Model-Based Inference for Small Area Estimation with Sampling Weights

Y. Vandendijck[a,*], C. Faes[a], R. S. Kirby[b], A. Lawson[c], N. Hens[a,d]

[a] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium*
[b] *Department of Community and Family Health, College of Public Health, University of South Florida, Tampa, FL*
[c] *Department of Public Health, University of South Carolina, Charleston, SC*
[d] *Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium*

Abstract

Obtaining reliable estimates about health outcomes for areas or domains where only few to no samples are available is the goal of small area estimation (SAE). Often, we rely on health surveys to obtain information about health outcomes. Such surveys are often characterised by a complex design, stratification, and unequal sampling weights as common features. Hierarchical Bayesian models are well recognised in SAE as a spatial smoothing method, but often ignore the sampling weights that reflect the complex sampling design. In this paper, we focus on data obtained from a health survey where the sampling weights of the sampled individuals are the only information available about the design. We develop a predictive model-based approach to estimate the prevalence of a binary outcome for both the sampled and non-sampled individuals, using hierarchical Bayesian models that take into account the sampling weights. A simulation study is carried out to compare the performance of our proposed method with other established methods. The results indicate that our proposed method achieves great reductions in mean squared error when compared with standard approaches. It performs equally well or better when compared with more elaborate methods when there is a relationship between the responses and the sampling weights. The proposed method is applied to estimate asthma prevalence across districts.

*Corresponding author
Email address: yannick.vandendijck@uhasselt.be (Y. Vandendijck)

Keywords: Integrated Nested Laplace Approximations, Model-Based Inference, Small Area Estimation, Spatial Smoothing, Survey Weighting

1. Introduction

In public health we are often interested in the question whether there are disparities in illness, behavioural risk factors or health conditions across areas. An increasing amount of information on individuals is collected in this respect. Bayesian methods in disease mapping based on census or population registry data are well developed and used in a fairly standard manner (see e.g., Elliott et al., 2001; Waller and Gotway, 2004; Lawson, 2013 for a review of the methods). Such population registry or census data obtains information pertaining to each member of the population of an area. Historically, focus was on the construction of cancer atlases and on mapping rare diseases based on registry data (see e.g., Kemp et al., 1985; Mason, 1995).

Since it is nearly always impossible to measure the health outcome of interest in every individual in the population, a survey is used to record information from a random sample of individuals from the population (Cochran, 1977). Such surveys are often characterized by a complex design, with stratification, clustering and unequal sampling weights as common features. Policy makers are often interested in a specific characteristic, such as the total number of diseased cases or the prevalence, per area. In small area estimation (SAE) one investigates how to obtain these area specific characteristics from survey data covering more than only the area of interest by using spatial smoothing methods.

In SAE, one needs to choose whether to base inference on *design-based*, *model-based* or *design-based model-assisted* approaches. In design-based inference the values of the health outcomes are assumed fixed, and inference is based on the randomization distribution of the sample inclusion indicators. Often a model is used in the construction of a design-based estimator (known as design-based model-assisted approaches). A popular design-based estimator is the Horvitz-Thompson (HT) estimator (1952) and its extensions that weigh sampled individuals with the associated sampling weight. These estimators play a dominant role in sample surveys, however, they often fail in SAE because the sample size per area could be very small or even zero inflating the mean squared error tremendously. This makes design-based estimators unreliable or not feasible to use (Rao, 2011). Additionally, because of the spatial nature of the problem, understanding the geographical

distribution of the health outcome is important. Model-based approaches that perform spatial smoothing, both those based on empirical and hierarchical Bayesian methodology, have shown to be more relevant in the handling of spatially correlated health survey data. In model-based approaches one conditions on the selected sample and the inference is based on the underlying model of the health outcome. Examples include Fay and Herriot (1979) which proposed a linear empirical Bayes model to estimate the income for small areas, while Datta and Ghosh (1991) considered a hierarchical Bayesian formulation instead. A number of extensions have been made, see Rao (2003) and Jiang and Lahiri (2006) for an overview. For binary data, MacGibbon and Tomberlin (1989) developed an empirical Bayes model using a logistic regression model with fixed and random effects. Stroud (1994), Ghosh et al. (1998) and Farrell (2000) described hierarchical Bayesian approaches to estimate small area proportions.

While model-based SAE is conceptually appealing, complex survey designs with the accompanying survey weights cause a difficulty in their practical implementation. Only relatively few approaches acknowledge the survey sampling mechanism and account for it in the model. Kott (1989) and Prasad and Rao (1999) described a design-consistent model-based estimator. Kott (1989) proposed an estimator which is a weighted combination of the HT estimator and the sample means of the different areas. Prasad and Rao (1999) proposed a pseudo-empirical best linear unbiased prediction estimator for the small area mean based on area level data. You and Rao (2002, 2003) used unit level data instead. Malec et al. (1997) described a hierarchical Bayesian model for binary survey data. They examined the use of sampling weights as a linear covariate in the model, after the inclusion of several post-stratification variables. Chen et al. (2014) proposed the use of a weight-adjusted Bayesian estimator that takes into account the effective sample size. Mercer et al. (2014) described a simulation study in which several methods for spatial smoothing in SAE, taking into account the sampling weights, are compared.

In this article, we describe a spatial predictive model-based approach to SAE for a binary health outcome in a complex survey with given sampling weights. We assume that the sampling weights on the sampled individuals are the only information available about the survey design. The goal is to estimate the prevalence of the health outcome for all small areas in the spatial domain. A hierarchical Bayesian model is used in which the health outcomes are regressed on the sampling weights. A non-parametric regression on the

weights is used to minimise possible bias of the regression function. Additionally, both unstructured and structured spatial random effects are introduced to model the geographical distribution of the health outcomes. The population distribution of the sampling weights is unknown as well, hence we must model the weights themselves to be able to perform predictions. Our proposed method extends ideas described in Si et al. (2015) that are useful for surveys outside the SAE context. We use integrated nested Laplace approximations in R for model estimation (Rue et al., 2009). The methods described in this article add a hierarchical Bayesian model-based prediction approach for data with associated sampling weights to the SAE literature.

The structure of the paper is as follows. In Section 2 we introduce notations and describe the traditional design-based approach to perform SAE from a health survey. Several model-based approaches summarized in Mercer et al. (2014) that are used here for comparison purposes in the simulation study are also described in Section 2. We describe our proposed model-based approach in Section 3, and provide some details on the implementation of the models in standard software. A simulation study comparing our methods to other design- and model-based methods is provided in Section 4. In Section 5, we analyse the 2001 Belgian Health Interview Survey to estimate asthma prevalence across districts. We conclude the paper with a discussion in Section 6.

2. Notation and Conventional Method of Analysis

2.1. Notation

Let Y_{ik} be a binary health outcome for individual i in small area k ($i = 1, \ldots, N_k$ and $k = 1, \ldots, K$) with N_k the population size in area k. We assume that N_k is known for each area. A sample of size n_k is drawn from each area k, where some of the n_k could be zero. Denote the sampled values by y_{ik}. Let $N = \sum_{k=1}^{K} N_k$ and $n = \sum_{k=1}^{K} n_k$ represent the total population and sample size, respectively. We shall focus on estimating the true prevalence, P_k, in each area k, namely

$$P_k = \frac{1}{N_k} \sum_{i=1}^{N_k} Y_{ik}. \tag{1}$$

Let R_{ik} denote the binary variable indicating whether the i^{th} individual in area k is sampled ($R_{ik} = 1$) or not ($R_{ik} = 0$). We use s_k to indicate the set of sampled individuals in area k and s'_k for those that are not sampled.

Table 1: Structure of datasets used in this article.

Response	Area	Sample Weight
y_{11}	1	w_{11}
y_{21}	1	w_{21}
$\vdots$	$\vdots$	$\vdots$
y_{12}	2	w_{12}
$\vdots$	$\vdots$	$\vdots$

To reflect the sampling design, weights w_{ik} are attached to each respondent's outcome. The weights are proportional to the inverse probability of inclusion in the sample for unit i in area k. These weights can reflect both or a combination of the complex survey design and post-stratification adjustments. In this paper, we assume that the sampling weights on the sampled individuals are the only information available on the design of the survey. This assumption seems limited (e.g., one cannot adjust for cluster sampling) but it is standard in publicly available datasets that only sampling weights are available with few to no information on the survey design. We further assume that all sampled individuals respond to the survey. A typical dataset will have the structure as presented in Table 1. Throughout this article, we use the normalized weights, denoted by $\tilde{w}_{ik}$, defined by

$$\tilde{w}_{ik} = n_k \frac{w_{ik}}{\sum_{i \in s_k} w_{ik}}. \tag{2}$$

The weights are called normalized because they sum up to the sample size n_k in area k.

2.2. Horvitz-Thompson Estimator

Design-based methods evaluate properties of estimators under the randomization distribution and assume that the measurements are fixed values. Inference is based on all possible samples that could be selected from the target population of interest under the considered sampling design. A common design-based estimator in SAE is the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) given by

$$\hat{p}_k^{HT} = \frac{\sum_{i=1}^{N_k} R_{ik} \tilde{w}_{ik} y_{ik}}{\sum_{i=1}^{N_k} R_{ik} \tilde{w}_{ik}} = \frac{1}{n_k} \sum_{i \in s_k} \tilde{w}_{ik} y_{ik}. \tag{3}$$

The variance of $\hat{p}_k^{HT}$ has the form

$$\widehat{\text{var}}(\hat{p}_k^{HT}) = \frac{1}{n_k}\left(1 - \frac{n_k}{N_k}\right)\frac{1}{n_k - 1}\sum_{i \in s_k} \tilde{w}_{ik}^2 \left(y_{ik} - \hat{p}_k^{HT}\right)^2. \tag{4}$$

The HT estimator is a design-unbiased estimator of P_k. It is a so-called *direct* estimator because it uses only the responses from the area of interest (Rao, 2003). Most surveys are not designed to yield appropriate direct estimates for all areas as the sample size in some areas can be too small to produce reliable or stable estimates. Another disadvantage is that no estimate can be obtained in those areas that are not included in the sample. In the next section, we present *indirect* estimates that borrow strength across the different areas by using the responses from all sampled areas.

