DATA SCIENCE JOURNAL

Chien, L-C, et al. 2018 Disparity of Imputed Data from Small Area Estimate Approaches – A Case Study on Diabetes Prevalence at the County Level in the U.S. *Data Science Journal,* 17: 8, pp. 1–11, DOI: https://doi.org/10.5334/dsj-2018-008

RESEARCH PAPER

Disparity of Imputed Data from Small Area Estimate Approaches – A Case Study on Diabetes Prevalence at the County Level in the U.S.

Lung-Chang Chien[1], Ge Lin[1], Xiao Li[2] and Xingyou Zhang[3]

[1] University of Nevada, Las Vegas, US

[2] University of Texas Health Science Center at Houston (UTHealth) School of Public Health, US

[3] U.S. Census Bureau, US

Corresponding Author: Lung-Chang Chien (lung-chang.chien@unlv.edu)

This paper assesses concordance and inconsistency among three small area estimation methods that are currently providing county-level health indicators in the United States. The three methods are multi-level logistic regression, spatial logistic regression, and spatial Poison regression, all proposed since 2010. Diabetes prevalence is estimated for each county in the continental United States from the 2012 sample of Behavioral Risk Factor Surveillance System. The mapping results show that all three methods displayed elevated diabetes prevalence in the South. While the Pearson correlation coefficients among three model-based estimates were all above 0.60, the highest one was 0.80 between the multilevel and spatial logistic methods. While point estimates are apparently different among the three small area estimate methods, their top and bottom of quintile distributions are fairly consistent based on Bangdiwala's B-statistic, suggesting that outputs from each method would support consistent policy making in terms of identifying top and bottom percent counties.

Keywords: Small area estimate; diabetes prevalence; multi-level logistic regression; spatial logistic regression; spatial Poisson regression

1. Introduction

Small area estimation (SAE) methods have been routinely used to generate poverty, employment and other economic indicators at census county and tract levels in the United States (US). Official data providers, such as the Census Bureau, and Bureau of Labor Statistics in the US approach SAE in stages: 1) proposing an appropriate SAE method, 2) evaluating and validating the proposed method, and 3) deploying the recommended method for specific SAE applications or SAE data releases. Certainly, stages 1) and 2) are often iteratively developed, and an initially proposed method may not proceed to stage 3. Such an approach had produced more than a dozen SAE methods for various small area applications (Rao and Molina 2015). Although the traditional synthetic methods coupled with direct and other indirect estimation methods are still in use, recent development point to model-based methods with or without auxiliary information as most promising in providing reliable and robust SAEs.

Model-based methods, however, can vary widely based on model-specification, auxiliary information selection, and model estimation (e.g., frequentist and Bayesian). In the last few years, there have been at least three applications of model-based SAE methods, all based on the Behavioral Risk Factor Surveillance System (BRFSS) data at the county level. A Bayesian unit-level model is currently used by the Centers for Disease Control and Prevention (CDC) to monitor changes in diabetes prevalence from 2004 onward (Cadwell et al. 2010). Then, a multi-level logistic regression model was developed to estimate chronic obstructive pulmonary disease prevalence (Zhang et al. 2014). This method was later used to produce prevalence estimates for 27 behavioral risk factors at the census tract level for the CDC-Robert Wood Johnson Foundation 500 city project (https://www.cdc.gov/500cities/about.htm). Then, a Bayesian space-time logit model was proposed to examine county level drinking patterns from 2002 to 2012 (Dwyer-Lindgren et al. 2015). An advanced estimation method was later used in various county-based risk factors and health outcome estimates, such as diabetes, cancer, cardiovascular disease and other major mortality patterns over time (Dwyer-Lindgren et al. 2016a; Dwyer-Lindgren et al. 2016b; Mokdad et al. 2017; Roth et al. 2017).

SAE methodological adoptions for major nationwide projects present several challenges, especially when the same data are used. First, by not going through the cycle of methodological development from proposing to validating a method, one risks of applying one application method as one fits for all without proper validation. The only exception in this regard is the multi-level logistic regression method that was later validated (Zhang et al. 2015). However, validating an application can be an endless endeavor; an application is appropriate for one health indicator may not necessarily appropriate for another. Second, when varied SAE methods are applied to the same health outcome at the same geographic level (e.g., county or census tract), it is not clear if one method is more appropriate to be used for the same dataset for all indicators, or different methods should be used for different indicators of the same dataset. When different SAE methods produce inconsistent county prevalence estimates, they would negatively affect decisions from local health agencies to direct limited resources to address purported health deficits. Finally, many auxiliary information used in SAE is themselves estimated by an SAE method with wide variation at the census tract and county levels. Examples include census poverty, household income variables, and many other variables from the American Community Surveys (Beaghen et al. 2012; Huang and Bell 2012). Furthermore, these variables are ever changing from time to time, and how to use auxiliary information and how they should be included in space-time SAE models have been a subject of SAE research (Rao and Molina 2015).

