Spatial smoothing models to deal with the complex sampling design and nonresponse in the Florida BRFSS survey

Peer-reviewed author version

WATJOU, Kevin; FAES, Christel; Kirby, R; AREGAY, Mehreteab; Carroll, R & VANDENDIJCK, Yannick (2019) Spatial smoothing models to deal with the complex sampling design and nonresponse in the Florida BRFSS survey. In: Spatial and spatio-temporal epidemiology (Print), 29 , p. 59 -70.

DOI: https://doi.org/10.1016/j.sste.2019.03.001
Handle: http://hdl.handle.net/1942/30594

Spatial smoothing models to deal with the complex sampling design and nonresponse in the Florida BRFSS survey

[1], K. Watjou[1], C. Faes[1], R.S. Kirby[2], M. Aregay[3], R. Carroll[4], and Y. Vandendijck[1]

[1] *I-BioStat, Hasselt University, Diepenbeek, Belgium*

[2] *Department of Community and Family Health, College of Public Health, University of South Florida, Tampa, USA*

[3] *National Institute of Environmental Health Sciences, USA*

[4] *Novartis Pharmaceutical Corporation, East Hanover, NJ, USA*

Abstract

Public health and governmental organizations have acknowledged the importance of obtaining information of various characteristics for small areas, such as counties. Spatial smoothing models have been developed to gain reliable information on the geographical distribution of the outcome of interest. When the geographical analysis is based on survey data, two issues pose challenges: (1) the complex design of the survey and (2) the presence of missing data due to non-response. We investigate the influence of missing data and the adjustment thereof in the context of the 2013 Florida Behavioral Risk Factor Surveillance System (BRFSS) health survey. We focus on the application and comparison of the Hajek ratio estimator and two model-based approaches for estimation of the spatial trend of the prevalence of having no health insurance coverage. The model-based methods are compared using the Deviance Information Criterion which show the benefits of modeling the weights as flexibly as possible. Methods are extended towards subgroup analyses and the estimation of area-specific standardized rates, where household incomes was identified as an important factor to include in the analysis.

Keywords: BRFSS, Complex Survey Design, Hierarchical Bayesian Modeling, Imputation Model, Missing Data, Subgroup Analysis, Standardized Rate.

1 Introduction

The geographical mapping of health outcomes is important to better identify risk factors for disease and targets for health care. Lots of efforts have been made in the development of hierarchical spatial smoothing models for mapping the spatial distribution of health measures (See e.g. Elliott et al., 2001; Walter and Gotway, 2004; Lawson, 2013). These methods are typically used to obtain reliable estimates of local disease risk based on counts of observed cases within small areas, accounting for population background information such as the regional age distribution. When investigating the geographic distribution of illnesses across areas, health surveys are an indispensable source of information, but they give rise to additional challenges.

Surveys often have a complex design, resulting in differences between the population and survey distribution. Therefore, weighting of each individual (or unit) in the sample is commonly done in survey sampling (Chambers and Skinner, 2003; Groves et al., 2004). Methods that account for these weights can be subdivided into design-based, model-based and model-assisted approaches (Rao, 2011). A well-known and commonly used design-based estimator is the Horvitz-Thompson estimator (1952). This method makes use of the design weights in order to weigh each observation in the sample, where these weights are the reciprocal of the sampling probabilities. While this estimator provides reliable inferences in large samples, it is ineffective when used in a setting with sparse sample sizes (Rao, 2011). Model-based methods are essentially a prediction problem, in which the prediction of non-sampled individuals takes into account variables used in the sampling process and auxiliary variables (Brewer, 1963; Royall, 1970). However, these models could become highly complex due to the inclusion of large numbers of variables. In addition, the key variables for inclusion of individuals may not be available in public-use data sets. Alternatively, the design weights themselves could be used as a proxy for these sampling variables (Gelman, 2007; Little, 2007; Chen et al, 2017). Beaumont (2008) proposed estimators which could improve the efficiency of an estimator derived under the design-based framework by smoothing the design

weights in an appropriate model. Vandendijck et al. (2016) extended this model to account for spatial correlation. Chen et al. (2015), Mercer et al. (2014) and Vandendijck et al. (2016) compare a series of methods which incorporate the design weights within a hierarchical spatial model. A number of these models were selected for comparison within the framework of this paper. We investigate the impact of the different methods in the setting of the BRFSS survey, accounting for missingness in the data and with interest in a subgroup analysis.

Missing data occur commonly in health surveys (Little, 1982). Respondents often refuse to answer a question or only answer it in part. The effect of non-response in the modeling process is two-fold (Carpenter and Kenward, 2013). Firstly, the reduction in sample size lowers the precision of the estimates and the amount of information in general. Secondly, when the population of the people who did respond to the question of interest differs systematically from the population that did not, statistical analyses may produce biased results if the missingness is unaccounted for. Over the years, approaches have been developed in order to deal with incomplete data, depending on the type of missingness (Rubin, 1976; Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Carpenter et al., 2006; Molenberghs and Verbeke, 2006; Molenberghs and Kenward, 2007). In this paper, we assume the scenario of missing at random (MAR), i.e. when missingness does not depend on unobserved data. The procedures that cope with MAR can be grouped under weighting methods, maximum likelihood and imputation methods (Rubin, 1976). In the context of complex surveys, weighting methods have been used most commonly, as this approach encompasses the weighting of people which have responded in order to compensate for those participants which did not. Imputation methods, while broadly applicable, have some difficulties when working with the complex survey designs (Carpenter and Kenward, 2013). In this paper we investigate both weighting and imputation methods to deal with the missingness and complex survey design using data from the Behavioral Risk Factor Surveillance System (BRFSS).

2 BRFSS Data

We investigate data from the Florida State 2013 BRFSS survey. BRFSS collects data from adult U.S. citizens on their risk behaviors and health practices that may affect their general health, and

general demographic information, such as gender, age, race, income and the county in which they reside.

In this paper, we focus on studying possible geographical differences in the proportion of adults that did not have any health insurance or coverage in 2013. Although efforts have been made to improve health care for all Americans, in recent years Florida has been one of the states with the lowest insurance coverage percentage. According to the U.S. Census Bureau, Florida had an uninsured rate of more than 19.1% in 2013, which is well above the national average of 13.4% (Smith and Medalia, 2014). We investigate whether there are differences within Florida, focusing on the population aged 18 - 64. American residents aged 65 and older are generally eligible to enroll into the national Medicare health insurance program, and is therefore not considered in the population. As there might be a relationship between having insurance coverage and family income, a subgroup analysis based on the income level is conducted. Seven household income levels are used in the analysis (Table 1).

