Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Shiwei Tang Date

Understanding and Modeling the Uncertainty in the American Community Survey: Detecting the Spatial Correlation in Uncertainty at the County Level with Conditional Autoregressive Models

By

Shiwei Tang
Master of Public Health

Biostatistics and Bioinformatics

Lance Waller, Dr
Committee Chair

Howard Chang, Dr
Committee Member

Understanding and Modeling the Uncertainty in the American Community Survey: Detecting the Spatial Correlation in Uncertainty at the County Level with Conditional Autoregressive Models

By

Shiwei Tang

B.S.
Florida Institute of Technology
2014

B.S.
Shanghai Ocean University
2014

Thesis Committee Chair: Lance Waller, Dr
Thesis Committee Member: Howard Chang, Dr

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2017

Abstract

Understanding and Modeling the Uncertainty in the American Community Survey: Detecting the Spatial Correlation in Uncertainty at the County Level with Conditional Autoregressive Models
By Shiwei Tang

Starting in 2010, the US Census Bureau replaced the Long Form with the American Community Survey (ACS), a rolling sample with annual and five- year data summaries. Because of the sampling design, the ACS reports the uncertainty of estimation via a margin of errors, which can be used to model the precision of results. Recent research illustrates latent spatial correlation in these margins of errors at different level of aggregation, and conditional autoregressive (CAR) models are popular way to incorporate such spatial correlations. In this thesis, we use spatial generalized linear mixed models (GLMMs), based on Poisson regression and the CAR model, to incorporate uncertainty in ACS-based population counts and covariates in small area estimates of disease risk. Both Markov chain Monte Carlo (MCMC) and Integrated Nested Laplace Approximation (INLA) are used to implement GLMMs, and we find both methods provide similar results. However, INLA is much more efficient in computation while MCMC provides slightly better interval coverages for fixed parameters.

Understanding and Modeling the Uncertainty in the American Community Survey: Detecting the Spatial Correlation in Uncertainty at the County Level with Conditional Autoregressive Models

By

Shiwei Tang

B.S.
Florida Institute of Technology
2014

B.S.
Shanghai Ocean University
2014

Thesis Committee Chair: Lance Waller, Dr
Thesis Committee Member: Howard Chang, Dr

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Biostatistics
2017

1. Introduction

When government agencies, research institutions, and companies seek to understand patterns and trends in the United States population to make plans and decisions, they often are interested in incorporating United States Census data to provide background reference values on population size, age distributions, demographics, etc. The US Census Bureau collects population data using several survey instruments. To balance the community's needs and efforts required to complete the survey, there are two types of questionnaires, which are Short Form and Long Form. Short Form, which only takes about 10 minutes to finish, asks some basic information such as sex, age, race and housing status. Long Form, which includes all questions from Short Form, contains many questions about population and housing. It takes about 38 minutes to finish and provides social-economic details. (U.S. Census Bureau 1999) Starting in 2010, the American Community Survey (ACS), a rolling sample of Americans, provides a smaller, but more frequent survey sample of the US population. The ACS results are published by US Census Bureau annually, and, according to current plans, will replace the Long Form completely by 2020. Based on detailed survey sampling techniques (U.S. Census Bureau 2016), the ACS reports both point estimates of demographic summaries and associated levels of uncertainty associated with the point estimates. Historically, Census estimates have been treated as fixed quantities in statistical and epidemiological models, but the ACS reports allow more detailed consideration of estimation uncertainty within these models.

The public health research community needs to cope with the new type of census data represented by the ACS including both point estimates and associated uncertainty, and

use the data in appropriate ways. Here we examine analytic techniques to incorporate the reported uncertainty in the ACS data in public health studies, specifically when estimating and mapping disease rates for small geographic areas (e.g., counties or census tracts). We are interested in assessing the impact compare two different estimation algorithms with a simulation study based on the ACS data.

1.1 Modeling

Disease mapping and small area estimation are related modeling techniques that have many applications in public health (Gelfand et al., 2010). These overlapping families of methods share two competing main goals. One involves estimating local values precisely for each area; the other is having "small" areas to provide local details within broad trends. However, these main goals of disease mapping are in direct contradiction with each other. Usually, a small area comes with a small sample size, which means that precision of results is decreased (Gelfand et al., 2010) and aggregating small areas to increase sample sizes sacrifices local detail. Small area methods seek to meet both goals by providing estimators that represent a weighted compromise between local and global data to improve local precision without entirely giving up local information.

Disease mapping methods extend small area estimation by assuming positive spatial correlation between nearby observations, allowing local estimates to borrow more information from nearby areas rather than borrow information equally across all areas. The spatial setting provides a smoothing of extreme values and an easy way to visualize geographic patterns of disease. As one example, popular conditional autoregressive (CAR) models allow us to borrow information from neighborhood areas, and provide more accurate local estimates than one would obtain ignoring neighboring values.

In spatial data, the CAR models often are implemented by Markov chain Monte Carlo (MCMC) algorithms (Gelfand et al., 2010) and, more recently, Integrated Nested Laplace Approximation (INLA) (Blanogiardo & Cameletti, 2015). Usually, MCMC is the first choice when we are looking for posterior sampling for Bayesian models such as the CAR model. However, the biggest challenge is obtaining convergence in posterior samples in a reasonable amount of time (Carroll et al., 2015). MCMC approaches can be very time consuming and put a heavy burden on CPU for simulation. Unlike MCMC, INLA is a new algorithm based on numerical rather than Monte Carlo integration to perform Bayesian analysis and calculate approximations of posterior marginal directly (Rue et al., 2009), allowing much quicker computation than MCMC. INLA has been widely used to analyze lattice data with CAR model, and examples are easy to find (Bivand et al., 2015).

1.2 Data Sources

Before 2010, the only official, reliable resource for demographic characteristics of the US population was the Decennial Census (DC). US Census Bureau (USCB) conducts this survey for the entire country with two different forms: the Long Form and the Short Form. The Short Form is given to all households and collects basic demographic information, e.g., number of people in housing units, age, race, and Hispanic origin (Herman, 2008). The Long Form, which is distributed to 1-in-6 households, collects the same questions as in Short Form with additional questions about socioeconomic topics including education, income, housing characteristics, employment, and disability status (Herman, 2008). Although government agencies, researchers, and companies rely on these data to make decisions, it is obvious that the data will be outdated when applied to

the last few years of the 10-year cycle. A new census approach was introduced in 2010 to replace the Long Form in order to provide more timely data.

The American Community Survey (ACS) is a complex sample conducted by US Census Bureau annually and yields dataset starting in 2010 for research and analysis. In 1996, US Census Bureau decided to switch to the ACS (Bazuin & Fraser, 2013). It aims to help decision makers understand the community at different scales (State, County or Census Tract) by providing basic information regarding households and individuals annually. The ACS is distributed to about 250,000 households every year on a rolling monthly basis. Compared to Decennial Census (DC), which is published every 10 years, the ACS provides demographic estimates with more frequent updates.