The current study intends to compare the three model-based methods to assess their concordance and inconsistency. A previous SAE study compared synthetic method, spatial data smoothing, and model-based regression analysis, and found that the model-based regression analysis was superior over the other two methods (Jia et al. 2004). Since the synthetic method was averaged over small area demographics, while spatial smoothing is a mechanic moving average, the superiority of the model-based regression analysis is naturally expected. In our study, we compared SAE methods that are all actively and continuously producing SAE products and scholarly publications using the BRFSS data, and they all seemed to be able to produce reliable prevalence estimates at the county level or even smaller geographic units in the US. In addition, earlier articles or recent extensions using BRFSS data at the county level can all be grouped under the three methods. For instance, the logistic regression approach used in estimating county level obesity in Mississippi (Zhang et al. 2011), and a newly proposed BRFSS-SAE method can be grouped under multi-level logistic regression (Pierannunzi et al. 2016). A two-step estimate of diabetes incidence applied essentially Cadwell's specification when it came to SAE (Barker et al. 2013). Another two-step estimate of undiagnosed diabetes is a computational improvement over the previous method (Dwyer-Lindgren et al. 2016b). We, therefore, chose the three methods as they represent current practices or the state of art of SAE using BRFSS data. Furthermore, since all three methods were applied to diabetes prevalence one way or the other, we chose diabetes as the outcomes for our comparative study.

2. Methods

2.1. Data

The study area was limited to 3,109 counties in the 48 states and the District of Columbia with a sample size of 455,406 respondents aged ≥ 18 years. The BRFSS 2012, the most recent year that the CDC released detailed county identifiers, was selected for SAEs. Respondents were regarded as having diabetes if they answered "yes" in the question: "Has a doctor, nurse, or other health professional ever told you had diabetes?" **Figure 1** presents county level diabetes samples over the study area. It shows that about 28.47% counties (n = 884) did not have any samples in the public use file. Although a large number of counties with missing samples presents challenges to SAE, they also present opportunities to compare model-based estimations for those counties. Note also that a number of counties (N = 48) had selected samples but no diagnosed diabetes. A half of counties had a sample rate (i.e., the number of selected samples divided by the number of population) less than 0.11%. Only around 1% of the counties had a sample rate higher than 1.36%.

Demographic controls include age groups (18–44, 45–64, 65+ years), race (non-Hispanic White, non-Hispanic Black, Hispanic, Hispanic, and others), and sex (male and female), all of them were used in the three methods. A total of 8, 582 respondents who did not report any of the three personal characteristics, living locations (state or county), and the diagnosis of diabetes were removed. **Table 1** shows that higher

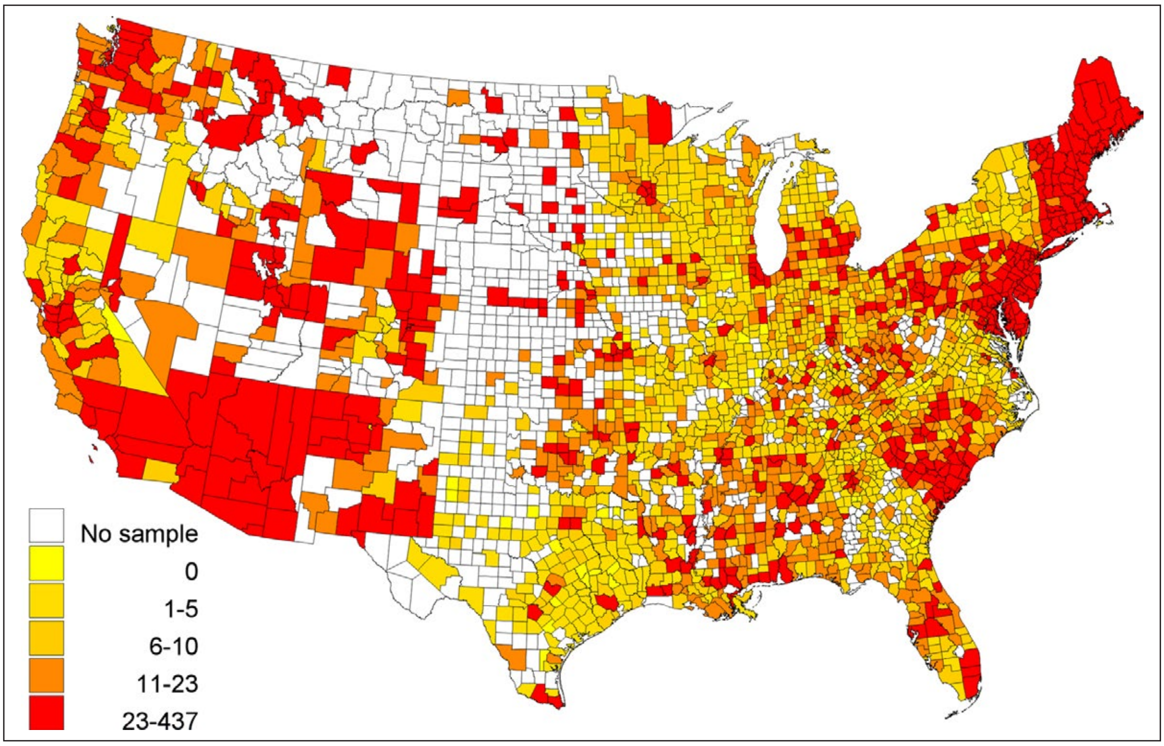

Figure 1: Geographic distribution of the number of diagnosed diabetes cases in the Behavioral Risk Factor Surveillance Survey among 3,109 U.S. counties in 2012. The color pattern was categorized by the quartiles of the number of diagnosed diabetes.

Table 1: Summary table of diabetes in the U.S.