Other design-based methods use a model for the construction of the estimators. The synthetic estimator (Gonzalez, 1973), for example, uses a linear model on several covariates fit by ordinary least squares to predict the mean for a particular area. Pfeffermann (2013) gives an overview of commonly used methods and new developments in the field of design-based small area estimation.

2.3. Model-based methods

We now focus on several model-based methods described in Mercer et al. (2014). The simplest approach, called naive binomial (NB), is to ignore the design and use the model

$$y_k|\tilde{p}_k \sim \text{Binomial}(n_k, \tilde{p}_k) \quad \text{and} \quad \text{logit}(\tilde{p}_k) = \beta_0 + u_k + v_k, \tag{5}$$

where $y_k = \sum_{i \in s_k} y_{ik}$, β_0 is an overall effect, $u_k \sim_{iid} \mathcal{N}(0, \sigma_u^2)$ are independent random effects taking into account extra heterogeneity amongst areas k, and v_k are spatially dependent random effects. It is assumed that v_k follows the commonly used intrinsic conditional autoregressive (ICAR) model (Rue and Held, 2005)

$$v_k|v_{k'} \sim \mathcal{N}\left(\frac{1}{m_k}\sum_{k' \in \text{ne}(k)} v_{k'}, \frac{\sigma_v^2}{m_k}\right), \tag{6}$$

where $\text{ne}(k)$ denotes the set of neighbours of area k and m_k is the number of neighbours. For identifiability reasons of the overall intercept β_0, the sum of the random effects v_k is constrained to zero (Eberly and Carlin, 2000). In this

article, we take the common approach to consider two areas as neighbours if they share a common boundary.

In model (5), the design weights of the survey are ignored. To account for the design, Mercer et al. (2014) proposed to model the empirical logistic transform of $\hat{p}_k^{HT}$, namely $y_k^L = \log\left[\hat{p}_k^{HT}/(1-\hat{p}_k^{HT})\right]$, using the model

$$y_k^L|\tilde{p}_k \sim \mathcal{N}\left(\text{logit}(\tilde{p}_k), \sigma_k^2\right) \quad \text{and} \quad \text{logit}(\tilde{p}_k) = \beta_0 + u_k + v_k, \tag{7}$$

where the variance σ_k^2 is set equal to $\widehat{\text{var}}(\hat{p}_k^{HT})/\left[\hat{p}_k^{HT}(1-\hat{p}_k^{HT})\right]^2$. Model (7) is further referred to as the logit-normal (LN) model.

As an alternative, Mercer et al. (2014) considered the arcsine square-root transformation as proposed by Raghunathan et al. (2007). This transformation, $y_k^A = \sin^{-1}(\sqrt{\hat{p}_k^{HT}})$, is an approximate variance stabilizing transformation for binary data. The model, called the arcsine (AS) model, is

$$y_k^A|\tilde{p}_k \sim \mathcal{N}\left(\sin^{-1}(\sqrt{\tilde{p}_k}), \sigma_k^2\right) \quad \text{and} \quad \sin^{-1}(\sqrt{\tilde{p}_k}) = \beta_0 + u_k + v_k, \tag{8}$$

where the variance σ_k^2 is equal to $(4n_k^*)^{-1}$ with n_k^* the so-called effective sample size calculated as $n_k^* = \hat{p}_k^{HT}(1-\hat{p}_k^{HT})/\widehat{\text{var}}(\hat{p}_k^{HT})$.

Some authors proposed the use of a weighted likelihood to take into account the sampling design, also referred to as pseudo-likelihood (PL) (e.g., see Skinner, 1989 and Congdon and Lloyd, 2010). Mercer et al. (2014) also considered a pseudo-likelihood approach by assuming that $y_k^P = \sum_{i \in s_k} \tilde{w}_{ik} y_{ik}$ has a binomial likelihood, namely

$$y_k^P|\tilde{p}_k \sim \text{Binomial}(n_k, \tilde{p}_k) \quad \text{and} \quad \text{logit}(\tilde{p}_k) = \beta_0 + u_k + v_k. \tag{9}$$

Finally, Mercer et al. (2014) described a recent approach proposed by Chen et al. (2014) that combines the pseudo-likelihood approach with the effective sample size. The model, called the effective sample size (ES) model, is

$$y_k^E|\tilde{p}_k \sim \text{Binomial}(n_k^*, \tilde{p}_k) \quad \text{and} \quad \text{logit}(\tilde{p}_k) = \beta_0 + u_k + v_k, \tag{10}$$

where $y_k^E = n_k^* \hat{p}_k^{HT}$. The rationale behind this model is that both numerator and denominator are adjusted for the sampling design.

For more details on the models described in this section, we refer to Mercer et al. (2014). Models (5), (7), (8), (9) and (10), in which global and local information is borrowed within the same model via independent and ICAR random effects for each region, are called convolution models (Besag et al., 1991).

3. Proposed Methods

In this section, we propose a hierarchical model for the observed outcomes y_{ik} (Section 3.1), and explain how to use this model to make predictions $\hat{y}_{ik}$ for non-sampled individuals in order to obtain an estimator of P_k (Section 3.2).

3.1. Hierarchical Model

A predictive model-based approach proposed by Royall (1970) is used to specify an estimator for P_k. The estimator is given by

$$\hat{p}_k = \frac{1}{N_k}\left(\sum_{i\in s_k} y_{ik} + \sum_{i\in s_k'} \hat{y}_{ik}\right), \tag{11}$$

where the first term sums outcome values of the sampled individuals, and the second term is a sum of the predicted values over the non-sampled individuals in area k. Royall (1970) argued that this model-based approach is more efficient than the design-based approaches, when the model to predict $\hat{y}_{ik}$ is correctly specified.

Several authors proposed the use of weight-smoothing models to predict $\hat{y}_{ik}$ by modelling the health outcome as a smooth-varying function of the survey weights (or probability of inclusion), and showed that these models give better estimates as compared to the HT estimator. For continuous data, Zheng and Little (2003, 2005) estimated the finite population total based on a non-parametric regression as a function of the inclusion probabilities in the likelihood framework. Chen et al. (2010) used a Bayesian p-spline predictive estimator to estimate the finite population proportion. Their model is a binary p-spline probit regression model with the inclusion probability as covariate.

We extend these ideas to small area estimation. The normalized sampling weights are used as a covariate in the model for the observed outcomes y_{ik}. We employ Bayesian hierarchical models consisting of three stages. At the first stage, the likelihood of the binary outcome is specified, namely

$$\begin{aligned} y_{ik}|\tilde{p}_{ik} &\sim \text{Bernoulli}(\tilde{p}_{ik}), \\ \text{logit}(\tilde{p}_{ik}) &= \eta_{ik}. \end{aligned} \tag{12}$$

At the second stage, the latent process η_{ik} is modelled as a function of the sampling weights and allows for between-area variation by using both

spatially independent and spatially dependent random effects (Besag et al., 1991). Two versions are considered:

Model 1: η_{ik} in (12) is modelled as

$$\eta_{ik} = \beta_0 + f(\tilde{w}_{ik}) + u_k + v_k. \tag{13}$$

Model 2: η_{ik} in (12) is modelled as

$$\eta_{ik} = \beta_0 + f(\pi_{ik}) + u_k + v_k, \tag{14}$$

where $\pi_{ik} = 1/\tilde{w}_{ik}$. Similar as described in Section 2.3, β_0 is an overall effect, u_k are independent random effects and v_k are spatially dependent random effects following the ICAR model.

Zheng and Little (2003, 2005) and Chen et al. (2010) used penalized splines to construct the non-parametric function $f(\cdot)$ in (13) and (14). Si et al. (2015) avoided parametric assumptions or specific functional forms for the $f(\cdot)$ function and used a Gaussian process (GP) prior. To specify $f(\cdot)$, we investigate both approaches, namely penalized splines using B-spline basis functions and a GP prior using a random walk model of order one (RW1).

For notational convenience, let $\tilde{w}_{[\delta]} = (\tilde{w}_{\delta_1}, \ldots, \tilde{w}_{\delta_L})$ denote the sorted set of unique values of all observed weights $\tilde{w}_{ik}$ in the sample, where L is the number of unique values. A RW1 model is a smoothing model constructed by assuming that the increments, $\triangle\tilde{w}_{\delta_l} = \tilde{w}_{\delta_l} - \tilde{w}_{\delta_{l-1}}$, follow a multivariate normal distribution with mean zero. The distribution of $\tilde{w}_{[\delta]}$ is thus proportional to

$$\exp\left\{-\frac{1}{2\sigma_w^2}\sum_{l=2}^{L}(\tilde{w}_{\delta_l} - \tilde{w}_{\delta_{l-1}})^2\right\}, \tag{15}$$

with variance parameter σ_w^2. For $1 < l < L$, the conditional distribution of the $(l+1)$th normalized weight $\tilde{w}_{\delta_{l+1}}$ depends only on $\tilde{w}_{\delta_l}$ and $\tilde{w}_{\delta_{l+2}}$, while the boundary weights $\tilde{w}_{\delta_1}$ and $\tilde{w}_{\delta_L}$ depend on their only neighbour point. To specify the RW1 model in (14) one simply replaces $\tilde{w}_{ik}$ by π_{ik} in (15).