A very special feature of the ACS is that it changes point-in-time statistics to a period average over 1, 3, or 5 years. Annual information is available for areas with more than 65,000 people. It covers data collected from January to December of the previous year. For areas with population size between 20,000 to 64,999 people, a three-year average is provided. For areas with a population size less than 20,000 people, a five-year average is the only available information (Herman, 2008). However, due to the decrease of the Census budget, the three-year estimation was discontinued at 2014 (Buff, 2015).

The ACS provides many advantages over the Decimal Census. First, it provides the most up-to-date estimates of basic social factors in the population on a relatively local scale. The ACS allows up-to-date data each year, and across the entire country. Second, the ACS data provides an associated margin of error for each local estimate. The margin of error allows users to analyze variability of the estimates, as well as the point estimates themselves. Third, with the yearly update, it is easier to examine the relationship between

population change and local or national events. Sampling weights are used at the person and housing unit level to adjust the importance of each complete survey, and the same weights of a survey apply to all estimates relating to that subject. For example, an Asian respondent whose income is $20,000 per year with sampling weight is 10 contributes $200,000 in aggregate income and 10 Asian residents to the estimated total population size (Spielman et al., 2015). Lastly, since the ACS dataset maintains a permanent, well-trained staff rather than a decennial pulse of data collectors, the collected data will be more accurate gathered by individuals with full understanding of the questionnaires. (Bazuin & Fraser, 2013).

With a limited budget, the USCB has to compromise between the precision, timeliness, and cost of their data collection efforts. Even with its advanced survey sampling design and best implementation efforts, bias and uncertainty will remain in the ACS and the goal is to quantify these and incorporate adjustments where possible. Although response to the survey is required by law, only 65% of selected ACS individuals reply. The USCB dispatches staff to follow up with a sample of non-respondents in person if individuals do not response by mail or Internet. However, not all non-respondents will receive personalized follow-up because of the high cost of doing so (Spielman et al., 2014).

The reported margin of error reflects the uncertainty associated between two independent samples from the same population. In a simple random sample, it is easy to estimate margin of error based on sample size and variation in the targeted population. Estimating margin of error is much more difficult in complex designs such as that of the

ACS survey, but the details have been carefully worked out and documented by the USCB (Spielman & Singleton, 2015).

We note that, because the design of the ACS is much more complex than the US census Long Form, it requires knowledge and care from users. Many users naively believe that the ACS provides the same information at the same accuracy as the decennial census data and ignore the reported margins of error (Napierala et al., 2017). The increased uncertainty in the ACS data comes directly from its smaller sample size compared to US Census Long Form. As Starsinic (2005) states, overall, Standard Error(SE) from the ACS is 1.41 times larger than the SE from US census Long Form at the county level and 1.8 to 2.0 times at the tract level.

In addition to sampling error, there are many factors influencing nonsampling error in the ACS. First, the ACS estimated population sizes may include random and systemic errors, especially in small areas. Another source of nonsampling error results from long data collection periods. Since the ACS is a rotating survey, some time passes between direct observation of data from particularly small areas. As a result, it is somewhat harder to directly interpret results since they represent summary estimates over periods of time rather than direct point estimates for each year. Similarly, the ACS also mismatches with population controls from Population Estimates Program (PEP), since PEP estimates the population on a single day rather than one- or five-year intervals of time (Napierala et al., 2017). There are also some systematic differences coming from data collection methods and questionnaire design between the ACS and US Census Long Form. However, compared to the Census 2000 Long Form, the ACS has higher overall response

rates with the same completeness rates. This helps the ACS to reduce nonsampling error (Napierala et al., 2017).

Recent research highlights latent spatial correlations in ACS uncertainty values. For example, when exploring associations between median household income and uncertainty in household income at the tract-level, analysts find that tracks with a high (or low) uncertainty are likely surrounded by tracts with similar uncertainty (Spielman et al., 2014). It is reasonable to assume that there are spatial correlations in reported ACS margins of error at the tract level. Spatial correlation, albeit at different spatial scales, may also be found at larger areas of aggregation (e.g., counties, states). When we are examining county or higher levels of the ACS data, we have more data within each area than in the tract-level data, for the same period of time. It is not always helping to reducing MOEs if we aggregate smaller areas in to larger areas as some aggregation will combine data from dissimilar types of neighborhoods. This question represents an open area of research, and recent research suggests that the aggregation recommendation from Census Bureau tends to over-correct for area-specific uncertainty (Spielman et al., 2015). Further research is clearly needed on this issue.

As we stated above, the ACS is a compromise of spatial and temporal data; the within area uncertainty is not uniform distribution across the full collection of small areas. In order to make the best use of ACS in spatially-referenced small area public health research, a better understanding of patterns in reported margins of errors and their impact on conclusions is required. In this thesis, we are interested in to finding a better way to understand and use reported ACS margin of error information to obtain accurate small area estimation, with particular focus on incorporating the ACS within Poisson

regression-based disease mapping models. More specifically, we are particularly interested in incorporating reported uncertainty in ACS-based population counts and covariates and comparing MCMC and INLA implementations of spatial generalized linear mixed models (GLMMs), based on Poisson regression and CAR model. We use a Poisson model as an approximation to a binomial model where the disease is rare. For the analysis, we use the following software and R packages: R version: 3.2.3, CARBayes version: 4.6, and INLA version:0.0.1485844051.

In this paper, we will discuss the set of our data simulation. Then, we will describe how to compare the results from MCMC and INLA. Next, we will provide our results. At last, we will discuss our finds including advantages and disadvantages of both algorithms.

2. Methods

2.1 Simulation model

In this paper, we focus on fitting Poisson regression disease mapping models with MCMC and INLA, and assess the impact on model results of uncertainty in ACS. To evaluate the impact of uncertainty in ACS, we simulate data and compare the results of both approaches.

As discussed above, a Poisson data model is frequently used in small area estimation of risk of a rare disease. We define y_i as the count outcome observed in the *i*th area, and we use the state of Georgia and its 159 counties as our study area. We assume that we have the population size (denoted p_i) for each area (county), and we set a fixed expected individual level rate (E). In our model, the outcome follows a Poisson distribution as follows:

$$y_i = Poisson(Ep_i\theta_i)$$

$$\log(\theta_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

where θ_i denotes the relative risk of ith area.

With a fixed expected rate for all individuals, we remove interindividual heterogeneity (not of interest here) and focus on variability due to uncertainty in the ACS, i.e., variability in p_i, x_{1i}, and x_{2i}. More specifically, we include two spatially varying variables, which represent percent of poverty (x_{1i}) and percent graduating from high school (x_{2i}), both reported at the county-level by the ACS. We simulated 200 datasets based on a fixed set of parameters (E = 0.05, β_0 = 0.1, β_1 = 0.01, β_2 = 0.02). Since we choose relative small β values and individual level risk E, the outcome mimics rare disease incidence.