	Diagnosed diabetes		**P-value***
	No **N (%)**	**Yes** **N (%)**	
Age			<.0001
18–45	105446 (96.65%)	3650 (3.35%)	
45–64	140475 (86.91%)	21159 (13.09%)	
65+	105983 (80.03%)	26440 (19.97%)	
Sex			<.0001
Male	140905 (86.67%)	21673 (13.33%)	
Female	214109 (87.72%)	29962 (12.28%)	
Race			<.0001
Non-Hispanic White	282189 (88.27%)	37485 (11.73%)	
Non-Hispanic Black	29536 (80.34%)	7227 (19.66%)	
Hispanic	22269 (86.33%)	3525 (13.67%)	
Others	16710 (86.55%)	2596 (13.45%)	

* Chi-square test, where 7519 missing values are excluded.

proportions of diagnosed diabetes were observed in elderly aged 65+ (19.97%), males (13.33%), and non-Hispanic Blacks (19.66%), with all the chi-square tests being significant at p-values < 0.0001.

County poverty rate, defined as percent of people living under the 100% of the federal poverty line was included as a county-level auxiliary variable. It was based on the 5-year estimate, according to American

Community Survey 2012 (Zhang et al. 2014). The average of the poverty percentage among 3,109 counties is 11.97% (standard deviation [SD] = 5.53).

2.2. Model specification

In the model specification process, we made sure that all models were specified as close as possible to the original SAE models. The multilevel logistic regression model, which is labeled Model 1, was specified identical to original model using the original SAS codes provided by the author (Zhang et al. 2014).

$$logit[P(Y_{isc}=1)]=\alpha+\beta_1(age)_i+\beta_2(race)_i+\beta_3(sex)_i+\beta_4(poverty)_c+\gamma_s+\tau_c, \qquad (1)$$

where Y_{isc} is a binary outcome variable of having diabetes (1 = yes, 0 = no) for individual i living in state s and county c. On the right hand side: α is an intercept, and (β_1, β_2, β_3, β_4) are slopes for fixed effects of age, race, sex, and poverty, respectively. The last two terms are random effects γ_s for state s and τ_c for county c.

In the absence of time effect, Model 2 can be specified by dropping time effect from the space-time logistic model in Dwyer-Lindgren et al. (2015), for which the original R codes were also provided:

$$logit[P(Y_{ic}=1)]=\alpha+\beta_1(age)_i+\beta_2(race)_i+\beta_3(sex)_i+\beta_4(poverty)_c+\tau_c+f_{spat}(c), \qquad (2)$$

which has a similar model construct to Model 1 without state random effect. In particular, county effect was attributed to a spatial uncorrelated random effect τ_c and a spatial function $f_{spat}(c)$, which is Markov random fields following an intrinsic conditional autoregressive prior (Kindermann and Snell, 1980). Due to the complexity of estimating two county effects, we applied the integrated Laplace approximation (INLA) to accelerate the model fitting, and improve the possibility of algorithm convergence (Rue et al. 2009).

Model 3 is a Bayesian Poisson model, which assumes that survey data are sampled from the complete population data (Cadwell et al. 2010). It estimates Y_{ijkc} as the number of diagnosed diabetes cases at age group i, race j, sex k in county c, which follows a Poisson distribution with a mean of μ_{ijkc}. Thus, Model 3 is specified:

$$\log(\mu_{ijkc})=\alpha+\beta_{1i}+\beta_{2j}+\beta_{3k}+\beta_4(poverty)_c+f_{spat}(c)+\log(n_{ijkc}), \qquad (3)$$

where β_{1i}, β_{2j}, β_{3k} are for age, race, and sex effects, respectively. The spatial function $f_{spat}(c)$ is still Markov random fields as in Model 2. The last term log (n_{ijkc}) is an offset corresponding to the logarithm of the at-risk population index by i, j, k, c. Similar to Model 2, we applied the INLA to estimate unknown parameters in Model 3. Note that the original model proposed by Cadwell et al. (2010) did not include the poverty as a confounding variable. To be consistent with the specifications of the first two models, we opted to include the poverty variable in Model 3.

2.3. SAE for diabetes area

To generate county-level diabetes prevalence from Model 1, we obtained estimated coefficients for four fixed effects (age, sex, race and poverty) and 3,157 random effects (48 states plus 3,109 counties). In particular, counties without samples had a county-level random effect (τ_c) imputed by the average of adjacent counties' random effects (Zhang et al. 2014). Hence, the probability of a diabetic person is:

$$P(Y_{isc}=1|age,race,sex,s,c,p)=\frac{\exp(\hat{\alpha}+\hat{\beta}_1(age)_i+\hat{\beta}_2(race)_i+\hat{\beta}_3(sex)_i+\hat{\beta}_4(poverty)_c+\hat{\gamma}_s+\hat{\tau}_c)}{1+\exp(\hat{\alpha}+\hat{\beta}_1(age)_i+\hat{\beta}_2(race)_i+\hat{\beta}_3(sex)_i+\hat{\beta}_4(poverty)_c+\hat{\gamma}_s+\hat{\tau}_c)}. \qquad (4)$$

The number of diabetes cases in each county can then be summed up by age, race and sex. After dividing by state-county specific population, we can obtain the SAE of diabetes prevalence straightforwardly:

$$p_{sc}=\frac{\sum_{age}\sum_{race}\sum_{sex}P(Y_{isc}=1|age,race,sex,s,c)\times(Pop|age,race,sex,s,c)}{\sum_{age}\sum_{race}\sum_{sex}(Pop|age,race,sex,s,c)}. \qquad (5)$$

