Haaglanden	7,32	1,79	6,33	1,00	-44,25
Hollands-Midden	5,65	1,60	5,43	0,90	-43,90
IJsselland	4,88	1,43	5,35	0,90	-37,26
Kennemerland	5,47	1,60	4,91	0,91	-43,35
Limburg-Noord	5,46	1,54	4,82	0,94	-39,28
Limburg-Zuid	5,70	1,56	5,04	0,94	-39,62
Midden- en West-Brabant	5,02	1,50	5,32	0,87	-41,84
Noord- en Oost-Gelderland	2,94	1,19	3,97	0,81	-31,89
Noord-Holland-Noord	4,69	1,52	4,96	0,87	-43,00
Rotterdam-Rijnmond	5,75	1,56	5,74	0,90	-42,03
Twente	6,41	1,71	5,50	0,99	-42,17
Utrecht	5,23	1,61	5,57	0,95	-40,81
Zaanstreek-Waterland	5,02	1,55	4,81	0,90	-42,26
Zeeland	3,66	1,40	4,26	0,88	-37,36
Zuid-Holland-Zuid	5,11	1,54	5,06	0,88	-42,71

and proportion of population aged 30 or over. Clearly these two covariates have most predictive power. Models with more covariates are penalized by AIC and BIC for their complexity, and have less predictive power according to CV, most likely due to overfitting. This model is subsequently used to produce estimates of violent crime victim rates for the 25 PZs. Results are given in Table 4. Direct and model based estimates are listed with corresponding 95% confidence margins. The last column shows the percentage difference between the margins. The model based estimates are always within the 95% confidence intervals of the direct estimates. The 95% confidence intervals of the model-based estimates are around 40% smaller than those of the direct estimates in many PZs. The reduction of the confidence interval is smallest in PZ Amsterdam-Amstelland, where it is around 13%.

4. Conclusions

Results of this initial experiment were positively received by domain experts. However, some issues were raised. A concern is that the target population of the survey is not equal to the population reporting criminal offences with the Police. For example Amsterdam has a large non-resident population of business people and tourists, who can all report crimes but are not in the NSM of PZ Amsterdam. A similar issue applies to typical holiday destinations such as Friesland and Zeeland, where there is a large non-resident population in the summer season. It can be expected that in these areas the predictive power of registered offences is less than in other PZs.

The promising findings discussed in this paper will be extended. Planned future work includes the use of unit-level models, and attention to temporal aspects such as detection of change from year to year. Survey variables other than violent crime victim rate will be included in the study.

References

- Boonstra, H. J., Van den Brakel J. A., Buelens B., Krieg S., Smeets, M. (2008) Towards small area estimation at Statistics Netherlands, *Metron*, LXVI, 1, 21-49
- Boonstra, H. J., Buelens B., Smeets M. (2009) Model selection for small area estimation, Technical report DMH-2009-01-22-HBTA, Statistics Netherlands, Heerlen
- Buelens, B., Boonstra H. J., Smeets M. (2009) Estimation of the variance components of area level models for small area estimation, Discussion paper DMH-2009-01-06-BBUS, Statistics Netherlands, Heerlen
- Datta, G. S., Rao J. N. K., Smith D. D. (2005) On measuring the variability of small area estimators under a basic area level model, *Biometrika*, 92, 1, 183-196
- Fay, R. E., Herriot R. A. (1979) Estimates of income for small places: an application of James-Stein procedures to Census data, *Journal of the American Statistical Association*, 74, 269-277
- Hastie, T., Tibshirani R., Friedman J. H. (2003) *The elements of statistical learning*, Springer, New York
- Särndal, C. E., Swensson B., Wretman J. (1992) *Model assisted survey sampling*, Springer, New York
- Rao, J. N. K. (2003) *Small area estimation*, John Wiley, New York