2.2 Fitted model

We next fit standard disease mapping models to our simulated data. To access spatial correlation in the data, we fit an intrinsic CAR model. In this study region, we have J = 159 separate areas and a set of outcomes $\mathbf{y} = (y_1 \ldots y_{159})$. We define an offset $\mathbf{O} = (O_1 \ldots O_{159})$, where O_i equals to Ep_i. Next, we create a matrix of covariates X = $(X_1,\ldots,X_{159})$ with a spatial random interecept $v = (v_1,\ldots v_{159})$, which is used to explain any spatial autocorrelation for which fixed effects fail to account.

To explore variability in ACS values, we calculate the standard error of variables by Marginal or Error (MOE)/1.645 for covariates. For each simulation, we generate values for, $x_i=(1,x_{1i},x_{2i})$ for each county, where $x_{ki} \sim N(E(x_{ki}), SE(x_{ki}))$ and $0<x_{ki}<1$. These in turn allow us to simulate outcome values y_i for $i = 1,\ldots,J$. (Note that the simulation model does not include any random effects.

Once we have the simulated outcome values, we fit the following hierarchical Bayesian disease mapping model:

$$\log\{E(y_i|\beta)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + v_i + u_i$$

with an intrinsic CAR prior for the spatial random effects

$$v_i|v_{-i}, \boldsymbol{w}, \tau \sim N\left(\frac{\sum_{j=1}^{J} w_{ij} v_j}{\sum_{j=1}^{J} w_{ij}}, \frac{\tau}{\sum_{j=1}^{J} w_{ij}}\right)$$

standard prior distributions for covariate effects

$$\beta_i \sim N(0,1000),$$

and extra Poisson variation defined by another random intercept following

$$u_i \sim N(0, \sigma^2)$$

and finally, a hyperprior on the variance components

$$\tau, \sigma^2 \sim Inverse - Gamma(a, b)$$

We note that $\boldsymbol{v}=(v_1,\ldots,v_{159})$ denotes a vector of random intercepts with spatial structure defined via a CAR structure with precision τ, and weight matrix $\boldsymbol{w}$. The weight matrix $\boldsymbol{w}$ is a symmetric J by J matrix where $w_{ij} = 1$ if area i and area j have a shared boundary, and $w_{ij} = 0$ otherwise. We generate the $\boldsymbol{w}$ matrix using the tigris and spdep packages in R. Tigris is a package that accesses United States Census Bureau geography files and downloads a shapefile of Georgia at the county level. The spdep package helps us define neighbors on this shapefile via the poly2nb function (Bivand et al., 2015).

To complete the model definition, we assign a normal prior distribution to each β_i with mean = 0 and variance = 1,000, and we also assign a gamma prior distribution to τ with parameters $a = 1$ and $b = 0.01$.

Both the R-INLA and CARBayes packages can handle this model well (Bivand 2015), and we compare their abilities to recover true associations (the βs) in the presence of measurement error (in p, x_1, and x_2) consistent with typical ACS margins of error.

2.3 Comparison Techniques

For this study, we fit the same model in both MCMC and INLA to each simulated data set. Both models were fitted in R. We compare the performance of MCMC and INLA by mean square error (MSE). We calculate MSE for parameter estimates, fitted y values, and relative risk.

$$\mathrm{MSE}(\beta_k) = \frac{1}{N}\sum_{n=1}^{N}\left(\beta_k - \beta_{kn,sim}\right)^2$$

$$\mathrm{MSE}(Y) = \frac{1}{N}\sum_{n=1}^{N}[\frac{1}{J}\sum_{j=1}^{J}\left(y_j - \mathrm{y}_{jn,sim}\right)^2]$$

$$\mathrm{MSE}(\theta) = \frac{1}{N}\sum_{n=1}^{N}[\frac{1}{J}\sum_{j=1}^{J}\left(\theta_j - \theta_{jn,sim}\right)^2]$$

K denotes the number of covariates, N is the number of simulation times, and *J is* the number of counties. These results are based on averaged on 200 simulated datasets with the same underlying parameters.

We also use graphical displays to examine differences between the simulated *y* and fitted *y* values averaged over our 200 simulated datasets. This measure determines how much the fitted *y* differs from simulated *y* values, even after allowing for spatially structured and unstructured excess variation.

2.4 Data Summaries

All data are primarily from the 2011-2015 ACS. It is the most recent ACS five-year estimation. We choose the state of Georgia as our research area, which contains 159 counties.

	Mean	Median	Min	Max
Population Size	62,935	22,731	1,721	983,903
Education (%)	19.7	20.1	6.1	39.0
SE of Education (%)	1.5	1.5	0.2	4.0
Poverty (%)	22.2	21.9	6.9	42.2
SE of Poverty (%)	2.0	1.9	0.3	5.6

Table 1. Descriptive statistics for Georgia at the county-level for the 2011-2015 ACS.

According to Table 1, the mean county population size in Georgia is 62,935 people (range = 1,721, 983,903), the mean of education (% of population who does not graduate from high school and above age 25) is 19.7% (SE=1.5 %), and poverty (% of population who is below poverty line) is 22.2% (SE=2.0 %).

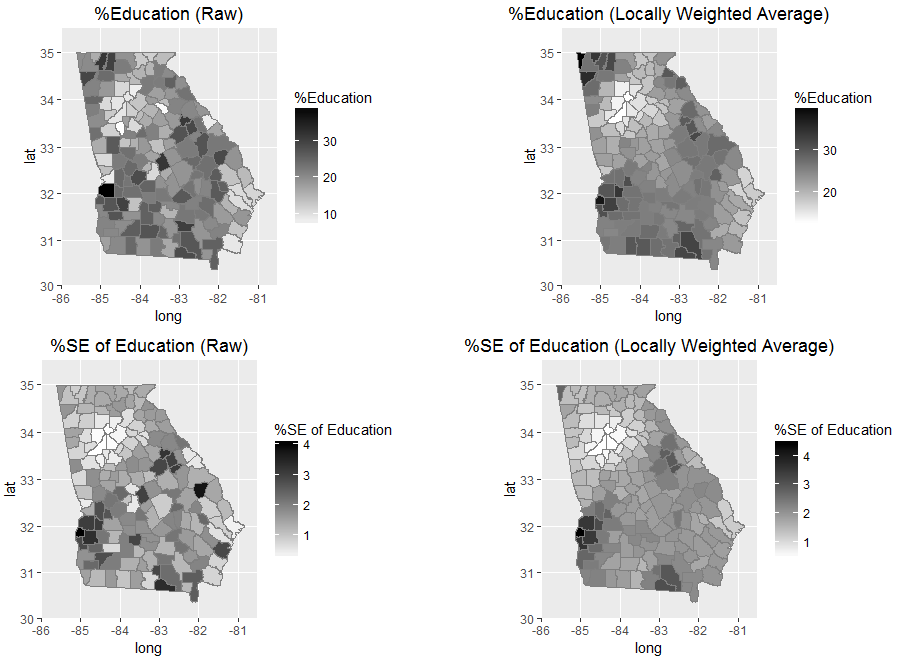

Figure 1. County-specific values from the 2011-2015 American Community Survey (left) and locally weighted average values (right) for the covariate % population who does not graduate from high school and above age 25. Covariate values in the top maps, standard errors on the bottom maps.