Category	Annual household income
1	$< \$10,000$
2	$\$10,000 - \$15,000$
3	$\$15,000 - \$25,000$
4	$\$25,000 - \$35,000$
5	$\$35,000 - \$50,000$
6	$\$50,000 - \$75,000$
7	$> \$75,000$

Table 1: *Categorization of the annual household income variable in the 2013 BRFSS data set*

A first complication in the analysis is the amount of incomplete data. When we restrict the results of the BRFSS data set to the aforementioned age interval, 999 (6.1%) participants respondents answered "Yes", 10709 (65.1%) answered "No" and 4749 (28.9%) answered "Not Sure/Don't know/Refused" with respect to the question "In the past 12 months was there any time when you did NOT have any health insurance or insurance coverage?". This means that the overall percentage of adults who are uninsured is around 8.5%. This percentage however, does not take into account the missingness, nor the sampling design. Figure 1 displays declines in both the proportion of individuals without insurance coverage (blue line) and the nonresponse rate as income increases (red line). The spatial distribution of the unweighted proportions of respondents who

answered ”Yes” to the “no health insurance coverage”-variable are shown in Figure 2. The color categorization in Figure 2 (and the rest of the paper) has been constructed using intervals of equal widths , with darker colors corresponding to a higher proportion of respondents with lack of health coverage. The average rate of no health coverage is 8.5% with a range of 2.8% to 13.7%. Though the rates are stable for the counties, there is some heterogeneity present in the data which will be investigated.

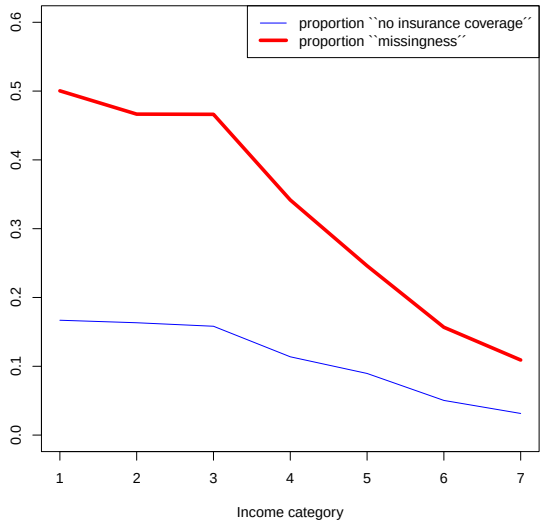

Figure 1: *Line plot showing the relation between the proportion of having no health insurance coverage and the amount of missingess at each level of the income variable*

A second complication in the analysis is that the observed sample sizes are small within the subgroups. Planned sample sizes at the county level vary between 79 (Sumter County) and 637 (Duval County) and sum up to a total of 16457 in the original data set. This resulted in actual sample sizes (complete cases) fluctuating between 59 and 473, summing up to a total of 11708. Figure 3 shows the geographic distribution of both the planned sample sizes (left panel, darker colors corresponding to larger sample size) and proportions of non-response (right panel, darker colors corresponding to more missingness).

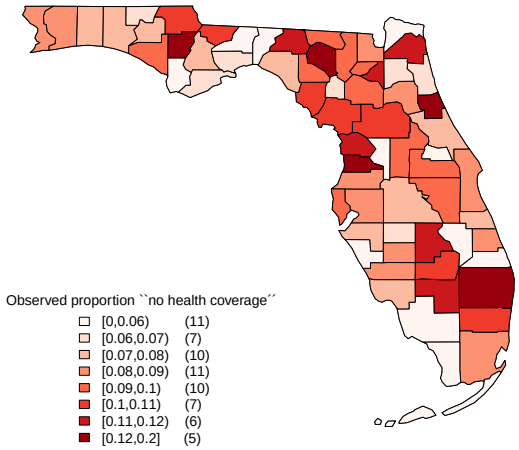

Figure 2: *Map of the observed county-specific proportions of the "no insurance coverage"-variable. The number of counties in each category is denoted between brackets.*

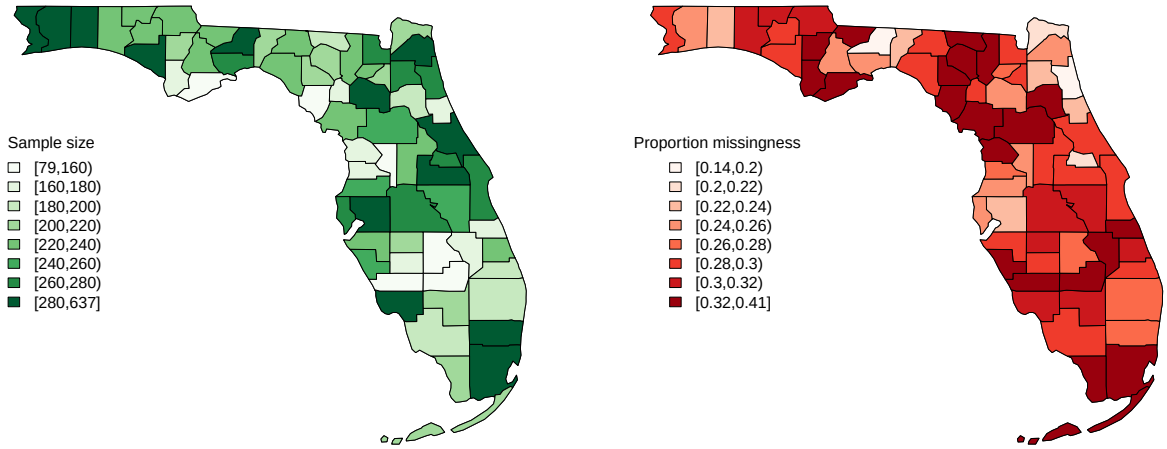

Figure 3: *Spatial distribution of the sample sizes in the original data set for the Florida counties in the (left) and the missingness proportions for the "no insurance coverage"-variable in the BRFSS study (right)*

A third complication in the analysis is the complex sampling design that was used to collect the data. In order to account for the sampling design, design weights were used in the estimation process. These design weights, `Design Wt` are constructed as a product of five components:

$$\texttt{Design Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{Nr Telephones}} \times \texttt{Nr Adults} \times \texttt{Post Strat Wt} \times \texttt{Raking Adjustment}, \quad (1)$$

where `Strat Wt` is calculated as the reciprocal of the probability of being sampled in a particular stratum, `Nr Telephones` signifies the number of land-line telephones in the selected respondent's household, `Nr Adults` represents the number of adults in the household. and `Post Strat Wt` is the post-stratification weight. Further details on the BRFSS questionnaire, sampling procedure and definition of the sampling weights for the BRFSS data set can be found at `https://www.cdc.gov/brfss/annual_data/2013/pdf/overview_2013.pdf`. No clustering was present for the variable of interest in the study population.

3 Methodology

3.1 Prevalence estimation and definition of the design weights

Denote Y_{ik} as the binary outcome variable for the i^{th} individual in county k ($i = 1,...,N_k$ and $k = 1,...,67$). We want to estimate the county-specific prevalence, defined as

$$P_k = \frac{1}{N_k} \sum_{i=1}^{N_k} Y_{ik}, \quad (2)$$

where N_k is the population size in county k for people aged 18 to 64, and $N = \sum_{k=1}^{67} N_k$ is the overall population size in the state of Florida for the specified age group. We define n_k as the planned sample size of the BRFSS sample in county k, with $n = \sum_{k=1}^{67} n_k$ denoting the total sample size ($n = 16457$). Denote r_{ik} as the binary variable, which indicates whether the variable of interest y_{ik} is observed ($r_{ik} = 1$) or is missing ($r_{ik} = 0$). Further, define s_k as the set of individuals in county

k which were sampled in the BRFSS study, with $|s_k| = n_k$, and s_k^* as the set of individuals for which the response variable was observed, where $|s_k^*| = m_k$ represents the number of respondents which answered the question of interest in county k.