In the penalized spline case, the function $f(\cdot)$ is of the form

$$f(\tilde{w}_{ik}) = \sum_{b=1}^{B}\theta_b B_b(\tilde{w}_{ik}), \tag{16}$$

where $B_1(\cdot), \ldots, B_B(\cdot)$ are the B-spline basis functions of degree d (Eilers and Marx, 1996). The smoothness of $f(\cdot)$ is achieved by imposing a penalty on

the regression coefficients $\boldsymbol{\theta}$ of the form $\lambda \boldsymbol{\theta}^T \mathbf{D}_q^T \mathbf{D}_q \boldsymbol{\theta}$, where λ is a smoothing parameter and $\mathbf{D}_q$ is the q-th order differencing matrix (Eilers and Marx, 1996). For a detailed description of B-splines, see for example Hastie et al. (2001) and Ruppert et al. (2003). We implement B-spline basis functions of degree two ($d = 2$) with a second order difference penalty ($q = 2$). We choose $B = 20$ to ensure enough flexibility and we use quantile-based knots. For fitting purposes, the penalized spline in (16) is expressed in general linear mixed model representation (Ruppert et al., 2003):

$$f(\tilde{w}_{ik}) = \tilde{w}_{ik}\beta_w + \sum_{b=1}^{B'} \alpha_b z_b(\tilde{w}_{ik}), \tag{17}$$

where β_w is an unknown coefficient, $z_b(\cdot)$ is a transformed spline basis of the B-spline basis functions and $\alpha_b \sim \mathcal{N}(0, \sigma_\alpha^2)$ (see the Supplementary Materials for more information).

In the last stage one needs to assign proper hyperprior distributions for the unknown parameters in stage 2, namely for β_0, σ_u^2, σ_v^2, σ_w^2 and σ_α^2. Vague priors are used for all parameters. A normal prior with large variance is used for β_0. Similar as Mercer et al. (2014) and Chen et al. (2014), we assign Gamma$(0.5, 0.008)$ priors on the precision parameters σ_u^{-2} and σ_v^{-2}. This gives a 95% range on the σ_u and σ_v scale of (0.056, 4.036). It is well-known that by taking both these hyperpriors to be vague, only the sum of the random effects $(u_i + v_i)$ and not their individual values are identified (Gelfand et al., 2010). A common choice for the prior on the precisions σ_w^{-2} and σ_α^{-2} is Gamma$(0.001, 0.001)$. However, as discussed in Wakefield (2009) this prior puts most of the prior mass of σ_w and σ_α to the right of the prior distribution and is therefore not recommended. To avoid this and as recommended by Wakefield (2009), we use Gamma$(1, 0.01)$ priors for both precisions σ_w^{-2} and σ_α^{-2}. These priors yield a 95% range on the σ_w and σ_α scale of (0.052, 0.628). We investigate in Sections 4 and 5 the sensitivity of our results to other prior distributions.

3.2. Prediction of Health Outcome for Non-sampled Individuals

Once estimates of β_0, u_k, v_k and $f(\tilde{w}_{ik})$ for model 1 are obtained, $\tilde{p}_{ik}$ in (12) is estimated by

$$\hat{p}_{ik} = \left\{1 + \exp(-(\hat{\beta}_0 + \hat{f}(\tilde{w}_{ik}) + \hat{u}_k + \hat{v}_k))\right\}^{-1}, \tag{18}$$

with a similar expression used for model 2 (replacing $\tilde{w}_{ik}$) with π_{ik}. However, note that for the non-sampled individuals no information on w_{ik}, or equivalently on $\tilde{w}_{ik}$, is available, and thus it is not possible to obtain estimates of $\hat{p}_{ik}$ for individuals in s'_k. In this section, we propose (i) a method to estimate weights $\tilde{w}_{ik}$ for the non-sampled individuals i in sampled area k; and (ii) a method to estimate P_k for non-sampled areas k.

In previous work of Zheng and Little (2003, 2005) and Chen et al. (2010) it was assumed that the inclusion probabilities (sampling weights) were known for all units in the population. These three papers assumed a probability-proportional-to-size sampling in which the inclusion probability is proportional to a size variable (e.g., dwelling size) measured for all population units. Alternatively, Si et al. (2015) proposed to model the sampling process by a Bayesian model to obtain weights $\tilde{w}_{ik}$ for non-sampled individuals.

In each area k, we map the unique values of the observed weights w_{ik} to form L_k strata, where L_k is the number of unique w_{ik} values in area k. Si et al. (2015) referred to these strata as *poststratification cells* since they are constructed based on the weights in the sample. It is assumed that the L_k poststratification cells observed in the sample in area k are the only possible strata in the population of this area. Si et al. (2015) argued that this assumption is generally not correct (for example, if weights are constructed by multiplying factors for different demographic variables, there may be some empty cells in the sample corresponding to unique products of factors that would appear in the population but not in the sample) but it allows one to proceed with (18) without additional knowledge of the process by which the weights were constructed.

For each k, let n_{lk} denote the sample size in poststratification cell l ($l = 1, \ldots, L_k$) in area k. It is assumed that the supplied weights w_{ik} are proportional to the inverse of the inclusion probabilities. Assuming independent sampling, the sampling process probabilities for an individual i in area k is given by a Bernoulli process, with probability

$$P(R_{ik} = 1) = c_k / w_{ik}, \tag{19}$$

where c_k is a positive normalizing constant to ensure that the expected number of observed individuals corresponds with the actual sample size n_k. Denote by N_{lk} the (unknown) population size in stratum l and area k. Since all individuals i in stratum l have the same weight, $w_{ik} \equiv w_{(l)k}$, the expected value of n_{lk} is $\mathrm{E}(n_{lk}) = c_k N_{lk} / w_{(l)k}$. Because $n_k = \sum_{l=1}^{L_k} n_{lk}$, it follows that

$c_k = n_k \frac{1}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}$, and as a result $\mathrm{E}(n_{lk}) = n_k \frac{N_{lk}/w_{(l)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}$. Therefore, it is assumed that the vector $(n_{1k}, ..., n_{L_k k})$ in area k follows a multinomial distribution conditional on n_k,

$$(n_{1k}, ..., n_{L_k k}) \sim \text{Multinom}\left(n_k; \frac{N_{1k}/w_{(1)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}, ..., \frac{N_{L_k k}/w_{(L_k)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}\right), \quad (20)$$

for each k, where the N_{lk} are unknown parameters. Because the N_{lk} are unnormalized in the above parametrization, we normalize them after fitting such that they sum to the population size in area k:

$$\tilde{N}_{lk} = \frac{N_{lk}}{\sum_{l=1}^{L_k} N_{lk}} N_k. \quad (21)$$

Knowledge of $\tilde{N}_{lk}$ can be used in (11), which can be written as

$$\hat{p}_k = \frac{1}{N_k}\left\{\sum_{l=1}^{L_k} n_{lk}\bar{y}_l + \sum_{l=1}^{L_k}\left(\tilde{N}_{lk} - n_{lk}\right)\hat{p}_{lk}\right\}, \quad (22)$$

where $\bar{y}_l = \sum_{i \in l} y_{ik}/n_{lk}$ and $\hat{p}_{lk}$ is obtained from (18) with weight $\tilde{w}_{(l)k}$. Equation (22) is the point estimate of P_k. The contributions of the L_k different cells to the estimation of P_k is clear from (22).

Inference of $\hat{p}_k$ is based on the posterior distribution of $\hat{p}_k$. The posterior is obtained by drawing B posterior samples from the posterior distributions of $\hat{p}_{lk}$ and $\tilde{N}_{lk}$ for $l = 1, \ldots, L_k$. Substituting these samples in (22) yields B posterior draws of $\hat{p}_k$. The Bayesian $100 \times (1-\alpha)\%$ credible interval (CI) of $\hat{p}_k$ is constructed by taking the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution.

Finally, we propose a procedure to estimate P_k in those areas where no individuals have been sampled (the so-called *off-sample* areas). Consider an area k^* that has not been sampled and thus has no observations available. To obtain estimates of p_{lk^*}, we consider all unique weights that are observed in the sample, $\tilde{w}_{[\delta]} = (\tilde{w}_{\delta_1}, \ldots, \tilde{w}_{\delta_L})$. These weights are used in (18) to get estimates of p_{lk^*} and the estimate of P_{k^*} is of the form

$$\hat{p}_k^* = \frac{1}{\sum_{l=1}^{L} \tilde{N}_{lk}} \sum_{l=1}^{L} \tilde{N}_{lk}\hat{p}_{lk^*}. \quad (23)$$

3.3. Implementation

The Bayesian hierarchical model for the outcomes (12) - (17) and the multinomial model (20) described in Sections 3.1 and 3.2 are fitted using the integrated nested Laplace approximations (INLA) approach by Rue et al. (2009). INLA yields a computationally convenient alternative to Markov chain Monte Carlo (MCMC) techniques. This method combines Laplace approximations and numerical integration in a very efficient manner to carry out a Bayesian analysis. A multinomial likelihood, needed to fit the multinomial model (20), is not directly available in INLA. Instead, we employ the multinomial-Poisson transformation of Baker (1994) to fit model (20). The sum-to-zero constraint of the random effects v_k is default when using INLA. Sampling using this constraint is achieved by considering the intrinsic Gaussian Markov random field representation of the ICAR model for which, in addition, a linear constraint is assumed (Rue and Held, 2005 and Rue et al., 2009; See the Supplementary Materials for more detailed information).

We used `R` version 3.2 to fit the models using the `INLA` package (Martino and Rue, 2009). Sampling from the posterior distributions obtained from INLA is done via the *inla.posterior.sample()* function. More details on the implementation and example code are given in the Supplementary Materials. In the Supplementary Materials, an `R` script to perform the analyses and a simulated dataset are attached.