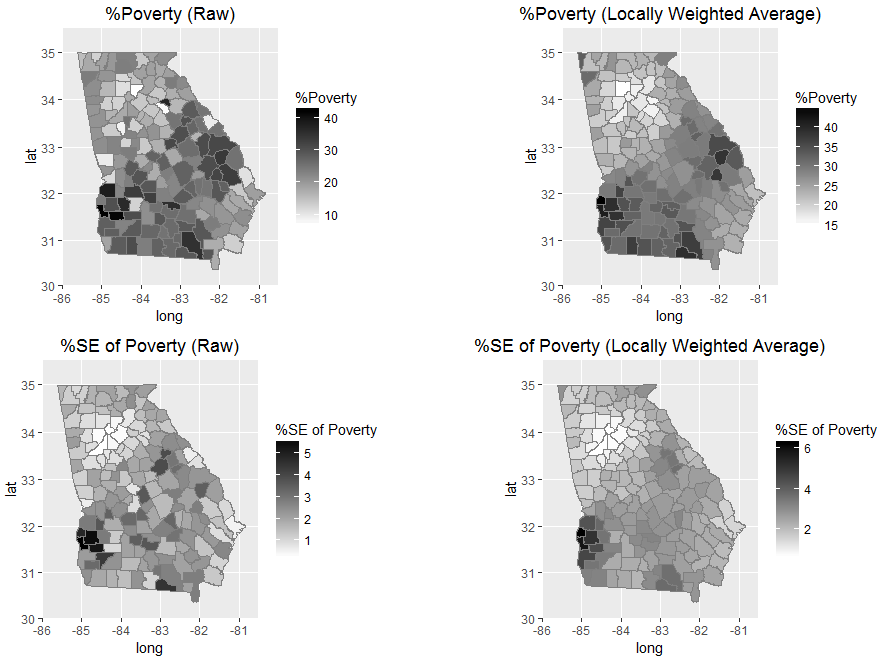

Figure 2. County-specific values from the American Community Survey (left) and locally weighted average values (right) for the covariate % in poverty. Covariate values in the top maps, standard errors on the bottom maps.

Since we are using percent to display the spatial heterogeneity at county level, we may face small number problems (Waller and Gotway 2004). As we state above, counties of Georgia have a very different population sizes. The percent of education and poverty covariates are based on population size, so some counties have more precise local estimates than others. This may obscure some spatial patterns in maps. Spatial smoothing helps reduce the noise by borrowing information from neighboring regions. As an initial descriptive display, we use locally weighted averages in Figures 1 and 2 to show smoothed values for each county by averaging the value of each county with that if its neighboring counties who share the boundary. We use adjacency to define neighbors. More specifically, if p_1, p_2, …,p_J are the percent, smoothed percent is :

$$\widehat{r_i} = \frac{\sum_{j=1}^{J} w_{ij} p_j}{\sum_{j=1}^{J} w_{ij}}$$

$$\text{Where } w_{ij} = \begin{cases} 1, \text{if county } i \text{ and } j \text{ share a boundary or } i = j, \text{and} \\ 0, \text{otherwise.} \end{cases}$$

According to Figure 1 and Figure 2, the smoothed maps illustrate broad spatial patterns of both the covariates percent of Education and Poverty and their standard errors with pockets of high values (and high associated SEs). These patterns motivate our inclusion of CAR-based random effects in our models below.

The other tool, which is widely used to detect spatial correlation, is Moran's I autocorrelation index. We find Moran's I is 0.294 for Education, 0.371 for Poverty, 0.377 for SE of Education and 0.468 for SE of Poverty (All p-values < 0.01). These results suggest significant spatial correlation in both predictors and standard errors.

We also examined the local components of Moran's I (Bivand et al., 2009). According to Figure 3, we find that about 20% of counties exhibit significant positive spatial correlation with their adjacent neighbors. We note that geographic distribution is similar for both covariates and their standard errors, and there are few "Hot Spots" in the metropolitan Atlanta area, which have large population sizes.

Figure 3. County-specific P-values from Local Moran's I. Covariate values in the top maps, standard errors in the bottom maps.

From these maps, we also find that the spatial patterns are very similar in all maps, and inversely related to local population size, but also may be due to mixing dissimilar types of neighborhoods in aggregating data from tract level to county level.