In the BRFSS, each observation y_{ik} is accompanied by a design weight w_{ik}^d, for which the calculation is expressed by equation (1). Design weights need to be adjusted when some observations are missing, in order to account for the reduction in sample size and any distributional imbalance compared to the original sample and population. This can be remedied by defining a new weight w_{ik}^*, which is the product of the design weight w_{ik}^d and a missingness weight w_{ik}^m. The latter weight w_{ik}^m can be defined as the so-called inverse probability of a respondent to answer the question. We model this probability by means of a logistic regression, taking into account the characteristics which might have an influence on the missingness. We use the following bernoulli likelihood for the missingness variable r_{ik} and model:

$$r_{ik} \sim \text{Bernoulli}(P(r_{ik} = 1)) \tag{3}$$

$$\text{logit}(P(r_{ik} = 1)) = \alpha + \beta X_{ij} + u_k + v_k, \tag{4}$$

where X_{ij} is a vector containing information on the j covariates which have a significant effect on the missingness process for individual i in area k, including age, race, gender, education, marital status and income. Additionally, random effects for correlated (u_k) and uncorrelated (v_k) heterogeneity can be included to allow for spatial heterogeneity. We assume a normal distribution for the uncorrelated random effects, i.e. $v_k \sim N(0, \sigma_v^2)$, and an intrinsic conditional auto-regressive model (ICAR) (Besag et al 1991, Rue et al. 2005) for the correlated heterogeneity:

$$u_k | u_{k',k\neq k'} \sim N\left(\frac{1}{|ne(k)|}\sum_{k'\in ne(k)} u_k^{'}, \frac{\sigma_u^2}{|ne(k)|}\right), \tag{5}$$

where $ne(k)$ denotes the set of neighbors for a given county k and $|ne(k)|$ is the number of neighbors. Following common convention, we consider two counties to be neighbors if they share a common boundary.

Parameter estimation for this model can be easily conducted in a Bayesian framework. In this case, priors need to be specified on all the parameters. We choose vague priors for all parameters. For the precision parameters σ_u^{-2} and σ_v^{-2}, we assigned Gamma(0.05,0.008) priors, similar to Mercer et al. (2014) and Chen et al. (2015).

When using these weights in the estimation process, it is often the case that one uses the normalized version instead. The final weights w_{ik}^* can be normalized in such a way that they sum up to the number of non-missing observations:

$$\tilde{w}_{ik}^* = m_k \cdot \frac{w_{ik}^*}{\sum_{i \in s_k^*} w_{ik}^*}. \tag{6}$$

In Section 3.2 and 3.3 we discuss area-specific methods which incorporate these weights. In section 3.4, we review an individual-level method, as proposed by Watjou et al. (2017), and in Section 3.5 an extension of this model towards a subgroup analysis is presented, as well as an area-specific direct standardized rates based on survey data.

3.2 Model 1: Hajek ratio estimator

Since surveys are often complex by construction, the sampling design of the survey needs to be taken into account in order to obtain valid estimators. Hajek (Hajek, 1971) introducing the Hajek ratio estimator (HR) when estimating the population proportion. This method is design-based, meaning that inference is performed based on the randomization distribution of all possible samples that could have been collected from the target population.

The county-specific HR estimator can be expressed as follows:

$$\hat{P}_k^{HR} = \frac{\sum_{i \in s_k^*} \tilde{w}_{ik}^* y_{ik}}{\sum_{i \in s_k} \tilde{w}_{ik}^*} = \frac{1}{m_k} \sum_{i \in s_k^*} \tilde{w}_{ik}^* y_{ik}. \tag{7}$$

The Hajek ratio estimator is a direct estimator, as it only uses the response values from the area of interest (Rao 2003). This may have implications for the variance of the estimator, as $\hat{P}_k^{HR}$ can become unstable when the sample size in county k is sparse. In this scenario it is better to

use methods which borrow information across different counties (see Models 2-3). Note that this estimator can take into account both the design of the study and the missingness in the data via the weights.

3.3 Model 2: Arcsine Root Normal estimator

Raghunathan et al. (2007) proposed to model the arcsine-square root normal (AN) transformation of the direct Hajek ratio estimator, $y_k^{AN} = \sin^{-1}\left(\sqrt{\hat{P}_k^{HR}}\right)$, using a hierarchical model. The advantage of the transformation is that it breaks the mean-variance relationship of the prevalence and allows the variance to be stabilized approximately (Efron and Morris 1979). The likelihood is based on the following approximate normal model (Mercer et al. 2014, Chen et al. 2014):

$$\begin{aligned} y_k^{AN}|P_k &\sim N\left(\sin^{-1}\left(\sqrt{P_k}\right), \sigma_k^2\right) \\ \sin^{-1}\left(\sqrt{P_k}\right) &= \beta_0 + u_k + v_k, \end{aligned} \tag{8}$$

where the variance $\sigma_k^2 = \frac{1}{4 \cdot m_k^E}$ depends on the effective sample size $m_k^E = \hat{P}_k^{HR}(1-\hat{P}_k^{HR})/\hat{\text{var}}\left(\hat{P}_k^{HR}\right)$. The resulting estimate of the prevalence is $\hat{P}_k = (\sin(\hat{\beta}_0 + \hat{u_k} + \hat{v_k}))^2$. The correlated and uncorrelated random effect, u_k and v_k, follow a N(0, σ_v^2) and ICAR(0, σ_u^2) distribution, respectively. Note that, by including both spatially structured random effects (v_k) and unstructured random effects (u_k), this model can borrow strength across neighboring areas, in addition to possible overdispersion in the areas. When no spatial trend is present, the unstructured random effect will dominate the analysis.

Inference is conducted in the Bayesian framework. We assigned a vague N(0,1) prior to the intercept β_0 and a Gamma(0.05, 0.008) distribution to the variance parameters σ_u^{-2} and σ_v^{-2}. Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)) was used in order to perform the analysis using an accurate approximation of the posterior distribution.

3.4 Model 3: Hierarchical weight-smoothing estimator

While previous methods are weighting-based methods, the last method is an imputation-based approach, combining ideas from imputation methods in missing data with weight-smoothing methods for complex surveys. In this context we refer to imputation as the mechanism in which we 'impute' or 'predict' values of the response outcome for individuals for which we have not observed the response outcome (either because he was not sampled or because he did not respond to the question).