To generate county-level diabetes prevalence from Model 2, we first calculated the probability of a person with diagnosed diabetes similar to Eq. (5) by replacing state random effect $\hat{\gamma}_s$ term with the estimated spatial function $\hat{f}_{spat}(c)$:

$$P\left(Y_{ic}=1 \mid age, race, sex, c, p\right)=\frac{\exp(\hat{\alpha}+\hat{\beta}_1(age)_i+\hat{\beta}_2(race)_i+\hat{\beta}_3(sex)_i+\hat{\beta}_4(poverty)_c+\hat{\tau}_c+\hat{f}_{spat}(c))}{1+\exp(\hat{\alpha}+\hat{\beta}_1(age)_i+\hat{\beta}_2(race)_i+\hat{\beta}_3(sex)_i+\hat{\beta}_4(poverty)_c+\hat{\tau}_c+\hat{f}_{spat}(c))}. \tag{6}$$

We then used the probability to calculate the SAE of diabetes prevalence:

$$p_c=\frac{\sum_{age}\sum_{race}\sum_{sex} P\left(Y_{ic}=1 \mid age, race, sex, c\right)\times(Pop \mid age, race, sex, c)}{\sum_{age}\sum_{race}\sum_{sex}(Pop \mid age, race, sex, c)}. \tag{7}$$

The approach to generating SAEs for Model 3 differs from Models 1 and 2. Note that we defined N_{ijkc} and Y_{ijkc} in Model 3 as age-race-sex-county specific at-risk population, and those with diabetes, respectively. Thus, we can derive Z_{ijkc}, the number of unobserved people with diagnosed diabetes, indexed by age, race, sex and county straightforwardly. Let P_c be the SAE of diabetes prevalence in county *c*, then it is the sum of the observed and unobserved, or $Y_{ijkc}+Z_{ijkc}$:

$$p_c=E\left(p_c \mid Y,n,N\right)=E\left(\frac{\sum_i\sum_j\sum_k(Z_{ijkc}+Y_{ijkc})}{\sum_i\sum_j\sum_k N_{ijkc}} \mid Y,n,N\right)=\frac{\sum_i\sum_j\sum_k\left[E(Z_{ijkc} \mid Y,n,N)+Y_{ijkc}\right]}{\sum_i\sum_j\sum_k N_{ijkc}}, \tag{8}$$

where $Z_{ijkc} \mid Y, n, N \sim \mathrm{POI}(\gamma_{ijkc})$, and the parameter γ_{ijkc} is defined as:

$$\gamma_{ijkc}=\left(\frac{\mu_{ijkc}}{n_{ijkc}}\right)\times\left(N_{ijkc}-n_{ijkc}\right)=\exp(\hat{\alpha}+\hat{\beta}_{1i}+\hat{\beta}_{2j}+\hat{\beta}_{3k}+\hat{\beta}_4(poverty)_c+\hat{f}_{spat}(c))\times(N_{ijkc}-n_{ijkc}). \tag{9}$$

Because the parameter γ_{ijkc} is exactly the expected value of a Poisson distribution, the SAE of diabetes prevalence based on Model 3 is:

$$p_c=\frac{\sum_i\sum_j\sum_k[\gamma_{ijkc}+Y_{ijkc}]}{\sum_i\sum_j\sum_k N_{ijkc}}. \tag{10}$$

2.4. Analysis of concordance and inconsistency

We first used county maps to provide visual descriptions of diabetes estimates by the three methods. The intention here is to see if the three methods provide consistent geographic patterns regardless of specific county prevalence rates. Moreover, we categorized the SAEs into five quintiles (top 20%, upper-middle 20%, middle 20%, lower-middle 20% and bottom 20% of observations) in each model, and calculated Bangdiwala's B-statistic (also known as Cohen's Kappa coefficients) to evaluate the proportion of counties in the same quintile categories (Bangdiwala and Shankar 2013). The observer agreement charts were also provided to visualize the proportion of concordance in each quintile between two models.

Model 1 was analyzed by the PROC GLIMMIX procedure in SAS v9.13. Model 2 & 3 was analyzed by the *inla* package in R software v3.13 (R Core Team). The agreement analysis was accomplished by the PROC FREQ procedure in SAS v9.13. Parameters were determined significance by a type I error of 0.05. In the preliminary analysis, we used both weighted and unweighted samples. Although both provided very close estimates, the average from the unweighted was closer to the national weighted average. The latter result was also reported in a validation study of multilevel logistic SAE (Zhang et al. 2015). For this reason, the unweighted sample was used for the estimations of three models.

3. Results

Results from SAEs of diabetes prevalence are shown in **Figure 2** in 5 quantiles. In general all three models captured the elevated prevalence in the South; all three models show reduced prevalence in West North Central and some northern Mountain states. Models 1 and 2 generated broadly similar geographic distributions, especially for counties with missing samples (refer to **Figure 1**). Models 2 and 3 had some similar geographic patterning when sample sizes for diabetes were relatively large (e.g., >24). The scatter plots in **Figure 3** reconfirmed those observations that Models 1 and 2 had closer SAEs than those compared to Model 3. The