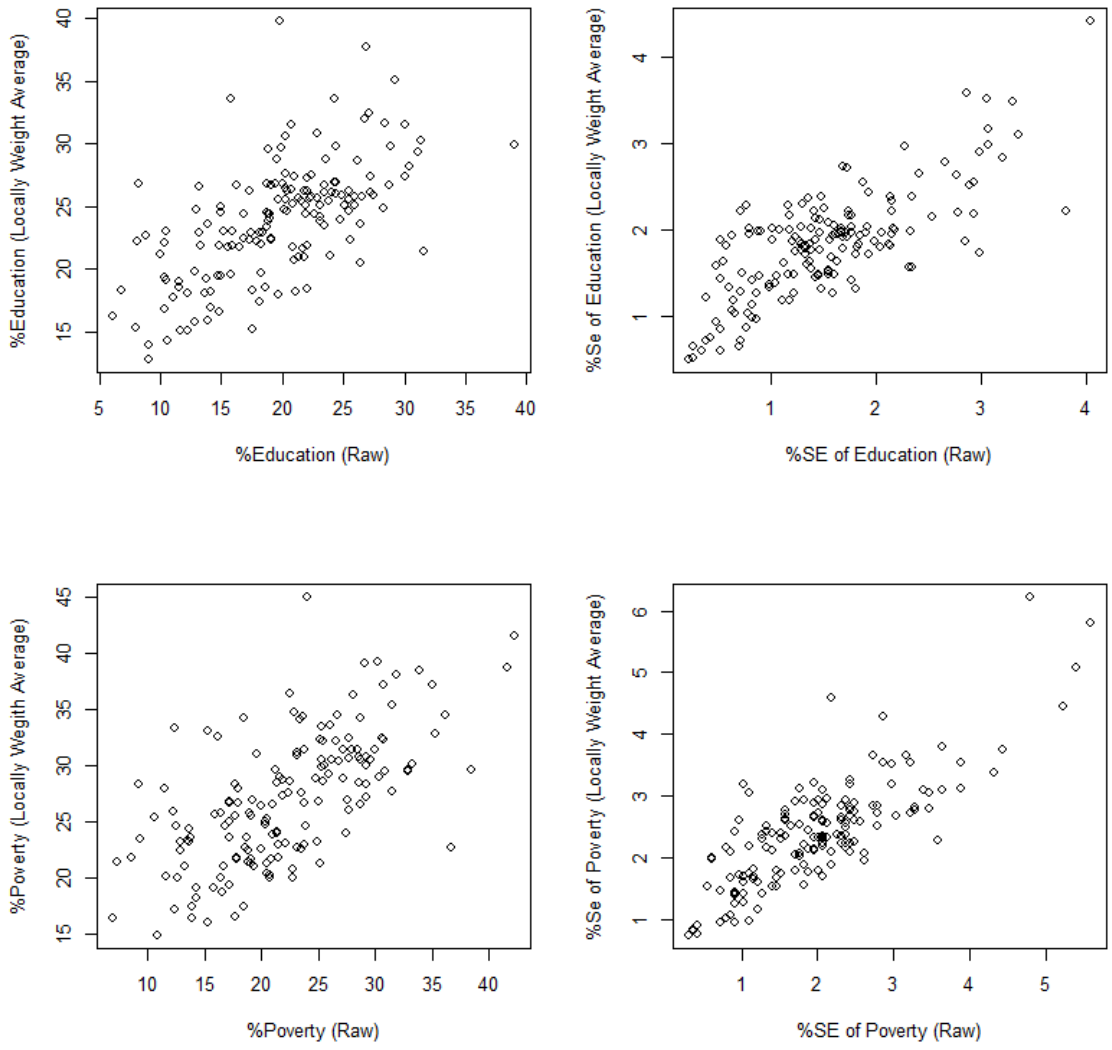

Figure 4. The relationship between raw data and locally weight average data.

According to Figure 4, we see that the locally weight averages do not oversmooth the results but rather highlight the spatial trends and correlations already existing in the original data.

It is worth noting that we use locally weight averages for display of potential spatial patterns, we still use the original raw data in our study.

3. Results

It is important that our MCMC model is converged. We checked 200 simulated datasets using the Geweke Convergence Diagnostic in CARBayes package. All datasets show that the model has been converged after 200,000 iterations.

	True Value	MCMC	INLA*
Intercept	0.1	0.144	0.144
95% Credible Intervals		(0.104, 0.184)	(0.111, 0.177)
Education	0.01	0.011	0.011
95% Credible Intervals		(0.009, 0.013)	(0.009, 0.012)
Poverty	0.02	0.017	0.017
95% Credible Intervals		(0.016, 0.019)	(0.016, 0.019)
Tau		0.002	0.002
Sigma		0.001	0.001
MSE (BETA0)		0.002	0.002
MSE (BETAEDU)		0.000001	0.000001
MSE (BETAPOV)		0.000001	0.000001
MSE (Y FITTED)		338.83	340.71
MSE (RELATIVE RISK)		0.31	0.31

*Mean was used instead of median.

Table 2. Parameter estimates from MCMC and INLA as well as the true values defining our simulated data values.

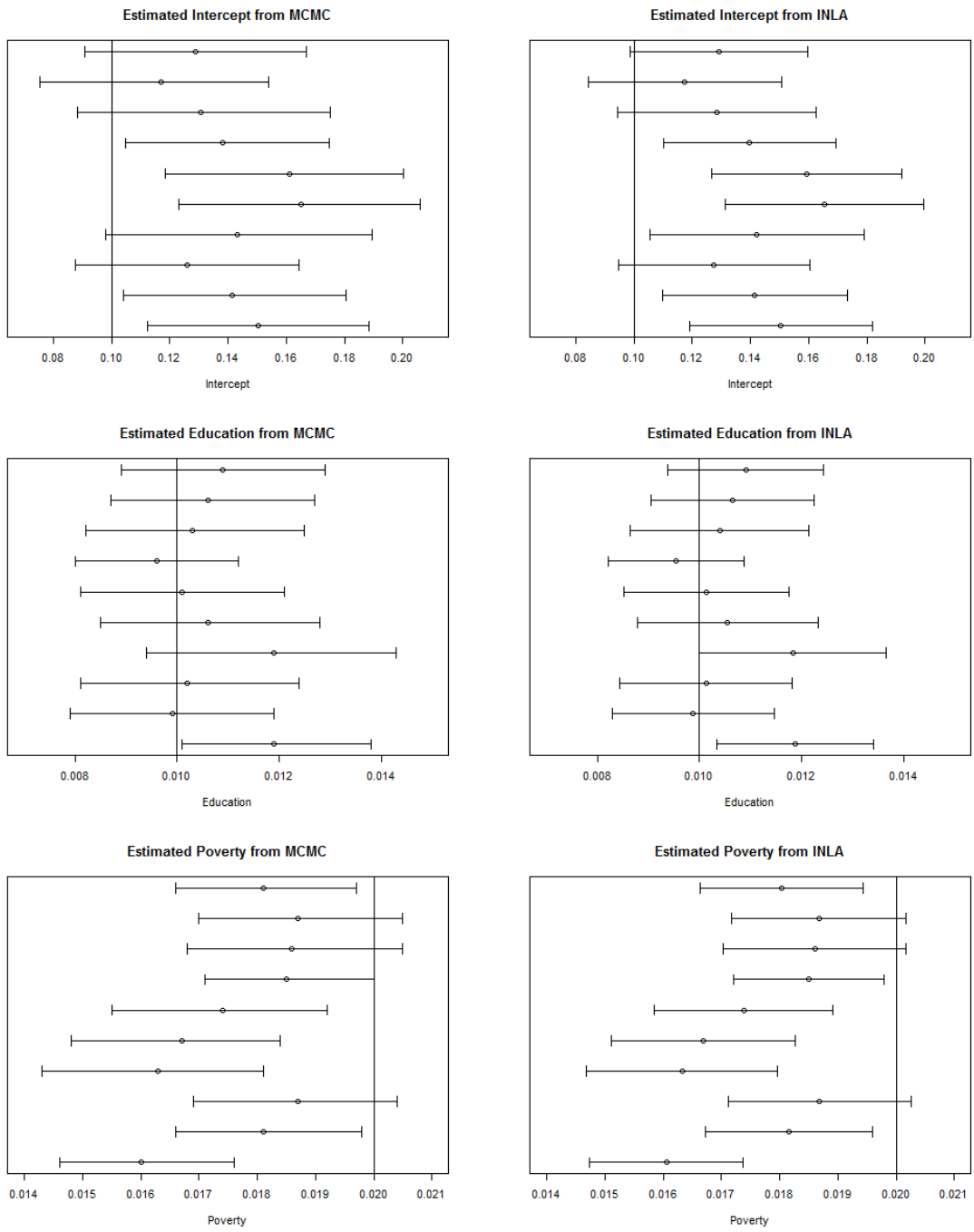

Figure 5. Ten example parameter estimates and 95% credible intervals from MCMC and INLA. Reference values indicated by vertical lines.