Similar to the ideas from Royall (1970) in the context of complex sampling designs and ideas from imputation methods for incomplete data, we first define the likelihood of the predictive model for the non-sampled and non-observed individuals:

$$y_{ik}|P_{ik} \sim \text{Bernoulli}(P_{ik}), \tag{9}$$

where P_{ik} is specified as a flexible function of all covariates that impact the design of the study and missingness indicator. This can also be referred to as the imputation model. Second, the predictive hierarchical estimator for the prevalence in area k is calculated as (Vandendijck et al, 2016)

$$\hat{P}_k = \frac{1}{\sum_{l=1}^{L_k} \tilde{N}_{lk}} \left(\sum_{l=1}^{L_k} n_{lk}\bar{y}_l + \sum_{l=1}^{L_k} \left(\tilde{N}_{lk} - n_{lk}\right) \hat{P}_{lk} \right), \tag{10}$$

where $\bar{y}_l = \frac{\sum_{i \in l} y_{ik}}{n_{lk}}$ is the sample average and $\tilde{N}_{lk}$ the estimated population size in post-stratification cell l of area k. These poststratification cells l represent the set of units which have the same normalized weight $\tilde{w}^*_{ik}$. Since these weights are unique for each unit in the BRFSS data set, the index i coincides with the index l: $\tilde{w}^*_{ik} \equiv \tilde{w}^*_{lk}$. Generally, an estimate for the prevalence P_{lk} in each post-stratification cell l and area k can be obtained from (11) using the unique normalized weights $\tilde{w}^*_{lk}$. Point estimates of this estimator $\hat{P}_{lk}$ are obtained by calculation of the posterior predictive mean, while the posterior standard deviations provide a measure of uncertainty. This estimator has shown to have good performances by Vandendijck et al. (2016) and Watjou et al. (2017).

Vandendijck et al. (2016) and Watjou et al. (2017) integrated this technique into the framework of small area estimation. Two predictive models were proposed:

$$MB_1 : \text{logit}(P_{ik}) = \beta_0 + f(\tilde{w}^*_{ik}) + u_k + v_k, \tag{11}$$

or

$$MB_2 : \text{logit}(P_{ik}) = \beta_0 + f(\tilde{\pi}^*_{ik}) + u_k + v_k, \tag{12}$$

where $f(.)$ is a non-parametric function, used for either the design weights $\tilde{w}^*_{ik}$ or the inclusion probabilities $\tilde{\pi}^*_{ik} = 1/\tilde{w}^*_{ik}$. In this paper we specify this function either by a random walk model of order one (RW1) or a penalized spline (SP). While these models have proven to work well in small surveys (Vandendijck et al. 2016, Watjou et al 2017), we want to investigate their usefulness in the context of the Florida BRFSS in this paper. In this model, the design weights incorporate both the design of the study and probability to be observed.

Watjou et al. (2017) expanded this hierarchical weight-smoothing method by modeling the design weights and the weights adjusting for non-response separately. As these two weights can be included as distinct covariates, we can isolate their respective effects as follows:

$$MB_3 : \text{logit}(P_{ik}) = \beta_0 + f_1(\tilde{w}^d_{ik}) + f_2(\tilde{w}^m_{ik}) + u_k + v_k. \tag{13}$$

In this formulation $\tilde{w}^m_{ik}$ is the normalized version of the missingness weights. The strength of this model is that it can distinguish between the design variables and the variables which influence non-response, as these two sets are not always identical.

As before, estimation is done in a Bayesian framework. Gamma(1,0.001)-distributions were specified for the precision parameters. We refer to Vandendijck et al. (2016) and Watjou et al. (2017) for additional details and specifications on this model.

3.5 Impact of subgroups

A disadvantage of the presentation of area-specific prevalence rates, is that known risk-factors (subgroups) can have an impact on the spatial trend. For example, if the outcome of interest is impacted by income, and the income distribution is different in different areas, than the difference we observe in the prevalence of the outcome of interest might be purely the effect of income. We address two approaches that can be followed: one can either focus on subgroup analyses (subgroup-specific prevalence) or work with a standardized rate. In the above example, in the subgroup analyses, the area-specific prevalence rates would be estimated for each of the income groups separately. Alternatively, a direct standardized rate can be obtained per area, accounting for the underlying income distribution. Both methods are based on the predictive models $MB_1 - MB_3$.

In a similar way as in (10), the prevalence $\widehat{P}_{gk}$ for subgroup g in area k can be estimated as

$$\widehat{P}_{gk} = \frac{1}{\widetilde{N}_{gk}} \left[\sum_{l(k)\in g} n_{lk}\bar{y}_{lk} + \sum_{l(k)\in g} (\widetilde{N}_{lk} - n_{lk})\widehat{p}_{lk} \right] = \frac{1}{\sum_{l(k)\in g} \widetilde{N}_{lk}} \left[\sum_{l(k)\in g} n_{lk}\bar{y}_{lk} + \sum_{l(k)\in g} (\widetilde{N}_{lk} - n_{lk})\widehat{p}_{lk} \right] \quad (14)$$

in which the summation only goes over the strata within the subgroup and where $\widehat{p}_{lk}$ is obtained from the assumed prediction model $MB_1 - MB_3$. The summations over $l(k) \in g$ stand for a summation of all the post-stratification cells in area k that correspond to subgroup g. As explained in Section 3.4, since the the poststratification cells are the sampled units in the BRFSS data set it holds that $y_{lk} \equiv y_{ik}$. This gives rise to subgroup-specific estimates of the prevalence, and allows to investigate whether there are differences among subgroups.

Alternatively, a direct standardized rate can be obtained by predicting the number of cases we would observe in a standard population, if the observed group-specific rates (as in area k) applied to the standard population. As standard population, we take the overall study population. The direct standardized rate for area k is given by

$$DSR_k = \frac{1}{N} \sum_g \widetilde{N}_g \widehat{P}_{gk} = \frac{1}{N} \sum_g \left(\sum_{l,k\in g} \widetilde{N}_{lk} \right) \widehat{P}_{gk} \quad (15)$$

with $\widetilde{N}_g$ is the overall estimated population in group g and $\widehat{P}_{gk}$ is the predicted prevalence in group

g, according to the risk in region k (and estimated as in equation (14)). This estimators also takes into account the design of the study. This rate can be interpreted as the area-specific risk, taking into account differences of the risk factor among areas.

The data analyses were performed on a HP Probook 650 G1. The time needed for a single data analysis varies between 1 and 3 hours, depending on the model. The simulations were done using the Flemish Super Computer, which allowed us to run simulations in parallel, reducing the computation time. All analyses were conducted in the free statistical program R, using primarily the libraries "INLA" and "survey".

4 Application to BRFSS data

4.1 Overall rate of no insurance coverage

All proposed methods are applied to the Florida BRFSS data to estimate the county-specific proportion of adults with no health insurance coverage. The top left panel in Figure 4 shows the geographical distribution of the estimated rate of no insurance coverage based on Model 1 (the HR estimator), while the top right shows the estimated rates based on Model 2 (the AN estimator). Note that darker colors correspond to a larger proportion of individuals in the population with a lack of health coverage. This map can be compared with Figure 2, showing the observed proportion of respondents with a lack of health coverage.

Note that the spatial trend of the model-assisted AN estimates is smoothed compared to the design-based HR estimator, due to the inclusion of the spatial random effect in the modeling process. Indeed, while the HR estimator is highly variable due to small sample sizes in some areas, the AN estimate better acknowledges for the heterogeneity, sharing knowledge across boundaries. Two clusters can be distinguished, exhibiting higher rates of adults with no health coverage: (*i*) in the north-central counties, close to the city of Gainesville and (*ii*) the counties surrounding Lake Okeechobee along with the neighboring counties on the east coast, such as Palm Beach and Miami-Dade (see Figure 8 and Table 4 in the Appendix).