4. Simulation Study

4.1. Simulation Setup

In this section we describe the setup of the simulation study to evaluate the performance of the different small area estimators described in this article. As geography, we took the administrative district division of Belgium (see Figure 1 and Section 5). The total region consists out of 43 districts. Population sizes stratified by five-year age-groups and gender (yielding a total of J=36 strata) at each district are available. The total population size is around ten million. Let x_a denote the indicator for the different age-groups ($x_a = 1$ for ages 0-4, $x_a = 2$ for ages 5-9,. . .,$x_a = 18$ for ages 85+). Let x_g be a gender indicator taking the values 0 or 1. Let $Y_{i(j)k}$ denote the response value of the ith individual belonging to the jth stratum in district k.

In each district k and for $i = 1, \ldots, N_k$, binary outcomes were simulated, namely $Y_{i(j)k} \sim \text{Bin}(p_{jk})$ where $p_{jk} = \Pr(Y_{i(j)k} = 1)$ is the prevalence of a

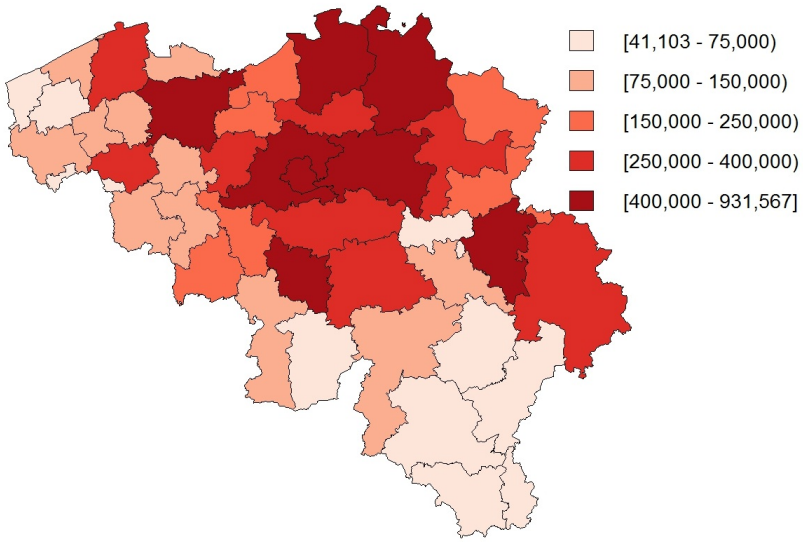

Figure 1: Map of Belgium divided in the 43 administrative districts with accompanying population size in each district.

certain health outcome for individuals belonging to stratum j in area k. The following six models were considered for the prevalences p_{jk}:

$$
\begin{aligned}
(P1) \quad & p_{jk} = 0.20, \\
(P2) \quad & \text{logit}(p_{jk}) = \text{logit}(0.10) + 0.15 x_{a,i(j)k}, \\
(P3) \quad & \text{logit}(p_{jk}) = \text{logit}(0.10) + 0.15 x_{a,i(j)k} - 0.50 x_{g,i(j)k}, \\
(P4) \quad & \text{logit}(p_{jk}) = \text{logit}(0.20) + u_k + v_k, \\
(P5) \quad & \text{logit}(p_{jk}) = \text{logit}(0.10) + 0.15 x_{a,i(j)k} + u_k + v_k, \\
(P6) \quad & \text{logit}(p_{jk}) = \text{logit}(0.10) + 0.15 x_{a,i(j)k} - 0.50 x_{g,i(j)k} + u_k + v_k.
\end{aligned}
$$

The effects $u_k \sim \mathcal{N}(0, 0.10)$ are spatially unstructured effects. The v_k are spatially correlated effects that are sampled from a zero mean ICAR model with a variance of 0.20. The random effects of this ICAR model were generated using INLA. In $(P1)$ the prevalence is constant over all strata and districts. In $(P2)$ the prevalence increases with age and is spatially independent. In $(P3)$ the prevalence also increases with age and is spatially independent, but women have a smaller prevalence than men. Models $(P4)$, $(P5)$ and $(P6)$ additionally assume that the prevalences vary across districts.

A survey sample was taken from the simulated population by the follow-

Table 2: Different scenarios presenting the hypothetical sampling proportions for different age and gender groups. These sampling proportions are used to construct probabilities q_{jk} used for the sampling of individuals from the generated population (see text). SRS: simple random sampling.

	Scenario	[0-20[y.	[20-35[y.	[35-50[y.	[50-65[y.	65+ y.
Male	$(S1)$	SRS	SRS	SRS	SRS	SRS
	$(S2)$	0.20	0.12	0.10	0.06	0.02
	$(S3)$	0.16	0.12	0.12	0.06	0.04
	$(S4)$	0.15	0.12	0.07	0.06	0.10
Female	$(S1)$	SRS	SRS	SRS	SRS	SRS
	$(S2)$	0.20	0.12	0.10	0.06	0.02
	$(S3)$	0.16	0.12	0.12	0.06	0.04
	$(S4)$	0.15	0.12	0.07	0.06	0.10

ing procedure:
(1) Select districts from which samples are drawn. First, we simulated a random number, n_{area}, from the set $\{39, 40, 41, 42, 43\}$. Next, we sampled n_{area} from the 43 districts with probability-proportional-to-size sampling where the size variables are the population sizes of the districts. In this manner, districts with a large population size were sampled with probability one, whereas districts with a small population size had a probability smaller than one. Note that a small number of off-sample areas were created in this manner.
(2) The total sample size was randomly sampled by drawing a random number from the set $\{4000, 4001, \ldots, 6000\}$. Next, a multinomial distribution with probabilities proportional to the district population size was used to draw sample sizes in each selected district, denoted by n_k.
(3) The sample sizes in the J strata of area k, denoted by n_{jk}, were generated from a multinomial distribution with total sample size n_k and probabilities q_{jk}. Four scenarios were considered and the probabilities q_{jk} were constructed using the hypothetical sampling proportions in different age groups as presented in Table 2. In scenario $(S1)$ simple random sampling (SRS) is performed which implies that $q_{jk} = N_{jk}/N_k$. In scenario $(S2)$, the hypothetical sampling proportion of the [0-20[year-old males is 0.20. In this [0-20[year-old males subgroup simple random sampling is performed which implies that $q_{jk} = 0.20 \times \frac{N_{jk}}{N_k^{(male,[0-20[\ y.)}}$ for the strata j belonging to this subgroup, where $N_k^{(male,[0-20[\ y.)}$ is the population size in district k of the [0-20[year-old males subgroup. For the other subgroups in Table 2, the probabilities q_{jk} were calculated in a similar manner. Finally, we randomly sampled n_{jk} individuals in strata j in area k from the simulated population.

For a simulated sample, the survey design weight of a sampled individual i of area k, w_{ik}^d, is equal to the inverse of the probability of inclusion π_{ik} of this individual in the sample. The calculation of the probabilities of inclusion is presented in the Supplementary Materials. The survey design weights were adjusted with a post-stratification factor f_{ps} to form the final weights used in the analysis, namely $w_{ik} = w_{ik}^d \times f_{ps}$, with the strata of the post-stratification defined by the age-groups [0-10[, [10-20[, [20-30[, [30-40[, [40-50[, [50-60[, [60-70[, [70-80[, 85+ and gender.

In sampling scenario $(S1)$ there is no relationship between the values of the weights w_{ik} and age. In scenarios $(S2)$ and $(S3)$ the values of the weights w_{ik} increase with the age groups considered in Table 2 and the weights in scenario $(S2)$ are more dispersed than in scenario $(S3)$. In scenario $(S4)$ lower weights are obtained for the age groups [0-20[, [20-35[and 65+, and higher weights for the age groups [35-50[and [50-65[. Plots of the distribution of the generated weights w_{ik} are given in the Supplementary Materials.

For each combination of a prevalence model and a sampling scenario (6×4 combinations) we ran S times through steps (1) - (3) to obtain S simulated datasets. Each simulated dataset contains only the outcome, the area indicator and the final weight w_{ik}. Two direct and nine indirect estimators were used to estimate the small-area prevalences from the simulated datasets. As direct estimators we used the unweighted mean (UM) and the HT estimator given in (3). The five indirect estimators described in Section 2.3 were used, namely the NB-estimator in (5), the LN-estimator in (7), the AS-estimator in (8), the PL-estimator in (9) and the ES-estimator in (10). Finally, we calculated the prevalence using the indirect estimator (11), or equivalently (22), using the models presented in (13) and (14), respectively further referred to as *model-based model 1* (M1) and *model-based model 2* (M2). We then further used a random walk model (RW1) or penalized splines (PS). For all indirect estimators we included both independent and spatial ICAR random effects into the model.

To evaluate the different estimates we compared two statistics: the estimated squared bias and the estimated mean squared error (MSE). Denote by P_k the true proportion in area k (which stays constant across the simulations) and denote by $\hat{p}_k^{(s)}$ the estimated proportion from the sth simulated

Table 3: The average squared bias and mean squared error for 11 estimators based on 100 simulated datasets using the prevalence models (P1), (P2), (P3) and the four sampling scenarios. Figures in bold denote the row minimum.