Table 2, Figure 5 and 8 show the results of parameter estimation compared to the true reference value for ten data simulations. The estimates from MCMC and INLA are very

similar, and the greatest difference (y-y_{fit}) is less than 30. Compared to the true value, both MCMC and INLA tend to overestimate the intercept and underestimate the coefficient associated with % below Poverty. We note the negative correspondence between the intercept and the Poverty coefficient (when the estimate of one is high, the other is low). While the coefficient associated with % with high school degree (Education) is slightly overestimated, most of 95% credible intervals contain the true value from each simulation. The overall coverage probability across 200 simulations for the intercept is 36.5% (MCMC) and 19.5% (INLA), for Education is 95.5% (MCMC) and 87.5% (INLA), and for Poverty is 11.5% (MCMC) and 6.5% (INLA). We also note that the MSE estimates from MCMC and INLA are almost the same, which means that both algorithms provide comparable fit of our models to the datasets.

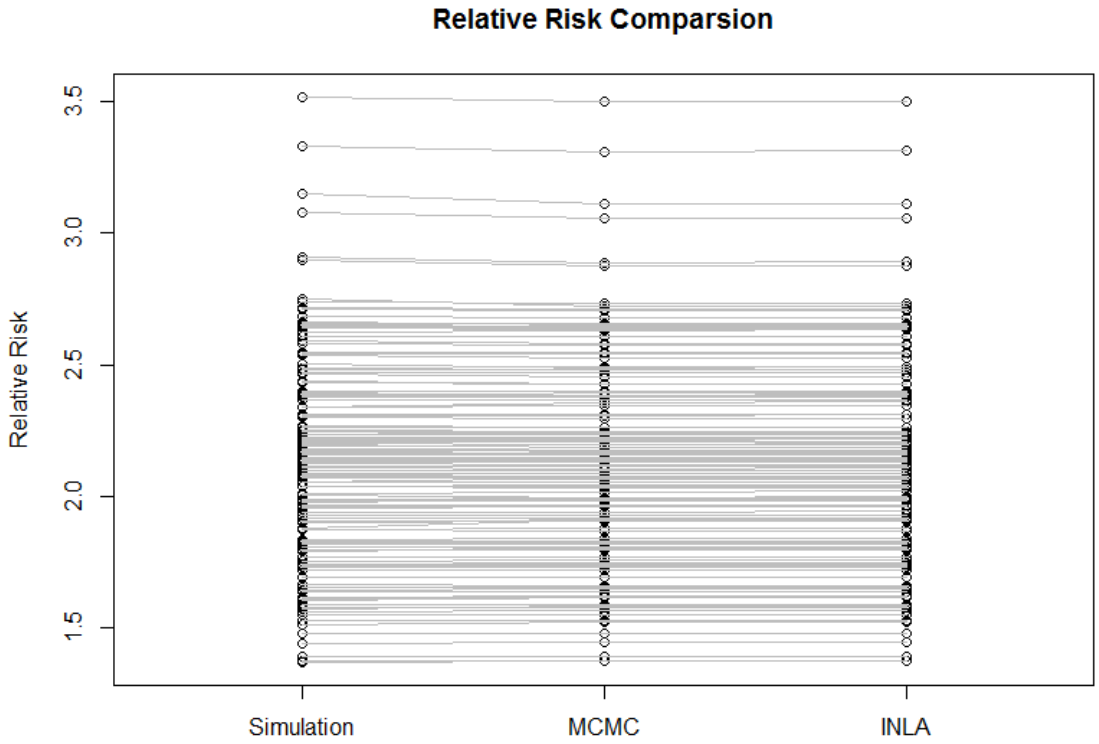

Figure 6. Estimated relative risk comparison between MCMC and INLA.

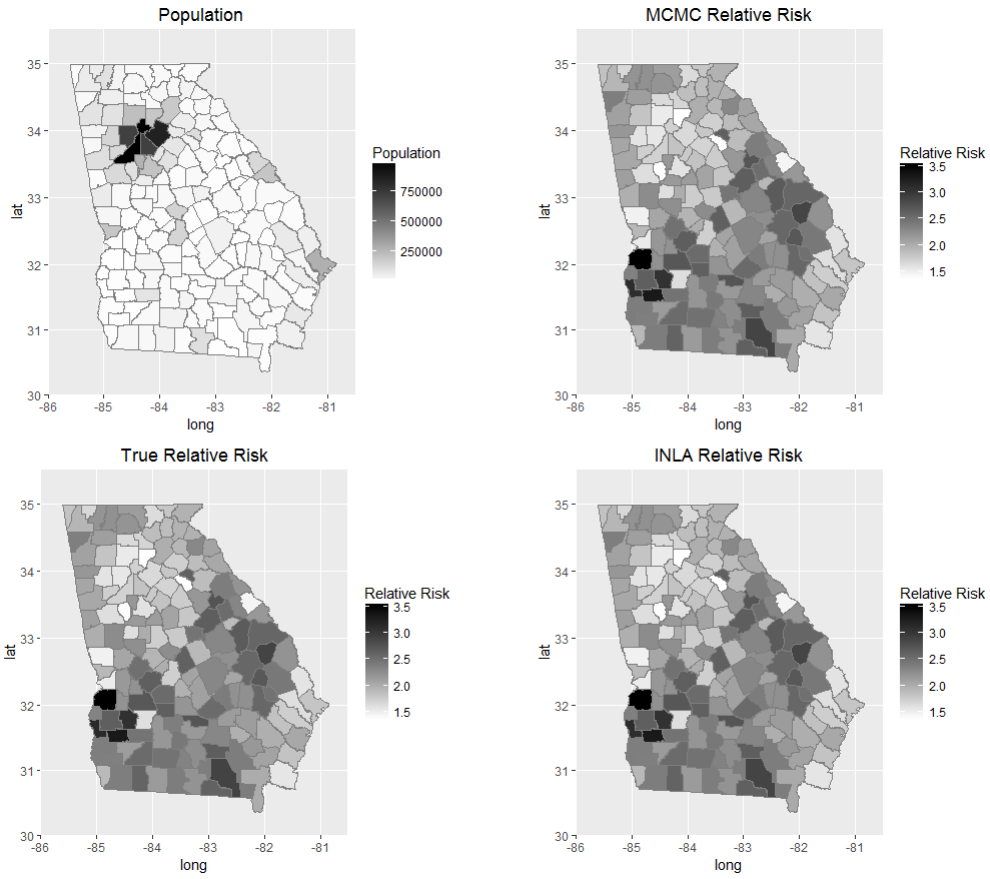

Figure 7. Relative risk mapping based on MCMC and INLA.