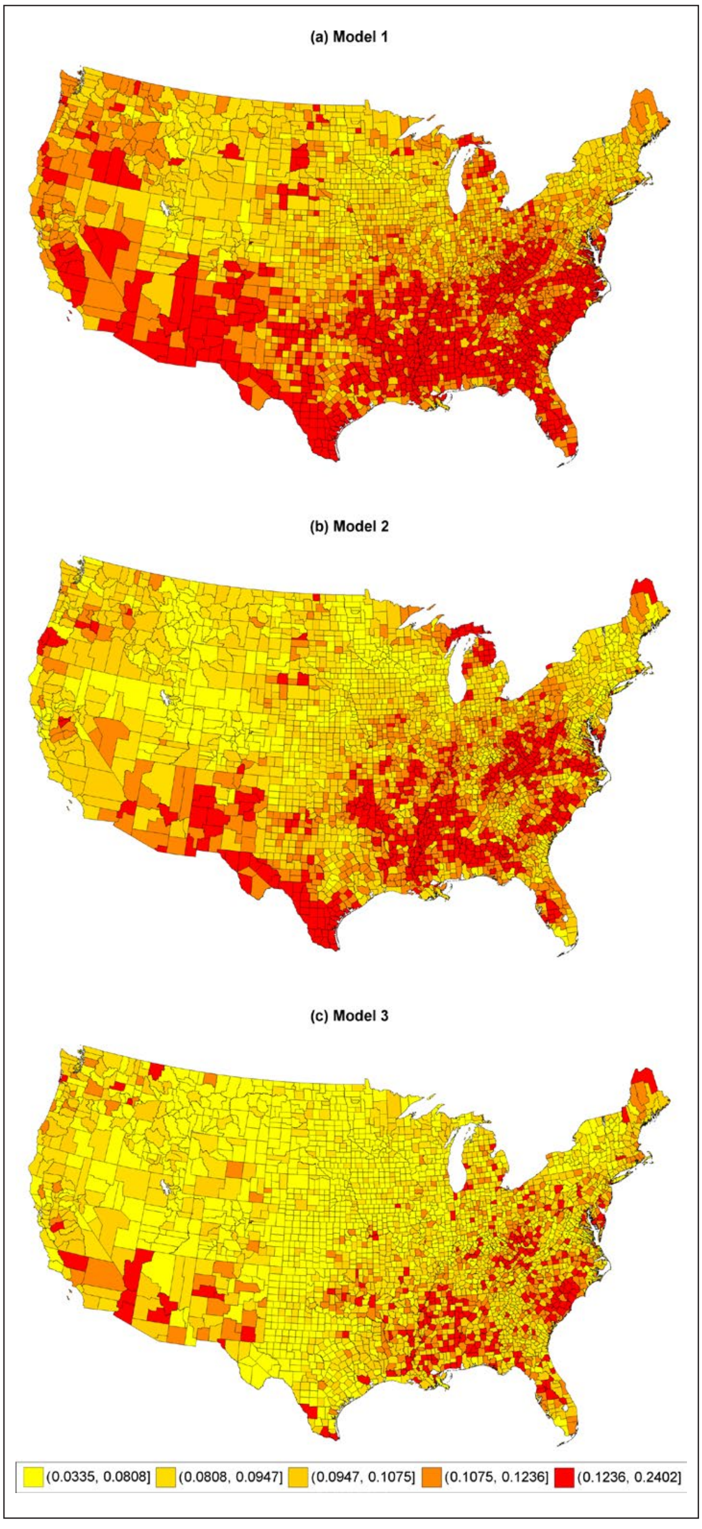

Figure 2: Comparing three SAEs of diabetes prevalence at the county level using quintiles.

correlation coefficient of SAEs between Model 1 and 2 was 0.85, while it dropped to 0.61 between Models 1 and 3. Likewise, the coefficient between Model 2 and 3 was 0.60. Even Model 1 and 2 estimates are broadly similar, we would not consider them close given the correlation coefficient was less than 0.90.

The weighted diabetes prevalence from the sample of the current study was 10.3%. Model 2 provides closest average (0.1042) among 3,109 counties (**Table 2**), while Model 1 had the highest average (0.1152). In terms of spread measures, Model 2 had the highest SD (0.0238), Model 1 resulted in the longest range by 0.1894 (0.0508, 0.2402), and Model 2 resulted in the longest interquartile range by 0.0323 (0.0870, 0.1193). Model 3 had the smallest average, SD, range, and interquartile range. Similar patterns appeared in the 2,225 counties with samples in the BRFSS. Among the other 884 counties without samples, Model 3 still had the smallest average of SAEs, and the SD became only a half compared to the SD of Model 3 in the counties with samples (0.0105 vs. 0.0226).

Table 3 shows the post-hoc comparisons of diabetes prevalence estimate, resulting in the largest mean difference in Model 1 by 0.0250 (95% CI = 0.0237, 0.0264) compared to Model 3 among 3,109 counties. For counties with samples, Model 1 still had the largest mean difference compared to Model 3, but the difference decreased to 0.0186 (95% CI = 0.0170, 0.0202). In particular, Model 2 and Model 3 had the least significantly difference by only 0.0073 (95% CI = 0.0057, 0.0089). For counties without samples in the BRFSS, Model 1 still had the largest difference by 0.0413 (95% CI = 0.0388, 0.0438) compared to Model 3, and this difference was over 2-fold of the difference between the two models among counties with samples in the BRFSS. A similar scenario occurs in the comparison between Model 2 and Model 3. Nonetheless, the difference between Model 1 and Model 2 had a tiny change between counties with and without samples in the BRFSS (0.0122 vs. 0.0104). ANOVA results in significant differences in the SAE of diabetes prevalence among the three models, regardless of presence and absence of samples in the BRFSS.