The weight-smoothing estimator is based on a predictive model, and different underlying mod-

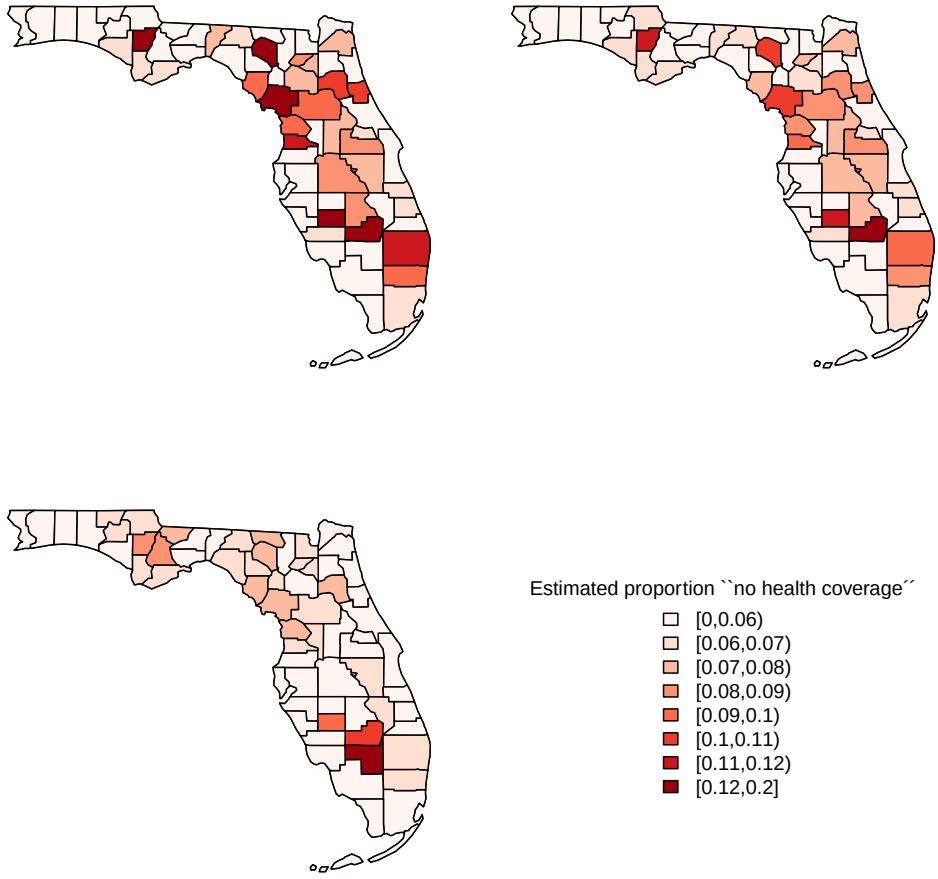

Figure 4: *Map of the estimated proportions of having no insurance coverage yielded by the HR estimator using adjusted weights (top left), the AR estimator using adjusted weights (top right) and the MB_3 model (SP) (bottom left).*

els can be suggested (such as those given by $MB_1 - MB_3$). The best model can however be selected via the Deviance Information Criterion (DIC) (Spiegelhalter, 2002). In Table 2, DIC values for the different weight-smoothing approaches are presented.

It is apparent that the methods which model the normalized design weights and missingness weights separately perform best in terms of goodness-of-fit; showing that the design characteristics and missingness characteristics have a different impact. In particular, the model which utilizes a spline specification for the design and missingness weights (MB_3), performs best in terms of DIC. Note that the same comparison with the design-based, model-assisted and model-based methods

	MB_1 (RW1)	MB_1 (SP)	MB_2 (RW1)	MB_2 (SP)
Adjusted weights	6747.50	6801.30	6683.27	6764.20
	MB_3 (RW1)	MB_3 (SP)		
Separate weights	6341.19	6328.08		

Table 2: *DIC values of the hierarchical weight-smoothing estimators for the "no insurance coverage"- variable with u_k*

cannot be made, as they each model a different response.

Results of the best fitting model, namely MB_3 (SP) is presented in Figure 4 (bottom left panel). As can be observed, further smoothing to take out uncertainty can be observed as we allow the weights to be modeled flexibly. The clusters close to the city of Gainesville and surrounding Lake Okeechobee are still visible, but are less extreme. Some other outlying counties with high or lower insurance coverage become more clearly visible based on this model. The counties Calhoun, Suwannee and Hernando, amongst others, have lower insurance coverage. The counties Nassau and Monroe, amongst others, have higher insurance coverage. Figure 5 shows the standard errors of the estimators provided in Figure 4. As expected, the standard errors of the HR estimator exhibits higher variability as compared to the AN and MB estimators. No major differences are seen for methods which use the adjusted weights or semi-adjusted weights however. The smallest variability is observed for the hierarchical weight-smoothing estimator.

Furthermore, the necessity of including a spatial random effect into the predictive model for the model-based estimators was also investigated by means of the DIC. These results can be consulted in Table 3 (in comparison to Table 2). In general it can be noticed, especially for the models which use the RW1 specification, that the inclusion of the spatial random effect does benefit the model. In the other scenarios, the benefit seems more moderate.

	MB_1 (RW1)	MB_1 (SP)	MB_2 (RW1)	MB_2 (SP)
Adjusted weights	6824.24	6812.19	6795.80	6681.91
	MB_3 (RW1)	MB_3 (SP)		
Separate weights	/	6327.28		

Table 3: *DIC values of the hierarchical weight-smoothing estimators for the "no insurance coverage"- variable without u_k*