		UM	HT	NB	LN	AS	PL	ES	M1 RW1	M2 RW1	M1 PS	M2 PS
Bias^2 ($\times 10^3$)												
(P1)	(S1)	0.02	0.03	0.00	0.01	0.01	0.00	0.00	**0.00**	0.00	0.00	0.00
(P1)	(S2)	0.04	0.05	**0.00**	0.04	0.01	0.00	0.01	0.00	0.01	0.00	0.00
(P1)	(S3)	0.04	0.05	0.00	0.02	0.01	0.00	0.00	0.00	0.00	**0.00**	0.00
(P1)	(S4)	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	**0.00**
(P2)	(S1)	0.03	0.03	0.03	0.04	0.02	0.03	0.03	0.02	**0.02**	0.02	0.02
(P2)	(S2)	4.01	0.09	4.06	0.04	0.04	0.03	**0.03**	0.66	0.22	0.14	0.13
(P2)	(S3)	1.76	0.05	1.75	0.04	**0.02**	0.02	0.02	0.15	0.07	0.05	0.05
(P2)	(S4)	0.26	0.06	0.21	0.03	**0.02**	0.02	0.02	0.10	0.09	0.12	0.12
(P3)	(S1)	0.04	0.04	0.03	0.04	**0.02**	0.03	0.03	0.02	0.02	0.02	0.02
(P3)	(S2)	3.05	0.10	3.13	0.04	0.02	**0.02**	0.02	0.57	0.22	0.11	0.09
(P3)	(S3)	1.32	0.05	1.33	0.04	**0.01**	0.02	0.02	0.13	0.07	0.05	0.04
(P3)	(S4)	0.17	0.05	0.12	0.03	**0.01**	0.02	0.02	0.07	0.07	0.08	0.08
MSE ($\times 10^3$)												
(P1)	(S1)	2.81	2.82	0.09	0.09	0.48	0.09	**0.09**	0.09	0.09	0.09	0.09
(P1)	(S2)	2.72	3.95	**0.09**	0.29	0.79	0.28	0.27	0.15	0.23	0.11	0.11
(P1)	(S3)	2.55	3.01	**0.08**	0.12	0.53	0.12	0.12	0.16	0.16	0.10	0.10
(P1)	(S4)	2.81	3.09	**0.09**	0.11	0.52	0.10	0.10	0.16	0.15	0.09	0.09
(P2)	(S1)	3.66	3.68	0.17	**0.16**	0.63	0.17	0.16	0.40	0.40	0.37	0.37
(P2)	(S2)	7.18	6.44	4.17	0.65	1.59	0.73	0.75	1.18	0.88	0.44	**0.41**
(P2)	(S3)	4.90	4.56	1.88	0.30	0.96	0.31	0.33	0.54	0.47	0.29	**0.28**
(P2)	(S4)	3.71	4.04	0.34	**0.20**	0.74	0.21	0.21	0.63	0.62	0.60	0.60
(P3)	(S1)	3.22	3.24	0.14	0.14	0.54	0.14	**0.13**	0.34	0.34	0.31	0.32
(P3)	(S2)	5.81	5.84	3.23	0.61	1.47	0.64	0.69	1.07	0.78	0.35	**0.32**
(P3)	(S3)	4.09	4.12	1.42	0.26	0.85	0.25	0.27	0.49	0.38	0.23	**0.22**
(P3)	(S4)	3.41	3.74	0.24	**0.13**	0.67	0.18	0.18	0.51	0.48	0.46	0.45

dataset. The statistics were calculated as:

$$\begin{aligned} \text{Bias}^2 &= \frac{1}{K}\sum_{k=1}^{K}(\bar{p}_k - P_k)^2, \text{ where } \bar{p}_k = \frac{1}{S}\sum_{s=1}^{S}\hat{p}_k^{(s)} \\ \text{MSE} &= \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{S}\sum_{s=1}^{S}(\hat{p}_k^{(s)} - P_k)^2\right). \end{aligned}$$

We also calculated the nominal coverage and the average length of the 95% credible intervals of the prevalence estimates.

4.2. Simulation Results

The squared bias and mean squared error results are presented in Table 3 and Table 4 with all results based on one hundred simulations (S=100).

We first discuss the results of the scenarios in which the small area prevalences are not spatially varying (Table 3). For scenario ($P1$) all estimators

Table 4: The average squared bias and mean squared error for 11 estimators based on 100 simulated datasets using the prevalence models (P4), (P5), (P6) and the four sampling scenarios. Figures in bold denote the row minimum.

		UM	HT	NB	LN	AS	PL	ES	M1 RW1	M2 RW1	M1 PS	M2 PS
Bias^2 ($\times 10^3$)												
(P4)	(S1)	**0.03**	0.03	0.52	0.71	0.41	0.52	0.55	0.52	0.52	0.52	0.52
(P4)	(S2)	**0.04**	0.04	0.58	0.79	0.43	0.48	0.58	1.37	1.35	0.58	0.58
(P4)	(S3)	**0.03**	0.04	0.56	0.74	0.41	0.52	0.57	1.02	0.81	0.56	0.56
(P4)	(S4)	**0.02**	0.02	0.50	0.69	0.37	0.48	0.52	0.62	0.66	0.50	0.50
(P5)	(S1)	0.03	**0.03**	0.73	0.92	0.56	0.73	0.75	0.65	0.65	0.69	0.69
(P5)	(S2)	4.08	**0.12**	4.74	0.91	0.59	0.73	0.74	3.58	2.30	0.94	0.91
(P5)	(S3)	1.93	**0.06**	2.59	0.90	0.58	0.76	0.74	3.09	1.65	0.81	0.81
(P5)	(S4)	0.28	**0.04**	1.03	0.91	0.54	0.70	0.73	1.43	1.29	0.76	0.76
(P6)	(S1)	0.02	**0.02**	0.65	0.86	0.50	0.65	0.68	0.58	0.58	0.61	0.61
(P6)	(S2)	3.27	**0.08**	3.92	0.81	0.50	0.62	0.63	2.90	1.89	0.80	0.77
(P6)	(S3)	1.54	**0.06**	2.19	0.81	0.52	0.68	0.66	2.69	1.32	0.72	0.72
(P6)	(S4)	0.23	**0.06**	0.89	0.81	0.47	0.60	0.64	1.26	1.04	0.67	0.67
MSE ($\times 10^3$)												
(P4)	(S1)	2.65	2.66	1.49	1.57	1.50	1.49	1.50	1.49	1.49	**1.48**	1.49
(P4)	(S2)	2.60	3.75	1.56	2.11	2.01	2.04	2.03	2.23	2.34	1.56	**1.56**
(P4)	(S3)	2.46	2.86	**1.50**	1.73	1.63	1.66	1.65	2.01	1.91	1.51	1.51
(P4)	(S4)	2.55	2.80	1.45	1.62	1.54	1.54	1.54	1.77	1.74	1.45	**1.44**
(P5)	(S1)	3.47	3.49	2.12	2.18	2.15	2.12	**2.12**	2.24	2.24	2.23	2.23
(P5)	(S2)	6.98	5.71	5.81	3.13	3.28	3.20	3.14	4.69	3.89	**2.56**	2.57
(P5)	(S3)	4.80	4.18	3.70	2.48	2.56	2.49	2.47	4.04	3.30	**2.27**	2.28
(P5)	(S4)	3.52	3.67	2.29	2.27	2.25	2.22	**2.21**	3.15	2.95	2.30	2.30
(P6)	(S1)	3.08	3.09	**1.82**	1.90	1.86	1.82	1.82	1.92	1.92	1.92	1.92
(P6)	(S2)	5.93	5.57	4.86	2.99	3.12	3.05	3.02	4.03	3.55	**2.25**	2.25
(P6)	(S3)	4.27	3.94	3.18	2.19	2.27	2.20	2.18	3.47	2.89	**2.03**	2.04
(P6)	(S4)	3.30	3.49	2.01	2.01	1.98	**1.94**	1.94	2.71	2.60	2.01	2.01

have a low squared bias. The MSE of the two direct estimators is large due to the increased variance associated with these estimators. The variance and thus the MSE of the indirect estimators is smaller, showing the benefit of spatial smoothing methods. For the prevalence model ($P2$) and ($P3$), except in the SRS case ($S1$), the unweighted mean and the naive binomial - methods ignoring the survey weights - have a large squared bias. Again, the MSE of the direct estimators is large. For sampling scenarios ($S2$) and ($S3$), it is observed that the proposed model-based approach using the penalized splines performs best in terms of MSE. This can be expected, since these scenarios imply a relationship between the design weights and the responses, namely both the prevalences and the values of the weights increase with age, and models (13) and (14) exploit that relationship. For ($S4$), on the contrary, our proposed methods perform less good than the methods LN, PL and ES. In this scenario there is no clear relationship between the design weights and the responses and thus using a model of the form (13) or (14) has no benefit

Table 5: The nominal coverage and length of the 95% credible intervals for 11 estimators based on 100 simulated datasets using the prevalence models (P4), (P5), (P6) and the four sampling scenarios. The results are averaged over all districts.

		UM	HT	NB	LN	AS	PL	ES	M1 RW1	M2 RW1	M1 PS	M2 PS
Nominal coverage												
(P4)	(S1)	0.96	0.96	0.95	0.94	0.96	0.95	0.95	0.95	0.95	0.95	0.95
(P4)	(S2)	0.96	0.94	0.94	0.89	0.92	0.91	0.90	0.80	0.80	0.94	0.94
(P4)	(S3)	0.95	0.95	0.94	0.92	0.95	0.93	0.93	0.85	0.88	0.94	0.94
(P4)	(S4)	0.96	0.96	0.95	0.94	0.95	0.94	0.94	0.91	0.92	0.95	0.95
(P5)	(S1)	0.95	0.95	0.95	0.94	0.96	0.95	0.95	0.95	0.95	0.95	0.94
(P5)	(S2)	0.70	0.95	0.63	0.88	0.88	0.89	0.88	0.64	0.78	0.96	0.96
(P5)	(S3)	0.84	0.95	0.80	0.92	0.92	0.92	0.92	0.72	0.85	0.96	0.95
(P5)	(S4)	0.94	0.95	0.92	0.94	0.95	0.94	0.94	0.84	0.87	0.94	0.94
(P6)	(S1)	0.95	0.95	0.94	0.94	0.96	0.94	0.94	0.94	0.94	0.94	0.94
(P6)	(S2)	0.73	0.93	0.66	0.87	0.86	0.87	0.86	0.65	0.78	0.95	0.96
(P6)	(S3)	0.86	0.95	0.81	0.92	0.92	0.92	0.92	0.72	0.85	0.95	0.95
(P6)	(S4)	0.94	0.95	0.92	0.93	0.95	0.94	0.94	0.85	0.87	0.94	0.94
Average length												
(P4)	(S1)	0.18	0.18	0.15	0.15	0.16	0.15	0.15	0.15	0.15	0.15	0.15
(P4)	(S2)	0.18	0.21	0.15	0.16	0.16	0.15	0.15	0.13	0.14	0.15	0.15
(P4)	(S3)	0.18	0.19	0.15	0.15	0.16	0.15	0.15	0.14	0.15	0.15	0.15
(P4)	(S4)	0.18	0.19	0.15	0.15	0.16	0.15	0.15	0.15	0.15	0.15	0.15
(P5)	(S1)	0.20	0.20	0.17	0.18	0.18	0.17	0.17	0.18	0.18	0.18	0.18
(P5)	(S2)	0.19	0.25	0.16	0.18	0.18	0.18	0.17	0.14	0.17	0.20	0.20
(P5)	(S3)	0.19	0.22	0.16	0.17	0.18	0.17	0.17	0.15	0.17	0.19	0.19
(P5)	(S4)	0.20	0.21	0.17	0.18	0.18	0.17	0.17	0.17	0.17	0.18	0.18
(P6)	(S1)	0.19	0.19	0.16	0.17	0.17	0.16	0.16	0.17	0.17	0.17	0.17
(P6)	(S2)	0.18	0.24	0.15	0.17	0.17	0.17	0.16	0.14	0.17	0.19	0.19
(P6)	(S3)	0.19	0.22	0.15	0.16	0.17	0.16	0.16	0.13	0.17	0.18	0.18
(P6)	(S4)	0.19	0.20	0.16	0.17	0.17	0.16	0.16	0.15	0.16	0.17	0.17

above the other indirect methods. We further observed a small decrease in the mean squared error between sampling scenarios ($S2$) and ($S3$) due to the influence of the less dispersed weights of the latter scenario.

The results with spatially varying prevalences are summarized in Table 4. It is observed that the HT estimator has a small bias over all scenarios. The unweighted mean is unbiased for ($P4$) and in the SRS scenarios ($S1$). The indirect estimators have a larger squared bias than the HT estimator since these methods shrink the small area prevalences towards the overall population prevalences through the spatial random effects. The estimators ignoring the survey weights, UM and NB, have larger bias for ($S2$), ($S3$) and ($S4$). In terms of MSE, the naive binomial and the proposed model-based approach using the penalized splines perform the best for ($P4$). For ($P5$) and ($P6$) again the proposed model-based approach performs well. Whereas for ($S4$) the methods PL and ES perform somewhat better.

The results of the nominal coverage and average length of the 95% credible

intervals are shown in Table 5. We only present the results of settings ($P4$), ($P5$) and ($P6$) here. Results for ($P1$), ($P2$) and ($P3$) are qualitatively similar (Supplementary Material). The HT estimator has a nominal coverage around 95% for all scenarios. The unweighted mean and naive binomial have poor coverages for the combinations ($P5$) and ($P6$) with ($S2$) and ($S3$) due to the bias of the estimators in these settings. The nominal coverage of the LN, AS, PL and ES methods have an undercoverage in some scenarios. The proposed model-based approach using the random walk does not perform well. The model-based approach using the penalized splines, on the other hand, has a coverage around 95% in all scenarios. The average length of the CIs of the indirect estimators is smaller than the length of the HT estimator CIs.

In general, the two methods ignoring the sampling weights (unweighted mean and naive binomial) produce poor estimates due to the bias of these methods. The HT estimator is unbiased but has a large variance, making it unsuitable for practical usage. The indirect estimators accounting for the weights produce estimates with both small bias and small mean squared errors. The model-based approach using the penalized splines described in Section 3 is preferred when there is a relationship between the survey weights and the responses, otherwise, the LN, PL and ES methods are preferable. The performance of the penalized splines is better than the random walk models. For the penalized spline model-based approach the difference in performance between models (13) and (14) is negligible.

In the Supplementary Material, we present the results with respect to the sensitivity of the results to other prior distributions. It was observed that the obtained results are insensitive to other (vague) prior distribution choices. In addition, we also present in the Supplementary Material the results with respect to the off-sample areas separately. These results are presented separately to be able to evaluate the described approaches with respect to the estimation and prediction in areas where no data is available. The conclusions from this off-sample analysis are similar as the conclusions discussed above.

5. Application to Belgian Health Interview Survey

Next, we focus our attention on empirical data measuring the prevalence of asthma across the 43 districts shown in Figure 1 using the 2001 Belgian Health Interview Survey (HIS). Data were collected in response to the question "Have you experienced asthma in the previous year?". In total, 12,003 individuals responded to this question. The number of respondents per district varied between 50 and 2,949, and 4 districts were not selected in the survey. In total 612 (5.1 %) individuals responded positive to the question. The 2.5% and 97.5% quantiles of normalized weights calculated via (2) are 0.35 and 2.49, respectively. The minimum and maximum normalized weights are 0.06 and 10.49, respectively.

We calculated the small areas prevalences of asthma using the eleven estimators that were considered in the simulation study in Section 4. Figure 2 presents the violin plots of the predicted prevalences of asthma by district using the different estimators. It is clear that the direct estimators

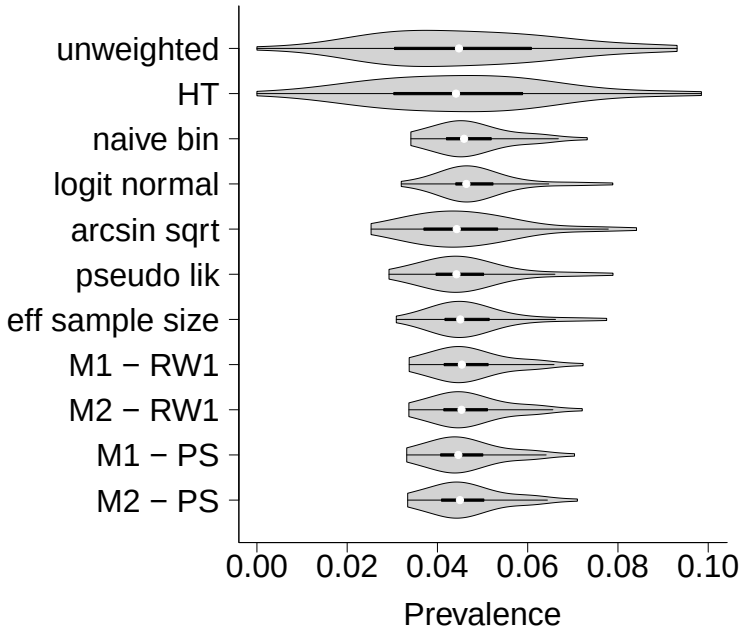

Figure 2: Violin plot of the predicted prevalence estimates of asthma, using various approaches, across the 43 districts in Belgium estimated from the Health Interview Survey of 2001.

(unweighted mean and HT estimator) have a large amount of heterogeneity amongst the districts with predicted asthma prevalences ranging from 0.0 to 0.10. This variability is substantially reduced by using the nine indirect approaches that use Bayesian hierarchical models. The shape and location of the violin plots are similar for the indirect estimators with the arcsin square root transformation and pseudo-likelihood binomial approach showing the most heterogeneity amongst districts. The results of the model-based approaches are fairly similar. In the Supplementary Materials, we give the point estimates and associated CI per district.

In Figure 3, we display maps of the predicted prevalences calculated using the unweighted mean, the HT estimator, the naive binomial approach and the proposed model-based approaches as described in Section 3 (lower panels). It is observed that the naive binomial approach yields quite similar results as the proposed model-based approaches. We expected this result, since the goal of the sampling design of the HIS (Demarest et al., 2001) is to obtain a sample which is as close as possible to simple random sampling. The estimated prevalences are highest in the districts of Nivelles and Soignies, with predicted prevalences (using the model-based model 1 with PS estimator) of 7.04% [95% CI: 5.00 - 9.98] and 6.25% [95% CI: 4.23 - 9.50], respectively. Figure 4 presents the estimated cell probabilities obtained via model (13) with PS (left panel) and the estimated cell size proportions (right panel) for the district of Nivelles. The 95% credible intervals of $\hat{p}_{ik}$ are calculated using (18) with the parameter estimates replaced by their posterior samples. It is observed that the relationship between the normalized weights and probability on asthma is non-linear. The credible intervals of $\hat{p}_{ik}$ are somewhat inflated near the boundary values of $\tilde{w}_{ik}$ since the variability of $\hat{f}(\tilde{w}_{ik})$ is inflated near the boundary. In the Supplementary Materials we also present the estimated cell size proportions obtained via model (13) with RW1. It can be observed that the PS approach is better able to capture a non-linear trend. Under RW1 the estimated probabilities $\hat{p}_{ik}$ are smaller than PS for small weights and approach a constant probability around 0.07 after a small initial drop. The strata size proportions $\tilde{N}_{lk}/N_k$ depend on the normalized weights in an irregular structure since the sample sizes in the strata are different. Figure 5 presents prevalence maps where the results are obtained via four spatial smoothing methods that account for the spatial weights that are described in Mercer et al. (2014). There are no important differences between these maps and the maps produced by the model-based approaches (Figure 3) as described in Section 3. The results of the model-based approaches are