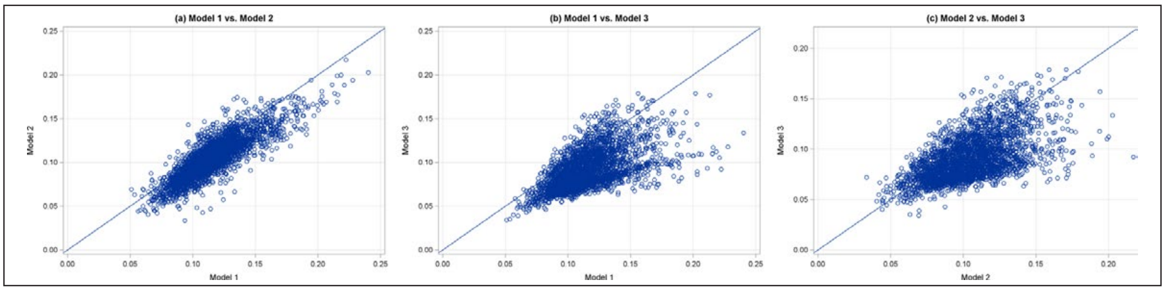

Figure 3: Scatter plots of small area estimates among Models 1, 2 and 3.

Table 2: Descriptive statistics of diabetes prevalence estimates from the three SAE models.

Model	Mean	SD	Minimum	Q1	Median	Q3	Maximum	F	P-value†
				All counties (N = 3,109)					
1	0.1152	0.0237	0.0508	0.0986	0.1121	0.1282	0.2402	910.22	<.0001
2	0.1042	0.0238	0.0335	0.0870	0.1024	0.1193	0.2171		
3	0.0902	0.0220	0.0342	0.0373	0.0855	0.1023	0.1789		
				Counties with samples in the BRFSS (N = 2,225)					
1	0.1146	0.0222	0.0536	0.0994	0.1124	0.1272	0.2210	377.93	<.0001
2	0.1034	0.0233	0.0405	0.0868	0.1024	0.1187	0.1940		
3	0.0960	0.0226	0.0351	0.0800	0.0931	0.1097	0.1789		
				Counties without samples in the BRFSS (N = 884)					
1	0.1167	0.0271	0.0510	0.0969	0.1105	0.1311	0.2402	831.05	<.0001
2	0.1063	0.0250	0.0335	0.0875	0.1023	0.1208	0.2171		
3	0.0755	0.0105	0.0342	0.0687	0.0736	0.0806	0.1337		

Abbreviation: SD = Standard deviation; Q1 = The first quartile; Q3 = The third quartile † The p-values were calculated from the analysis of variation.

Table 3: Mean difference comparison in the SAE of diabetes prevalence among Models 1, 2 and 3.

Comparison	Difference	95% CI
All counties (N = 3,109)		
Model 1 vs. Model 2	0.0110	(0.0096, 0.0124)
Model 1 vs. Model 3	0.0250	(0.0237, 0.0264)
Model 2 vs. Model 3	0.0140	(0.0127, 0.0154)
Counties with samples in the BRFSS (N = 2,225)		
Model 1 vs. Model 2	0.0112	(0.0096, 0.0128)
Model 1 vs. Model 3	0.0186	(0.0170, 0.0202)
Model 2 vs. Model 3	0.0073	(0.0057, 0.0089)
Counties without samples in the BRFSS (N = 884)		
Model 1 vs. Model 2	0.0104	(0.0079, 0.0129)
Model 1 vs. Model 3	0.0413	(0.0388, 0.0438)
Model 2 vs. Model 3	0.0309	(0.0284, 0.0333)

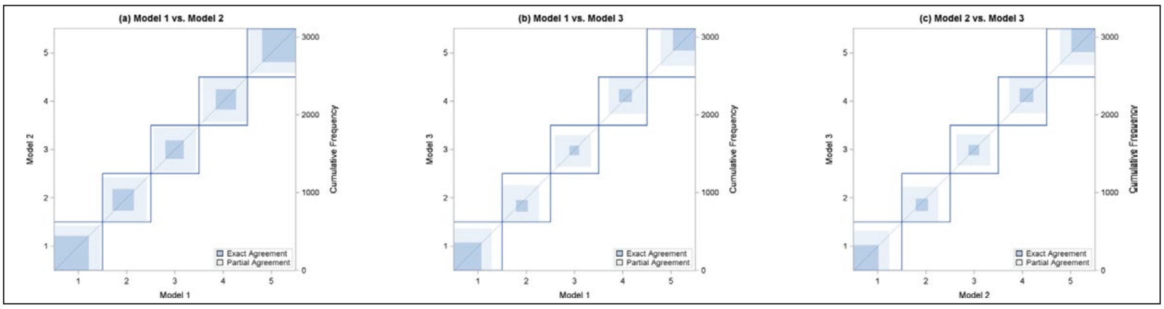

Figure 4: The observer agreement charts of categorized small area estimates among Models 1, 2 and 3.

The concordance analysis intends to show similarity in diagonals or closer to diagonals in a contingency table of 5 by 5 quantiles. A greater concordance is shown in both tails than in the middle quantiles. The SAEs among the three models had a smaller proportion of concordance in the second, third, and fourth sections (see in **Figure 4** in the 20th percentile to the 80th percentile in terms of agreement and near agreement areas). **Figure 4(a)** has the highest Bangdiwala's B-statistic by 0.3800, indicating that 38% of the 3,109 counties had the SAEs of diabetes prevalence in the same quintile sections from Model 1 and 2. However, the proportions decreased to 17.87% from Model 1 and Model 3, and 17.61% from Model 2 and Model 3 (see light blue and white areas in **Figure 4(b)** and **(c)**) indicating more inconsistent SAEs between Model 1/2 and Model 3. Since the top and bottom quantiles had little dis-concordance, using each method alone for small area estimate comparisons toward means would not generate considerable inconsistencies as far as counties with greatly above or below the national mean (e.g., 20%) are concerned.