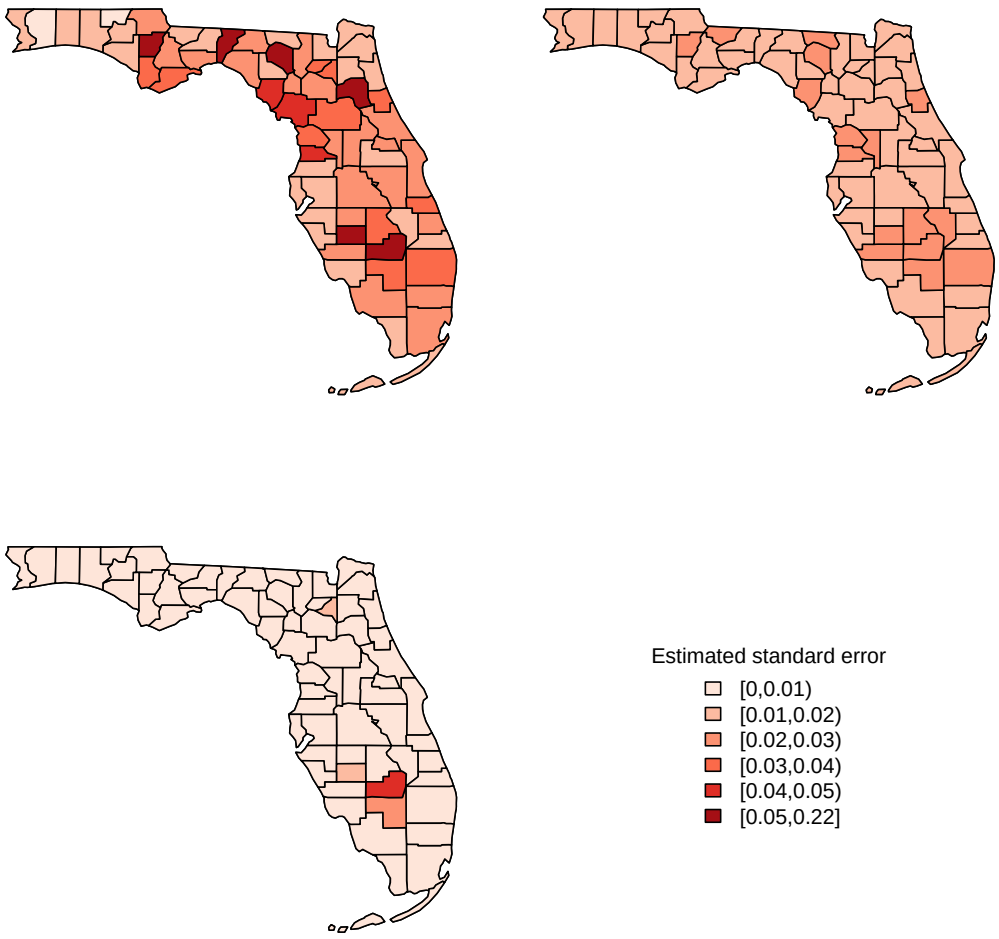

Figure 5: *Map of the estimated standard errors of the estimates for having no insurance coverage yielded by the HR estimator using adjusted weights (top left), the AR estimator using adjusted weights (top right) and the MB_3 model (SP) (bottom left).*

4.2 Rate of insurance coverage by income

In the previous analysis, focus was on the overall rate of adults with no insurance coverage. However, as the income distribution is not the same in all counties, the geographical trend could possibly be affected by this factor. Indeed, individuals with lower household incomes, have lower probability to have insurance coverage (see Figure 1). One of the strengths of the weight-smoothing models is that they are able to provide estimates for different subgroups, here for the income-specific strata.

Figure 6 illustrates the estimated proportion of having no insurance coverage at each of the seven levels of annual household income, based on model MB_3 (SP). First of all, one could observe an important trend of the rate of insurance coverage with income, as the magnitude of the estimated proportion of having no insurance coverage declines when the income level increases. This is not unexpected, since unemployed people or people with low income have more difficulties covering the rising insurance costs. Focusing on income groups 1-3, individuals which can likely be classified as 'poor', the rate of no insurance coverage varies among counties between 13% and 25%. In income groups 4-5, the rate of no insurance coverage decreases to a range between 6% and 19%. In the highest income groups, which are the 'not poor' individuals, the uninsurance rates vary between 0% and 11%.

Second, geographical differences among the counties are less visible, although there are some differences between counties for the middle income groups. Model MB_3 shows large variability between the counties for income groups 3 and 4. For these income groups, the high non-insurance coverage rate cluster around Lake Okeechobee, as observed before, is visible.

4.3 Standardized Rate of insurance coverage

While the subgroup analysis already gives a clear indication on the geographic distribution of having no health insurance coverage, this can be further investigated using the direct standardized rate. Figure 7 provides us with a map of the direct standardized rate estimates, produced by the best model (MB_3 (SP)). These are the rates of no insurance coverage if the income-specific rates would apply on the Florida population. The rates over the different counties are comparable, in the sense that they are standardized for the income distribution. As with the subgroup analysis, it could be observed that the cluster in the northern part of Florida has been smoothed out, whereas the cluster of counties with higher rates of no insurance coverage along the southeast coast is still apparent. The counties with the worst insurance coverage are Palm Beach and Hendry.

While counties such as Calhoun, Suwannee and Hernando were presented as areas of lower insurance coverage rate (based on the overall rate), it can be observed in this plot that these areas do not have an elevated standardized rate. This shows that income is the main reason for higher rates of no insurance coverage in these areas. Similarly, the counties Nassau and Monroe were

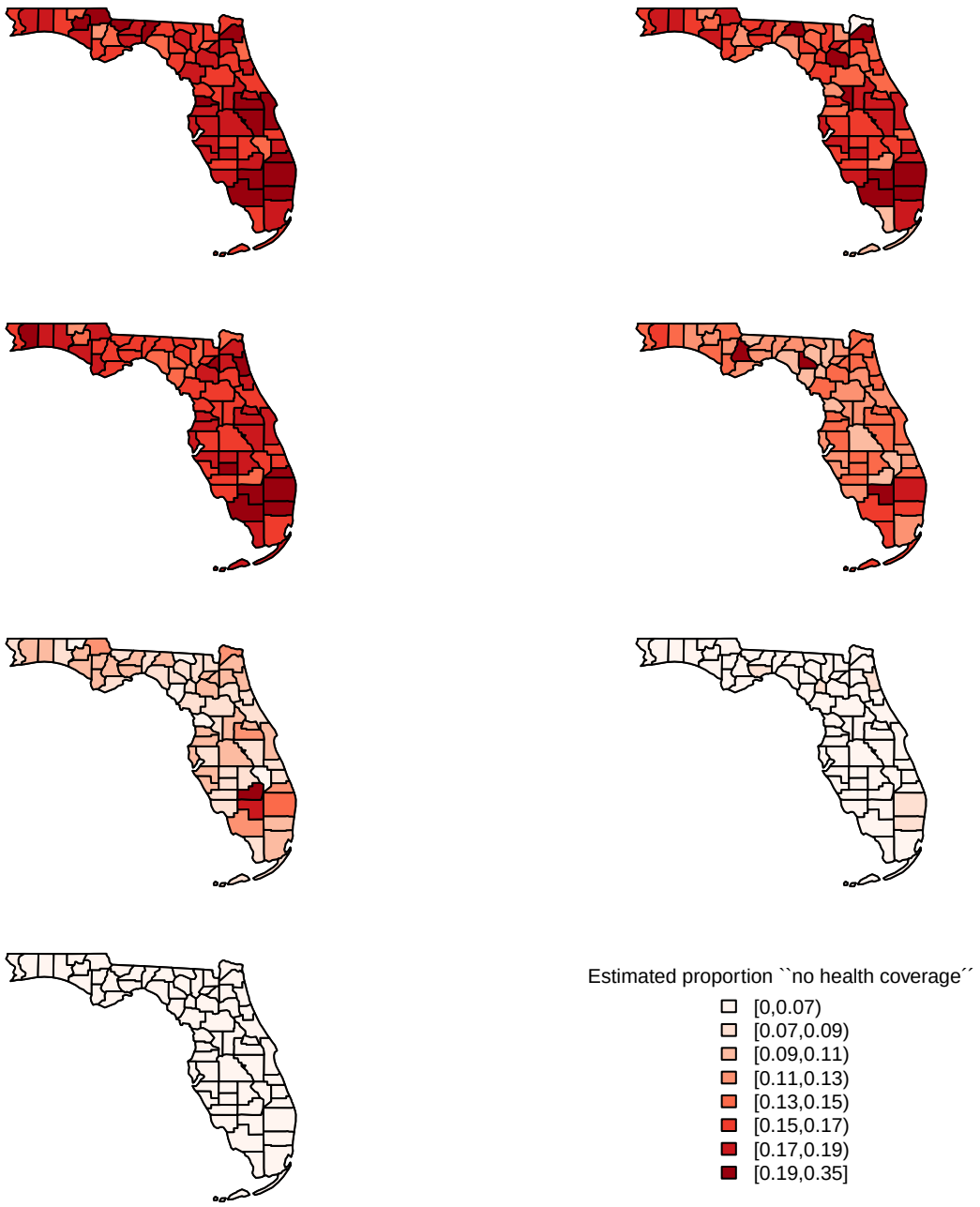

Figure 6: *Geographical distribution of the estimated prevalence of having no insurance coverage, produced by model MB_3 (SP), for Income equal to 1 till 7, going from left to right, top to bottom.*

previously categorized as counties with higher insurance coverage, but the direct standardized rate is higher for these areas as compared to other areas. This difference stresses the discrepancy between low- and high-income individuals in these counties.

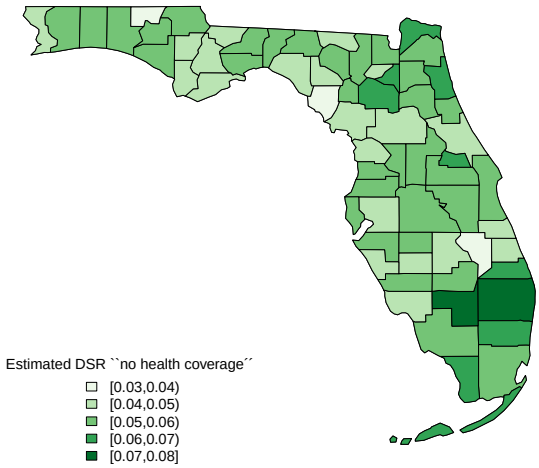

Figure 7: *Geographical distribution of the Direct Standardized Rate of having no insurance coverage, produced by model MB_3 (SP).*

5 Discussion and Conclusions

Missing data has gained a lot of attention over the last decades, resulting in a wide range of methodologies. A variety of methodologies have been developed in the context of missing data in survey samples under the missing at random (MAR) assumption, such as imputation approaches, full information direct likelihood and weighting methods (Rubin, 1987; Little and Rubin, 1987). In this paper, we adjusted for non-response by recalibrating the design weights, in order to counteract any distributional shift the missing data may have caused. When applying this to the 2013 Florida BRFSS sample, we compared the impact of this adjustment on the design-based HR estimator, the model-assisted AN estimator and several weight-smoothing models, where the latter methods modeled the design weight directly as a covariate into the model. For the county-specific prevalences of ”no insurance coverage”, this resulted in a decrease for both the HR and AN estimators. For the hierarchical weight-smoothing models, the effect of the adjustment was illustrated in a decrease of the DIC values. We showed the flexibility of the weight-smoothing model, by extending it to a subgroup analysis. Further, it allows us to calculate a directly standardized rate, to correct

for known risk factors in the analysis, and make areas comparable.

In Section 4.1, the model-based models were compared by means of the DIC values. However, the results have to be interpreted with caution since the DIC introduced by Spiegelhalter (2002) is valid under random sampling. Thus it is advisable that other model selection tools are investigated. In the context of complex survey Lumley and Scott (2015) introduced an adjustment to the classical AIC and BIC. The adjustment for DIC will be a topic for further research. Watjou et al (2017) performed a simulation study to compare the performances of the design-based, model-assisted and model-based methods described in Section 3 based on summary statistics such as the mean squared error and the bias squared. From their results it could be concluded that the model-based estimators generally perform better than their design-based and model-assisted counterparts.

The use of a flexible model is crucial in the analysis, allowing for possible non-linearity of the sampling weights. While both random-walk models and spline models allow for flexible trend, the spline model was preferred in this context. Due to the complex sampling design, a large amount of different sampling weights results, leading to instability in the estimation of this model. A possible way to mitigate these issues is to reduce the number of unique values for the weights by means of grouping the weights. As this could have an impact of the performance of the model, this should be investigated carefully.

The model-assisted AN estimator and the weight-smoothing models borrow information across neighboring counties. However, problems can arise when working with sparsely populated areas. Alternative models have been proposed to account for this. Goovaerts (2010, 2017) proposed a geostatistical model whereby rate data was estimated using poisson kriging.

A drawback of the proposed predictive model, is that the predictive model does not incorporate the uncertainty in the missingness weights. As an alternative, a spatial joint estimation model in which a measurement model is linked to a missingness model, accounting for the design aspects in the survey, is a topic of further research.

Acknowledgements

Support from the National Institutes of Health is acknowledged [award number R01CA172805]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. For the analyses we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI.

References

1. Basu, D., Godambe, V.P. and Sprott, D.A. (Ed.) (1971), *An essay on the logical foundations of survey sampling, Part One. In Foundations of Statistical Inference.* Toronto, ON: Holt, Rinehart and Winston, 202-233.

2. Beaumont, J. F. (2008), "A new approach to weighting and inference in sample surveys," *Biometrika*, 95 (3), 539-553.

3. Besag, J., York, J. and Mollié, A. (1991), "Bayesian image restoration with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, 43 (1), 1-59.

4. Brewer, K.R.W. (1963), "A Model of systematic sampling with unequal probabilities," *Australian & New-Zealand Journal of Statistics*, 5 (1), 5-13.

5. Carpenter, J.R., Kenward, M.G., and Vansteelandt, S. (2006), " A comparison of multiple imputation and doubly robust estimation for analyses with missing data," *Journal of the Royal Statistical Society, Series A*, 169, 571-584.

6. Carpenter, J. and Kenward, M. (2013), *Multiple Imputation and its Application*, Chichester: Wiley.

7. Chambers, R.L. and Skinner, C.J. (2003), *Analysis of survey data*, New York: Wiley.

8. Chen, C., Wakefield, J.C. and Lumley, T. (2015), "The use of sampling weights in Bayesian hierarchical models for small area estimation," *Spatial and Spatio-Temporal Epidemiology*, 11, 33-43.

9. Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R.J.A., Sedransk, J. and Thompson, M. (2017), "Approaches to Improving Survey-Weighted Estimates," *Statistical Science*, 32 (2), 227-248.

10. Efron, B. and Morris, C. (1979), "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70 (350), 311-319.

11. Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (2001), *Spatial Epidemiology: Methods and applications.* London: Oxford University.

12. Gelman, A. (2007), "Struggles with survey weighting and regression modeling," *Statistical Science*, 22 (2), 153-164.

13. Goovaerts, P. (2010), "Geostatistical analysis of county-level lung cancer mortality rates in the Southeastern United States," *Geographical Analysis*, 42 (1), 32-52.

14. Goovaerts, P. (2017) Geostatistical interpolation of rate data using poisson kriging. In: Shekhar S., Xiong H., Zhou X. (eds) Encyclopedia of GIS. Springer, Cham.

15. Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004), *Survey Methodology*, New York: Wiley.

16. Hajek, J. (1971), Comment on "An essay on the logical foundations of survey sampling" by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

17. Horvitz, D.G. and Thompson, D.J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47 (260), 663-685.

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

19. Little, R.J.A. (1982), "Models for nonresponse in sample surveys," *Journal of the American Statistical Association*, 77, 237-250.

20. Little, R.J.A. and Rubin, D.B. (1987), *Statistical analysis with missing data*, New York: Wiley.

21. Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data (2nd ed.)*, New York: Wiley.

22. Little, R.J.A. (2007), Comment on "Struggles with survey weighting and regression modeling," *Statistical Science*, 22 (2), 171-174.

23. Lumley, T. and Scott, A. (2015), "AIC and BIC for modeling with complex survey data," *Journal of Survey Statistics and Methodology*, 3 (1), 1-18.

24. Mercer, L., Wakefield, J.C., Chen, C. and Lumley, T. (2014), "A comparison of spatial smoothing methods for small area estimation with sampling weights," *Spatial Statistics*, 8, 69-85.

25. Molenberghs, G. and Kenward, M. (2007), *Missing Data in Clinical Studies*, Chichester: Wiley.

26. Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, New York: Springer.

27. Pfeffermann, D. (2007), Comment on "Struggles with survey weighting and regression modeling," *Statistical Science*, 22 (2), 179-183.

28. Pfeffermann, D. (2013), "New important developments in small area estimation," *Spatial Statistics*, 28, 40-68.

29. Raghunathan, T., Xie, D., Schenker, N., Parsons, V. , Davin, W., Dood, K. and Feuer, E. (2007), "Combining information from two surveys to estimate county-level prevalance rates of cancer risk factors and screening," *Journal of the American Statistical Association*, 102 (478), 474-486.

30. Rao, J.N.K. (2011), "Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal," *Statistical Science*, 36 (2), 240-256.

31. Royall, R. (1970), "On finite population sampling theory under certain linear regression models," *Biometrika*, 52 (2), 377-378.

32. Rubin, D.B. (1976), "Inference and missing data," *Biometrika*, 63 (3), 581-592.

33. Rubin, D.B. (1987), *Multiple imputation for nonresponse in surveys*, New York: Wiley.

34. Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman and Hall/CRC Press.

35. Rue, H., Martino, S. and 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 (2), 319-392.

36. Smith, J.C. and Medalia, C. (2014), Health Insurance Coverage in the United States: 2013. Current Population Reports, `http://www.nber.org/cps/hi/2014redesign/p60-250.pdf`.

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

38. Vandendijck, Y., Faes, C., Kirby, R.S., Lawson, A. and Hens, N. (2016), "Model-based inference for small area estimation with sampling weights," *Spatial Statistics*, 18, 455473.

39. Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

40. Waller, I.A. and Gotway, C.A. (2004), *Applied spatial statistics for public health data*, Hoboken, NJ: Wiley.

41. Watjou, K., Faes, C., Lawson, A., Kirby, R.S., Aregay, M., Carroll, R. and Vandendijck, Y. (2017), "Spatial small area smoothing models for handling survey data with nonresponse," *Statistics in Medicine*, 36 (23), 37083745.

Supplementary Materials

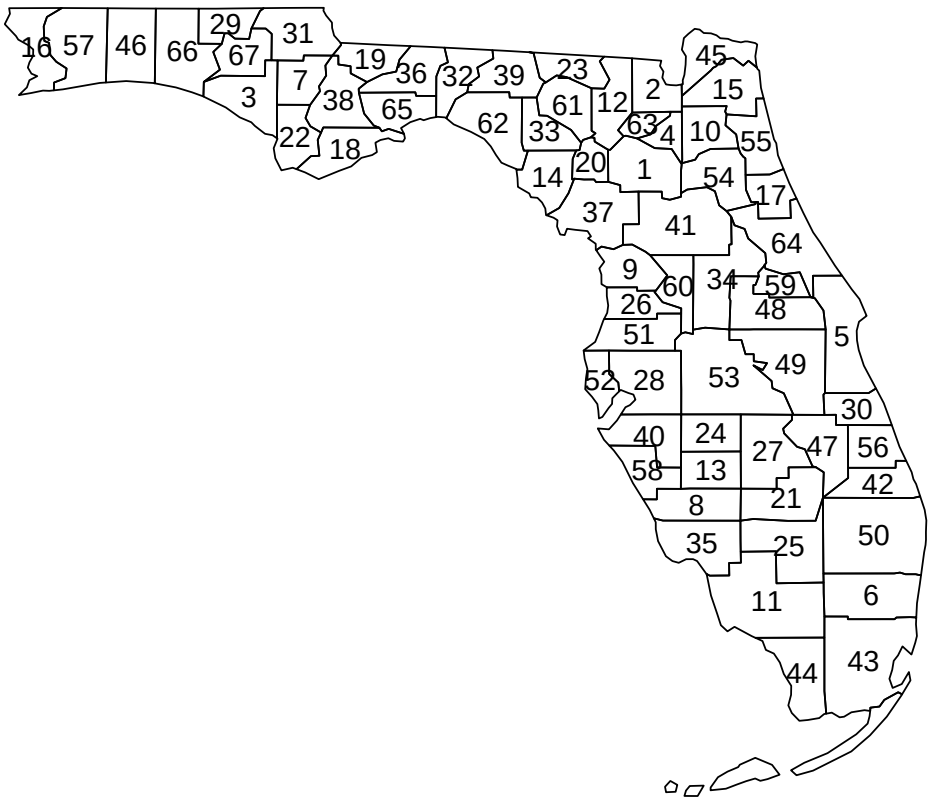

Figure 8: *Map of indices for counties in the state of Florida.*

Index	County	Index	County	Index	County	Index	County
1	Alachua	18	Franklin	35	Lee	52	Pinellas
2	Baker	19	Gadsden	36	Leon	53	Polk
3	Bay	20	Gilchrist	37	Levy	54	Putnam
4	Bradford	21	Glades	38	Liberty	55	St. Johns
5	Brevard	22	Gulf	39	Madison	56	St. Lucie
6	Broward	23	Hamilton	40	Manatee	57	Santa Rosa
7	Calhoun	24	Hardee	41	Marion	58	Sarasota
8	Charlotte	25	Hendry	42	Martin	59	Seminole
9	Citrus	26	Hernando	43	Miami-Dade	60	Sumter
10	Clay	27	Highlands	44	Monroe	61	Suwannee
11	Collier	28	Hillsborough	45	Nassau	62	Taylor
12	Columbia	29	Holmes	46	Okaloosa	63	Union
13	DeSoto	30	Indian River	47	Okeechobee	64	Volusia
14	Dixie	31	Jackson	48	Orange	65	Wakulla
15	Duval	32	Jefferson	49	Osceola	66	Walton
16	Escambia	33	Lafayette	50	Palm Beach	67	Washington
17	Flagler	34	Lake	51	Pasco		

Table 4: *Counties in the state of Florida in alphabetical order and the associated index corresponding to Figure 8* .