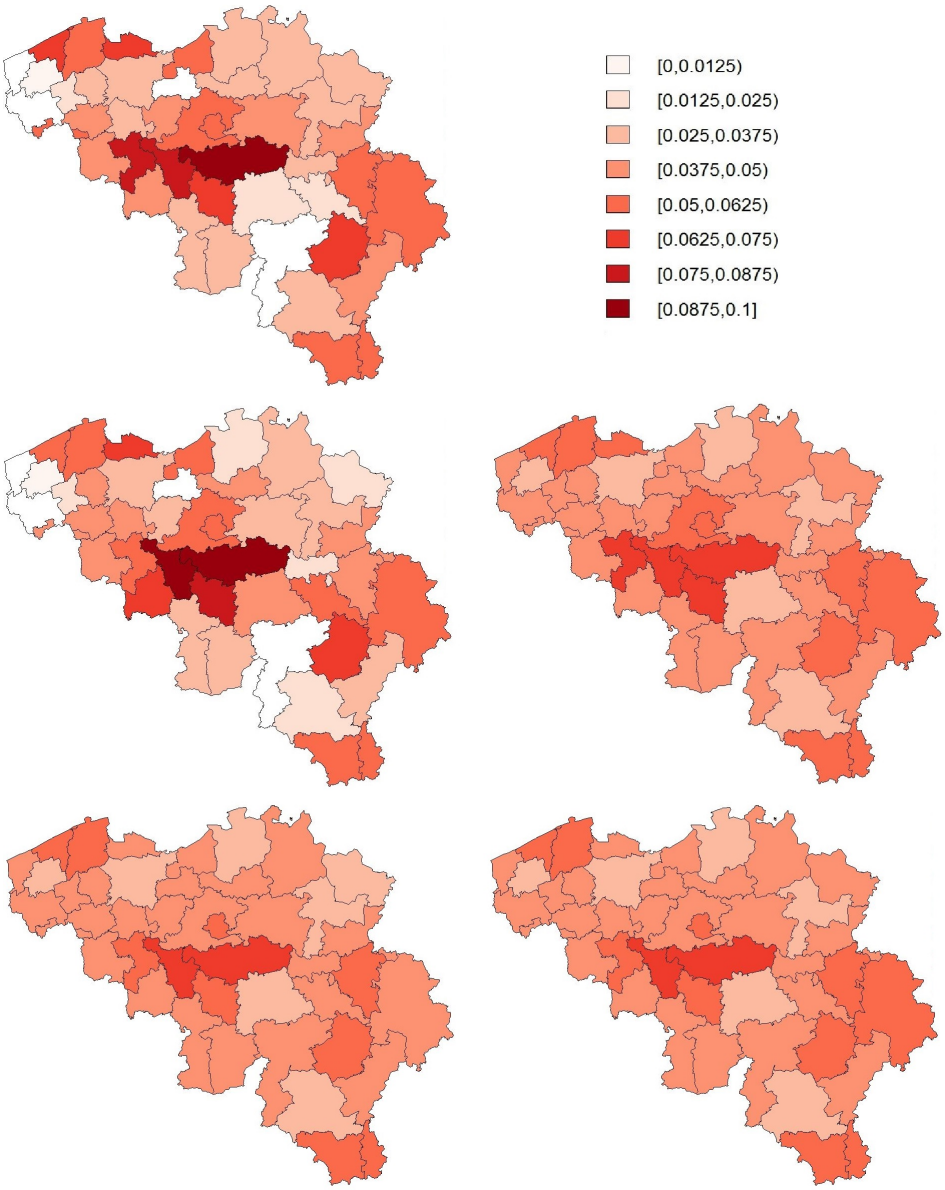

Figure 3: Predicted asthma prevalences by district in Belgium using the 2001 Belgian Health Interview Survey. The obtained estimates are the unweighted mean (top left), the Horvitz-Thompson estimator (middle left), the naive binomial approach (middle right), the model-based approach using model (13) with RW1 (bottom left) and the model-based approach using model (13) with PS (bottom right).

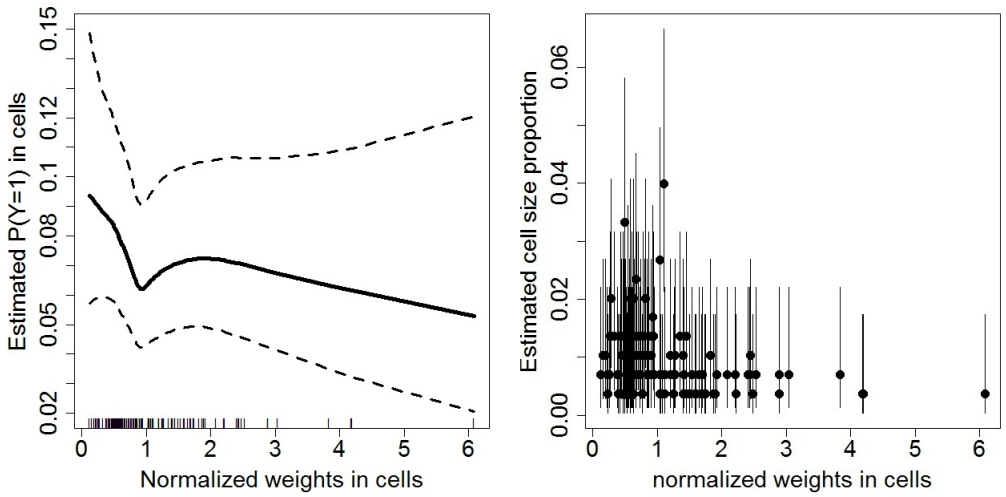

Figure 4: Estimated cell probabilities obtained via (13) with PS together with the 95% credible intervals (left panel) and cell size proportions obtained via (20) (right panel) for the district of Nivelles using the model-based approach. In the right panel, the dots indicate the posterior mean and the black vertical lines are the 95% credible intervals.

not influenced by the choice of the prior distributions (see Supplementary Materials).

6. Discussion

We have presented a predictive model-based approach for the estimation of small area estimates from a health survey in which the survey weights of the sampled individuals are the only information available on the survey design. Our approach uses a hierarchical Bayesian model in which the health outcomes are regressed via a non-parametric function on the normalized survey weights to obtain predictions of the outcome for the non-sampled individuals. The hierarchical model accounts for the spatial distribution by using both spatially unstructured and spatially structured random effects. Simultaneously, the survey weights themselves are modelled to estimate the survey weights of the non-sampled individuals. In the simulation study, the benefits of using indirect methods that account for the survey weights based on hierarchical models are clearly observable. The simulation study indicates that our proposed model-based approach performs well, especially when there

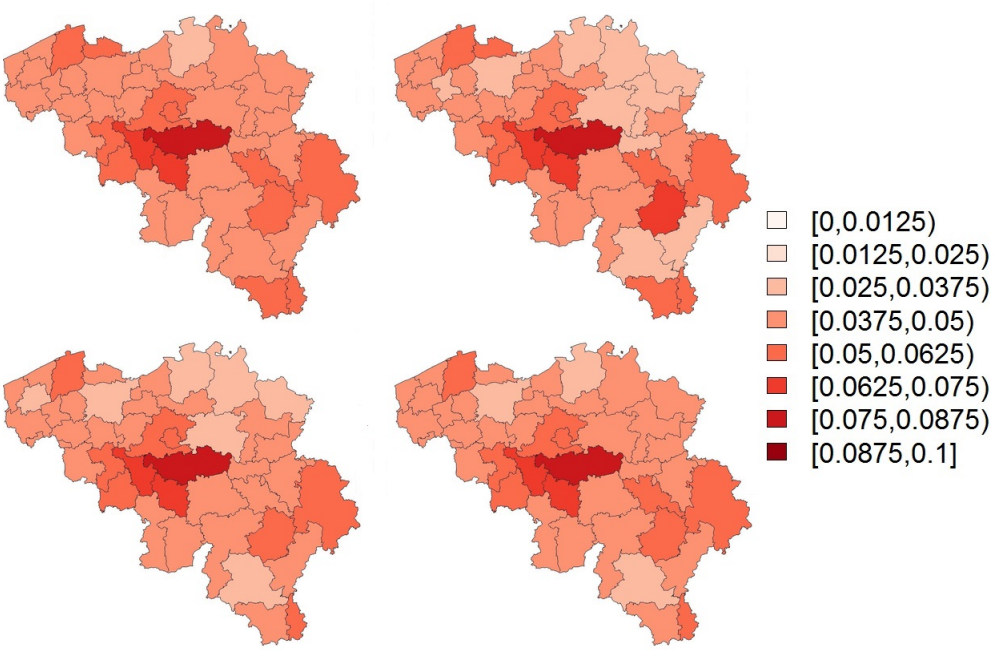

Figure 5: Predicted asthma prevalences by district in Belgium using the 2001 Belgian Health Interview Survey. The obtained estimates are obtained via four estimators described in the overview paper of Mercer et al. (2014): the logit normal estimator (top left), the arcsin square root transformation estimator (top right), the pseudo-likelihood binomial estimator (botom left) and the effective sample size adjusted estimator (bottom right).

is a relationship between the responses and the sampling weights. The 95% credible intervals of the described model-based approach provide good nominal coverage results with short average interval lengths. The results are not sensitive to different choices of prior distributions for the hyperparameters. The proposed methodology is based on Si et al. (2015) describing similar approaches to estimate the finite population mean from a survey. We extended their work to the context of small area estimation.

Both a first-order random walk and a penalized spline were considered for the regression of the health outcome on the survey weights. A first-order random walk can be seen as the Bayesian counterpart of P-splines in which abrupt jumps between two successive spline parameters $\beta_m - \beta_{m-1}$ are penalized (Brezger and Lang, 2008). A first-order random walk was preferred over a second-order random walk since it was observed that the second-order

random walk yielded unstable results for some datasets in the simulation study. Overall, we observed in the simulation study that the penalized spline approach outperforms the random walk. Knot selection for the penalized spline, both in terms of number of knots and placement of the knots, is out of the scope of the manuscript. Techniques described in, for example, Ruppert et al. (2003) could be used for this purpose. A sensitivity analysis (results not shown) indicated that our choice of $B = 20$ was sufficient for the analyses presented here.

To implement our methods we used the `inla` package within the `R` computing environment. In this paper, we have observed that `inla` is very accurate and yields trustworthy results. However, in general we advise researchers to check their results carefully, especially for rare binary events and small sample sizes, since `inla` can produce inaccurate results in these cases (Fong et al., 2010). The procedure to obtain estimates for off-sample areas is easily implemented in `inla` by extending the data with the appropriate set of weights and treat the response data as missing. In this manner, predictions for these responses are done as part of the model fitting. We opted for implementing the estimation for the mean and the distribution estimation for the weights separately to make sure that we can make use of standard `inla` functions. Part of the uncertainty is not accounted for in this manner, however, from the simulation study we observed that the nominal coverage of the 95% credible intervals of the proposed methods are satisfactory.

This paper presented methodology for binary distributed health outcomes. Generalization to other distributions of the health outcome could be done without much effort. The model is also easily extended to include additional predictors. In addition, this proposed methodology can also be used for surveys in which the variables on which the sampling design is based are known. Suppose for example that the survey design selects people according to age, gender and educational level. It seems unlikely that census data per small area which is fully cross-classified over these three variables is available (only marginals but no full cross-classifications). The proposed approaches described in the paper are useful in this example by modelling the health outcome on these three variables and simultaneously making inference about the unknown population sizes in the cross-classified cells. This is a topic that is currently under investigation. Investigating how the proposed models can be used and how they perform for the calculation of subgroup means is another interesting topic of future research.

Another possible extension is replacing the spatially structured random

effects by a smooth, non-parametrically specified trend. This was first proposed by Opsomer et al. (2008). This would be useful when more specific information on the spatial locations (more specific than the small area of interest) of the sampled individuals is available.

The hierarchical model approaches described in this paper borrow strength from neighbouring areas. These models are useful to detect hot spot clustering. However, their ability to detect localized hot spots (clusters) is questionable because the models include a global smoothing mechanism (Lawson and Denison, 2002). In the literature, some authors described Bayesian approaches addressing the hot spot (cluster) identification problem. We refer to Lawson and Denison (2002) for an overview.

The described approaches assume that all unique values of the weights have been observed. Si et al. (2015) argued that this is a possible concern of the proposed methods when it is known that some large weights in the population have not occurred in the sample. In this case, this should be taken into account in the setup of the model. A second concern arises in multipurpose health surveys in which many health outcomes are of interest. Our proposed methods should be repeated separately for each outcome. This is computationally intensive, however, we believe that in the modern era where cluster and parallel computing is available this should not be a major obstacle.

Acknowledgements

Support from a doctoral grant of Hasselt University is acknowledged (BOF11D04FAEC to YV). Support from the National Institutes of Health is acknowledged [award number R01CA172805 to CF]. Support from the University of Antwerp scientific chair in Evidence-Based Vaccinology, financed in 2009–2015 by a gift from Pfizer, is acknowledged [to NH]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. This research is supported in part by funding under grant NIH R01CA172805 [CF, RK, AL].

Supplementary Materials

The reader is referred to the Supplementary Materials for more information on the implementation of the described models, additional results of the simulation study, and additional results for the application study.

References

Baker, S. G., 1994. The multinomial-Poisson transformation. The Statistician 43 (4), 495–504.

Besag, J., York, J. C., Mollie, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43, 1–59.

Brezger, A., Lang, S., 2008. Simultaneous probability statements for Bayesian P-splines. Statistical Modeling 8, 141–186.

Chen, C., Wakefield, J., Lumley, T., 2014. The use of sample weights in Bayesian hierarchical models for small area estimation. Spatial and Spatio-temporal Epidemiology 11, 33–43.

Chen, Q., Elliott, M. R., Little, R. J. A., 2010. Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. Survey Methodology 36 (1), 23–34.

Cochran, W. G., 1977. Sampling techniques. Hoboken: John Wiley & Sons, Inc.

Congdon, P., Lloyd, P., 2010. Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. Journal of Data Science 8, 235–252.

Datta, G. S., Ghosh, M., 1991. Bayesian prediction in linear models: Applications to small area estimation. Annals of Statistics 19, 1748–1770.

Demarest, S., Tafforeau, J., Van Oyen, H., Bruckers, L., Molenberghs, G., Tibaldi, F., Van Steen, K., 2001. Health Interview Survey 2001: Protocol for the sampling design. Brussel, Wetenschappelijk Instituut Volksgezondheid, Afdeling Epidemiologie.

Eberly, L. E., Carlin, B. P., 2000. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Statistics in Medicine 19, 2279–2294.

Eilers, P. H. C., Marx, B. D., 1996. Flexible smoothing with B-splines and penalties (with discussion). Statistical Science 11, 89–121.

Elliott, P., Wakefield, J., Best, N., Briggs, D. (Eds.), 2001. Spatial epidemiology: Methods and applications. Oxford: Oxford University Press.

Farrell, P. J., 2000. Bayesian inference for small area proportions. The Indian Journal of Statistics 62, 402–416.

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 (366), 269–277.

Fong, Y., Rue, H., Wakefield, J., 2010. Bayesian inference for generalized linear mixed models. Biostatistics 11 (3), 397–412.

Gelfand, A. E., Diggle, P. J., Fuentes, M., Guttorp, P. (Eds.), 2010. Handbook of Spatial Statistics. Boca Raton: Chapman & Hall/CRC.

Ghosh, M., Natarajan, K., Stroud, T., Carlin, B., 1998. Generalized linear models for small-area estmation. Journal of the American Statistical Association 93 (441), 55–93.

Gonzalez, M. E., 1973. Use and evaluation of synthetic estimators. In: Proceedings of the Social Statistics Section. American Statistical Association, Washington, D.C., pp. 33–36.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. New York: Springer-Verlag.

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.

Jiang, J., Lahiri, P., 2006. Mixed model prediction and small area estimation. Test 1, 1–96.

Kemp, I., Boyle, P., Smans, M., Muir, C. S., 1985. Atlas of cancer in Scotland, 1975-1980: Incidence and epidemiological perspective. Lyon, IARC publication no 72.

Kott, P., 1989. Robust small domain estimation using random effects modelling. Survey Methodology 9, 1–12.

Lawson, A. B., 2013. Bayesian disease mapping: Hierarchical modeling in spatial epidemiology, second edition. Boca Raton: Chapman & Hall/CRC.

Lawson, A. B., Denison, D. G. T., 2002. Spatial Cluster Modelling. Boca Raton: Chapman & Hall/CRC.

MacGibbon, B., Tomberlin, T. J., 1989. Small area estimates of proportions via empirical Bayes techniques. Survey Methodology 15, 237–252.

Malec, D., Sedransk, J., Moriarity, C. L., LeClere, F. B., 1997. Small area inference for binary variables in the National Health Interview Survey. Journal of the American Statistical Association 92 (439), 815–826.

Martino, S., Rue, H., 2009. Implementing approximate Bayesian inference using integrated nested Laplace approximation: A manual for the INLA program. Available from: http://www.r-inla.org/download.

Mason, T. J., 1995. The development of the series of U.S. cancer atlases: Implications for future epidemiologic research. Statistics in Medicine 14, 473–479.

Mercer, L., Wakefield, J., Chen, C., Lumley, T., 2014. A comparison of spatial smoothing methods for small area estimation with sampling weights. Spatial Statistics 8, 69–85.

Opsomer, J., Claeskens, G., Ranalli, M. G., Kauermann, G., Breidt, F. J., 2008. Non-parametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society. Series B 70, 265–286.

Pfeffermann, D., 2013. New important developments in small area estimation. Statistical Science 28, 40–68.

Prasad, N. G. N., Rao, J. N. K., 1999. On robust small area estimation using a simple random effects model. Survey Methodology 25, 67–72.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Raghunathan, T., Xie, D., Schenker, N., Parsons, V., Davis, W., Dood, K., Feuer, E., 2007. Combining information from two surveys to estimate

county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association 102, 474–486.

Rao, J. N. K., 2003. Small Area Estimation. Hoboken: John Wiley & Sons, Inc.

Rao, J. N. K., 2011. Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. Statistical Science 26, 240–256.

Royall, R. M., 1970. On finite population sampling theory under certain linear regression models. Biometrika 57, 377–387.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Boca Ratton: Chapman and Hall/CRC Press.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B 71, 1–35.

Ruppert, D., Wand, M. P., Carroll, R. J., 2003. Semiparametric Regression. Cambridge: University Press.

Schrödle, B., Held, L., 2011. Spatio-temporal disease mapping using INLA. Environmetrics 22, 725–734.

Si, Y., Pillai, N. S., Gelman, A., 2015. Bayesian nonparametric weighted sampling inference. Bayesian Analysis 10, 605–625.

Skinner, C., 1989. Analysis of complex surveys. Wiley, Chichester, Ch. Domain means, regression and multivariate analysis, pp. 59–87.

Stroud, T., 1994. Bayesian inference from categorical survey data. Canadian Journal of Statistics 22, 33–45.

Wakefield, J., 2009. Multi-level modelling, the ecologic fallacy, and hybrid study designs. International Journal of Epidemiology 38, 330–336.

Waller, I. A., Gotway, C. A., 2004. Applied spatial statistics for public health data. Hoboken: John Wiley & Sons, Inc.

You, Y., Rao, J. N. K., 2002. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. The Canadian Journal of Statistics 30, 431–439.

You, Y., Rao, J. N. K., 2003. Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. Journal of Statistical Planning and Inference 111, 197–208.

Zheng, H., Little, R. J. A., 2003. Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. Journal of Official Statistics 19, 99–117.

Zheng, H., Little, R. J. A., 2005. Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. Journal of Official Statistics 21, 1–20.