4. Discussion

Currently, various SAE methods are used by federal agencies and research institutions, and it is difficult to gauge their estimation performance. In this paper, we have compared SAEs from thee model-based methods that are all actively producing SAEs at county or smaller area units. Our comparisons focused on spatial patterns or relative measures (e.g., quintiles) generated from each method, rather than point estimates. Overall, the three methods were able to point to the elevated diabetes prevalence observed in southern states. In addition, both top and bottom quintiles categories had highest concordance, which is of less concern, because either top or bottom of quantiles are matter the most in terms of policy making. In other words, when a county is high in diabetes prevalence compared to the national average, it less likely goes wrong if another SAE method is used.

Partly due to data limitation from the public use BRFSS file, we had 884 counties without identifiers in our SAEs. Since spatial Poisson regression in Model 3 depends on samples in spatial units, missing county samples would cause more problems for SAEs. Indeed, separate comparisons, both in map displays and post-hoc analyses, showed substantial discrepancies in counties with missing identifiers, and in general, they were under estimates compared to Model 1 and 2 estimates. Even though requesting unsuppressed BRFSS with all county identifiers is possible through each state (Li and Lin 2014), the problem remains for many small counties not to be sampled in the BRFSS.

Model-based SAE methods often include small area auxiliary information, and it tends to improve model predictability. However, different ways of using small area auxiliary variables present challenges for model comparisons. It also produces inconsistent model parameter estimates in space-time models, as both importance and reliance of auxiliary information change over the time dimension. Since socioeconomic status (SES) relates to many health outcomes, it is expected to be included in model-based SAEs in some form. In the original articles, only Model 1 used poverty, and Model 2 used education level, while model 3 did not use any SES variables. In the current study, we included the *poverty* variable for all three models to aid comparisons. Even we are not sure if poverty or educational level is more appropriate as a proxy for SES, we are pretty certain that not including poverty in model 3 would lead its SAEs to be far more inconstant from those generated from Models 1 and 2 than we currently observed. Future work should refine census based auxiliary variables for SAE. Studies should also expand to include potential administrative records, such as hospital discharge data, visits to physician clinics to either assist or produce separate SAEs, while guarding confidentiality.

Nationwide SAE of BRFSS at the county level can be calibrated to reflect state level estimates. In the absence of the full geographic sample, and to be true the three SAE model, we opted not to calibrate their estimates. Perhaps, for this reason, the three SAEs provided quite different national averages in diabetes prevalence, with Model 1 being the highest and Model 3 being the lowest. That is also why we placed less emphasis on comparing point estimates from the three methods. In real world practice, it might be preferable to calibrate the sample for each state to ensure each state average from SAEs match the overall state prevalence without SAE (Mohl 1996). Future work should compare point estimates of SAE methods and how sample sizes or population size of small areas would after the confidence intervals. It is especially pertinent when considering that auxiliary information tends to be estimates with smaller population areas having a greater margin of errors, suggesting varied uncertainties when conduct cross regional comparative SAE studies for a particular disease or health outcomes.

To some degree, all three methods considered spatial effects. The multilevel method uses the state-county hierarchy to account for some spatial effects while Models 2 and 3, use Markov random fields to work through geographically connected boundaries among all counties. While Model 1 does not consider local variation, models 2 and 3 were unable to incorporate geographic entities that completely separately or do not have local connectivity to other entities, such as Alaska or Hawaii. For this reason, we were unable to include Alaska and Hawaii due to their geographic separation. In addition, none of the three SAE methods considered spatial clustering. However, we know when data are fairly complete (e.g., births and deaths), model-based estimates would be substantially biased when spatial clustering or spatial association effects were not removed (Lin and Zhang 2012). Future studies should examine the effects of spatial clusters or clustering effects on SAEs, and remedies to reduce potential biases. Some of clustering effects could simply be due to spatial unit mismatches, in which case, we could address them through MAUP methods already developed.

5. Conclusion

All three methods were able to display elevated county-level diabetes prevalence in the South. While their point estimates were very highly correlated, the highest coloration was between the multilevel and spatial logistic methods (r = 0.86), suggesting much higher consistency compared to the spatial Poisson regression method. While there are apparent differences in point estimates among the three SAE methods, their top and bottom 20 percent distributions are fairly consistent. Each method outputs would support consistent policy making in terms of top and bottom percent counties for diabetes prevalence.

Acknowledgement

We thank for the consulting support from Institute of Health Metrics and Evaluation at the University of Washington.

Competing Interests

The authors have no competing interests to declare.

Disclaimer

The views expressed on statistical issues in this paper are those of the authors and not necessarily those of Economic Research Service, U.S. Department of Agriculture.

References

Bangdiwala, S I and **Shankar, V** 2013 The agreement chart. *BMC Med Res Methodol* 13: 97–97. DOI: https://doi.org/10.1186/1471-2288-13-97

Barker, L E, Thompson, T J, Kirtland, K A, Boyle, J P, Geiss, L S, McCauley, M M and **Albright, A L** 2013 Bayesian Small Area Estimates of Diabetes Incidence by United States County, 2009. *J Data Sci* 11: 269–280.

Beaghen, M, McElroy, T, Weidman, L, Asiala, M and **Navarro, A** 2012 Interpretation and use of American Community Survey multiyear estimates. *Statistics*: 03.