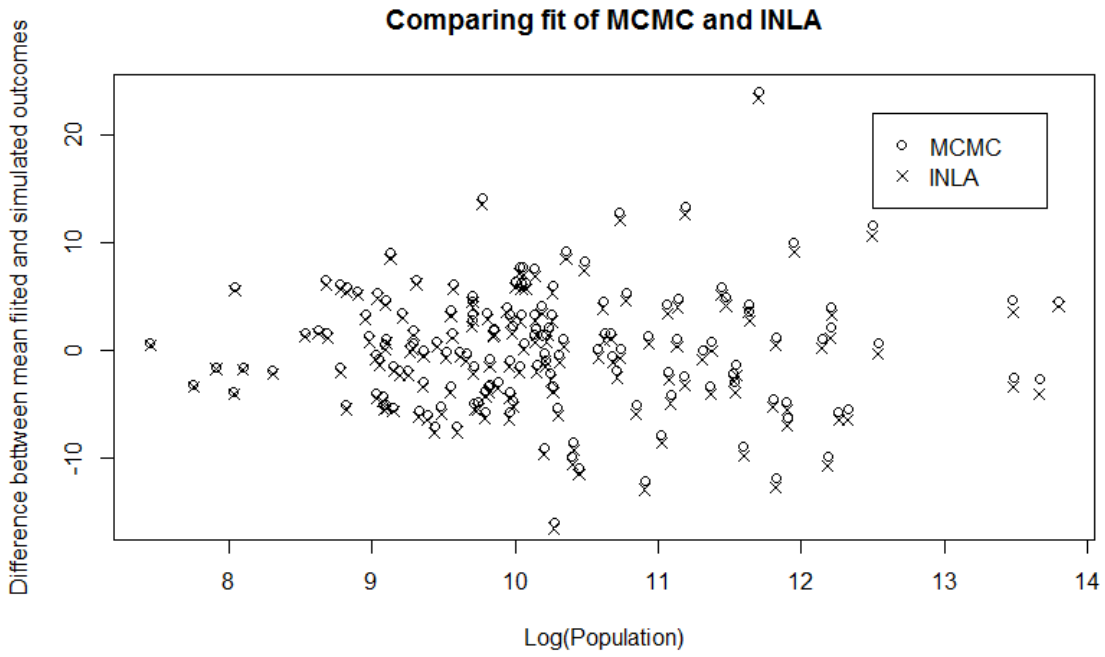

Figure 8. Relationship between log (population size) and difference in mean fitted and simulation outcomes.

According to Figure 6, 7, and 8, we notice that there is very little difference in the county-specific outcome and county specific relative risk estimates, and both results are close to the true values underlying the simulation, suggesting very little bias in overall estimation due to the offsetting bias between the intercept and poverty effects. What's more, there is not much difference in precision regardless of population size of country. The map also shows that both MCMC and INLA results show very similar spatial patterns, which are very closed to true value.

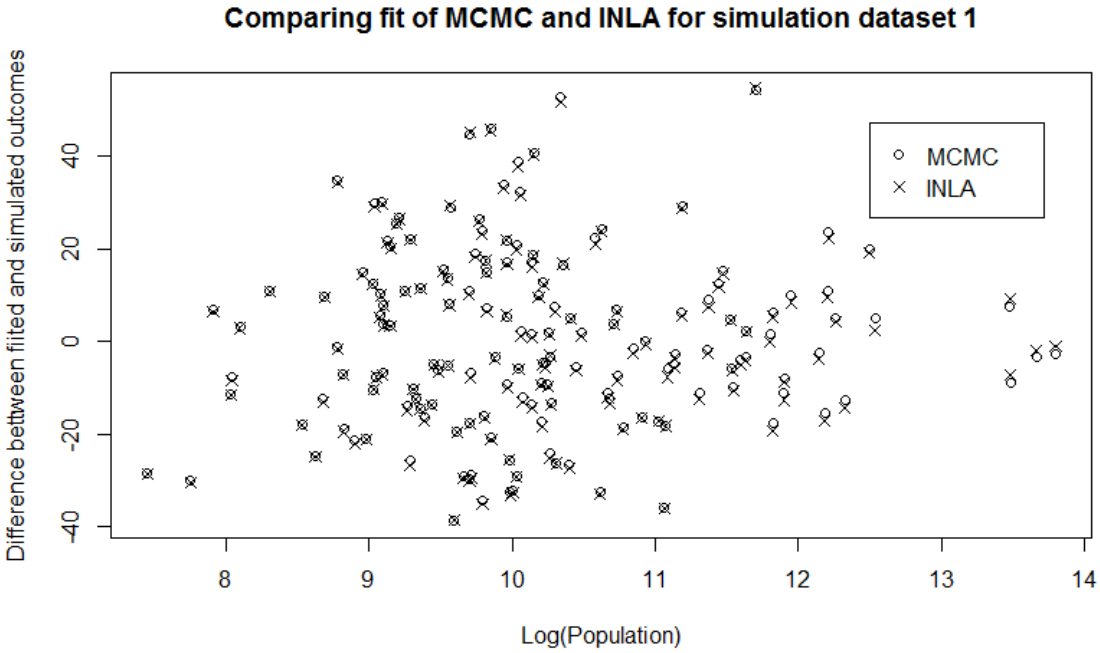

Figure 9. Relationship between log(Population) and difference in fitted and simulation outcomes based on dataset 1.

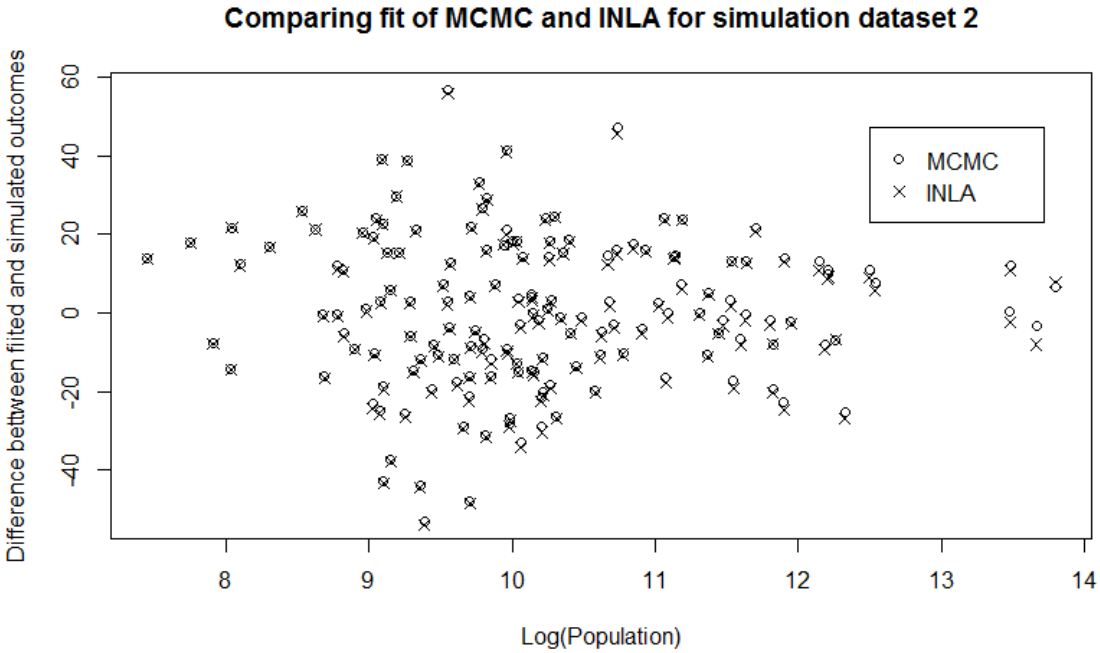

Figure 10. Relationship between log(Population) and difference in fitted and simulation outcomes based on dataset 2.

We also plot outcome comparison for 2 randomly chosen dataset and they provide similar results (Figure 9 and 10), but the greatest difference is larger than the difference from the mean value.

4. Discussion

As we see above, we find little difference in performance between INLA and MCMC suggesting either approach works well to fit standard disease mapping models. The parameter estimates are almost identical for both algorithms at default setting of CARBayes. While the estimates are similar, INLA runs much faster than MCMC. The computing time for 200 simulation datasets in INLA is less than 1 hour, while we found it takes about half day using MCMC to ensure convergence. We mention that MCMC provides better estimation for fixed effects compared to INLA, with respect to coverage probabilities for the intercept and the Poverty effect. This is somewhat surprising, since the estimates for the relative risk are very similar. The overall coverage probabilities for fixed effects in our 200 simulations are better in MCMC than in INLA.

Our results suggest that the spatial patterns of relative risk from both MCMC and INLA methods are very accurate even in the presence of ACS-type observation error, which means that the CAR model effectively adjusts for the effect of uncertainty in our dataset when we focus on local estimates of relative risk and outcomes, even though our data reveal measureable spatial correlations in local ACS values and reported margins of error. Even we did not simulate our dataset with spatial covariance, our simulated datasets still contain some spatial correlations, and these correlations are "soaked up" by the CAR random intercepts. So, it is reasonable to use CAR model to reduce uncertainty in the ACS dataset, and we can focus on outcomes and relative risk.

There are also some limitations to our approach model. As noted in earlier research, the non-spatial covariate effects are not well estimated in either MCMC or INLA (Carroll et al., 2015), suggesting the disease mapping models have better prediction than

estimation properties. Lawson (2015) suggests that this problem may be more of an issue wheen the magnitude of the true value is small. In our results, we also find similar performance to that reported in Lawson (2015)., We find that including the CAR terms in our model allows us to borrow neighborhood information to improve estimates of local values, even in the presence of spatially autocorrelated margins of error. It appears that neighboring information works well to stabilize fixed effect estimates to provide accurate overall predictions of local relative risk, but the specific fixed effect estimates merit more study. In our case, we found an underestimated Poverty term with an overestimated intercept, lessening the impact of the standard error on our local predictions.

There are several areas for future work. Some revised CAR models, such as the Leroux CAR Model (Leroux et al., 2000), help to balance the strength in random effects and fixed effects through an additional parameter and may have better estimation performance than the standard CAR model used in our research. The other thing we need to mention is that CAR model assume that we have the same strength of spatial correlation among the whole area, which is almost impossible in real world. A localized CAR model may help this problem and is worth examining in more detail in future work (Lee 2013).

5. Conclusion

Overall, both MCMC and INLA perform well to estimate model parameters and predict the local relative risk and outcomes in simulation datasets with random errors, and, like others, we find the INLA implementation to be considerably faster than MCMC. Our initial results are promising but more work is needed to better assess and understand the impact of using CAR disease mapping models.

Reference

Bazuin, J. T., & Fraser, J. C. (2013). How the ACS gets it wrong: The story of the American Community Survey and a small, inner city neighborhood. *Applied Geography*, *45*, 292-302.

Bivand, R., Müller, W. G., & Reder, M. (2009). Power calculations for global and local Moran's I. *Computational Statistics & Data Analysis*, *53*(8), 2859-2872.

Bivand, R., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *American Statistical Association.*

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Buff, B. (2015). The ACS 3-year Demographic Estimates Are History. Retrieved from http://apdu.org/2015/02/03/the-acs-3-year-demographic-estimates-are-history/

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., & Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and spatio-temporal epidemiology*, *14*, 45-54.

Herman, E. (2008). : an introduction to the basics. *Government Information Quarterly*, *25*(3), 504-519.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, *55*(13), 1-24.

Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (Eds.). (2010). *Handbook of spatial statistics*. CRC press.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials* (pp. 179-191). Springer New York.

Napierala, Jeffrey, and Nancy Denton. "Measuring Residential Segregation With the ACS: How the Margin of Error Affects the Dissimilarity Index." *Demography* 54.1 (2017): 285-309.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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 (statistical methodology)*, *71*(2), 319-392.

Spielman, Seth E., David Folch, and Nicholas Nagle. "Patterns and causes of uncertainty in the American Community Survey." *Applied Geography* 46 (2014): 147-157.

Spielman, Seth E., and David C. Folch. "Reducing uncertainty in the American Community Survey through data-driven regionalization." *PloS one* 10.2 (2015): e0115626.

Spielman, S. E., & Singleton, A. (2015). Studying neighborhoods using uncertain data from the American community survey: a contextual approach. *Annals of the Association of American Geographers*, *105*(5), 1003-1025.

Starsinic, M. (2005). American Community Survey: Improving reliability for small area estimates. In *Proceedings of the 2005 joint statistical meetings on CD-ROM* (pp. 3592-3599).

U.S. Census Bureau (1999). *The Long and Short of It: Why Does the Census Ask So Many Questions?*.(D-3239 (Rev. 6-99)) Retrieved from https://www.census.gov/dmd/www/pdf/d3239a.pdf.

U.S. Census Bureau (2016). *American Community Survey: Multiyear Accuracy of the Data (5-year 2011-2015)* Retrieved from https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/MultiyearACSAccuracyofData2015.pdf.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data* (Vol. 368). John Wiley & Sons.