Cadwell, B L, Thompson, T J, Boyle, J P and **Barker, L E** 2010 Bayesian small area estimates of diabetes prevalence by US county, 2005. *J Data Sci* 8: 173–188.

Dwyer-Lindgren, L, Bertozzi-Villa, A, Stubbs, R W, Morozoff, C, Kutz, M J, Huynh, C, Barber, R M, Shackelford, K A, Mackenbach, J P, van Lenthe, F J, Flaxman, A D, Naghavi, M, Mokdad, A H and **Murray, C J** 2016a US County-Level Trends in Mortality Rates for Major Causes of Death, 1980–2014. *JAMA* 316: 2385–2401. DOI: https://doi.org/10.1001/jama.2016.13645

Dwyer-Lindgren, L, Flaxman, A D, Ng, M, Hansen, G M, Murray, C J and **Mokdad, A H** 2015 Drinking Patterns in US Counties From 2002 to 2012. *Am J Public Health* 105: 1120–1127. DOI: https://doi.org/10.2105/AJPH.2014.302313

Dwyer-Lindgren, L, Mackenbach, J P, van Lenthe, F J, Flaxman, A D and **Mokdad, A H** 2016b Diagnosed and Undiagnosed Diabetes Prevalence by County in the U.S., 1999–2012. *Diabetes care* 39: 1556–1562. DOI: https://doi.org/10.2337/dc16-0678

Huang, E T and **Bell, W R** 2012 An Empirical Study on Using Previous American Community Survey Data Versus Census 2000 Data in SAIPE Models for Poverty Estimates. *Statistics*: 04.

Jia, H, Muennig, P and **Borawski, E** 2004 Comparison of small-area analysis techniques for estimating county-level outcomes. *Am J Prev Med* 26: 453–460. DOI: https://doi.org/10.1016/j.amepre.2004.02.004

Kindermann, R and **Snell, J L** 1980 *Markov random fields and their applications,* Providence, R.I.: American Mathematical Society. DOI: https://doi.org/10.1090/conm/001

Li, T and **Lin, G** 2014 Examining the role of location-specific associations between ambient air pollutants and adult asthma in the United States. *Health & Place* 25: 26–33. DOI: https://doi.org/10.1016/j.healthplace.2013.10.007

Lin, G and **Zhang, T** 2012 Examining Extreme Weather Effects on Birth Weight From the Individual Effect to Spatiotemporal Aggregation Effects. *J Agric Biol Environ Stat* 17: 490–507. DOI: https://doi.org/10.1007/s13253-012-0102-1

Mohl, S 1996 Understanding calibration estimators in survey sampling. *Surv Methodol* 22(2).

Mokdad, A H, Dwyer-Lindgren, L, Fitzmaurice, C, Stubbs, R W, Bertozzi-Villa, A, Morozoff, C, Charara, R, Allen, C, Naghavi, M and **Murray, C J** 2017 Trends and Patterns of Disparities in Cancer Mortality Among US Counties, 1980–2014. *JAMA* 317: 388–406. DOI: https://doi.org/10.1001/jama.2016.20324

Pierannunzi, C, Xu, F, Wallace, R C, Garvin, W, Greenlund, K J, Bartoli, W, Ford, D, Eke, P and **Town, G M** 2016 A Methodological Approach to Small Area Estimation for the Behavioral Risk Factor Surveillance System. *Prev Chronic Dis* 13: E91. DOI: https://doi.org/10.5888/pcd13.150480

Rao, J N K and **Molina, I** 2015. *Small area estimation.* 2nd edition. ed. Wiley Series in Survey Methodology. John Wiley & Sons, Inc., Hoboken, New Jersey. DOI: https://doi.org/10.1002/9781118735855

Roth, G A, Dwyer-Lindgren, L, Bertozzi-Villa, A, Stubbs, R W, Morozoff, C, Naghavi, M, Mokdad, A H and **Murray, C J L** 2017 Trends and Patterns of Geographic Variation in Cardiovascular Mortality Among US Counties, 1980–2014. *JAMA* 317: 1976–1992. DOI: https://doi.org/10.1001/jama.2017.4150

Rue, H, Martino, S and **Chopin, N** 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Series B* 71: 319–392. DOI: https://doi.org/10.1111/j.1467-9868.2008.00700.x

Zhang, X, Holt, J B, Lu, H, Wheaton, A G, Ford, E S, Greenlund, K J and **Croft, J B** 2014 Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am J Epidemiol* 179: 1025–1033. DOI: https://doi.org/10.1093/aje/kwu018

Zhang, X, Holt, J B, Yun, S, Lu, H, Greenlund, K J and **Croft, J B** 2015 Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *Am J Epidemiol* 182: 127–137. DOI: https://doi.org/10.1093/aje/kwv002

Zhang, Z, Zhang, L, Penman, A and **May, W** 2011 Using small-area estimation method to calculate county-level prevalence of obesity in Mississippi, 2007–2009. *Prev Chronic Dis* 8: A85.

How to cite this article: Chien, L-G, Lin, G, Li, X and Zhang, X 2018 Disparity of Imputed Data from Small Area Estimate Approaches – A Case Study on Diabetes Prevalence at the County Level in the U.S. *Data Science Journal,* 17: 8, pp. 1–11, DOI: https://doi.org/10.5334/dsj-2018-008

Submitted: 27 November 2017 **Accepted:** 09 March 2018 **Published:** 06 April 2018

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS