12-15-2014

# Simulation Based Evaluation of Multiscale Small Area Health Models

Purbasha Dasgupta
*University of South Carolina - Columbia*

# SIMULATION BASED EVALUATION OF MULTISCALE SMALL AREA HEALTH MODELS

by

Purbasha Dasgupta

Bachelor of Science
Mumbai University, 2005

Master of Science
Mumbai University, 2005

---

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatisticcs

The Norman J. Arnold School of Public Health

University of South Carolina

2014

Accepted by:

Andrew Lawson, Director of Thesis

James Hussey, Reader

Bo Cai, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Lawson for his continuous support throughout my thesis preparation. His guidance has helped me throughout my research and writing of this thesis. I would also like to thank Dr. James Hussey and Dr. Bo Cai for being in my thesis committee, and for their comments and suggestions.

I also like to express my gratitude to all the faculty members, staff, and fellow students in the Department of Epidemiology and Biostatistics for their help and support. It has been a pleasant experience to study in this department.

Finally, I would like to thank my family and friends for their love and support. This thesis is a result of their endless help and encouragement.

## ABSTRACT

The effects of *scale* on the analysis of spatial data, often referred to as the modifiable areal unit problem in spatial studies, is one of the issues often encountered in small area health models. These spatial effects of scale are also seen in the areas of disease mapping where data are usually available in counts. Often there is a need to consider the different scales of aggregation that exist within count data, since inferences based on analyses can vary if we change the definition of the unit of analysis. This thesis provides a framework that describes the distribution of relative risk across a hierarchy of multiple scales. With the help of simulation studies, we explore a methodology that allows us to estimate and compare measures of relative risk in Poisson-based models for count data. The proposed method will be illustrated through the Georgia Oral Cancer data set (2004), which has count data at two levels: County and Public health district.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

## INTRODUCTION

Spatial health studies focus on describing and understanding the spatial variation in disease risk in geographically referenced data. One of the main focus areas is disease mapping, where the general objective is to describe the overall disease distribution on a map ,and highlight areas of elevated or lowered mortality or morbidity risk.

Disease mapping, sometimes known as ‘spatial epidemiology’, is commonly defined as the estimation and presentation of summary measures of geo-referenced health outcomes, and has a long history in public health and the study of diseases in human population. The goals of disease mapping include simple description of health risk, hypothesis generation, geographical allocation of health care resources, and assessment of spatial inequalities and estimation of background variability in underlying risk in order to place epidemiological studies in context. In disease mapping the spatial distribution of the disease is of particular importance (Lawson 2013). One of the main questions that arises is how best to analyze the incidence or prevalence of disease when we are given geographical data at different levels of aggregation (e.g., county level data, district level data, state level data, etc.). Different aggregation levels in public health data are commonly available, but seldom studied in tandem. We need methods for ascertaining how regional variations in health outcomes could relate to sub-regional outcomes in geo-

referenced data. Many studies use Bayesian hierarchical models but do not address this aggregation issue.

The Bayesian approach using Markov Chain Monte Carlo (MCMC) algorithms produce stable estimates in the spatially arranged regions. It allows us to investigate the unexplained heterogeneity in the disease maps and even accounts for the spatial and regional variability which have substantial effects on the relative risk estimation. This results in smoother estimates as the standard errors are stabilized over space. Smoothing makes it easier to visualize the pattern of disease distribution.

The geographic data in Public Health studies is often available at different spatial units or at different scales of aggregation ,and is commonly observed in counts for a set of regions. Spatial correlation is known to exist between these different scales and needs to be accounted for in the analyses as it violates the assumption of independence among observations. Due to correlation we may also lose statistical precision because the effective sample size is reduced when the sample provides redundant information. Spatial correlation occurs when neighboring areas have some similarity in their outcome ,and when a dependency exists between the values of a variable measured at these proximal locations. Many approaches that account for this correlation in Bayesian statistics make use of models that are log linear in risk and that allow for the inclusion of uncorrelated random effects and correlated spatial random effects. The inclusion of these two components expresses the overall spatial dependence observed in the data.

One of the standard spatial models that is widely used ,and that includes these two components is the conditional autoregressive (CAR) model. It was first introduced and studied by Besag et al. (1991). The CAR model is a disease mapping technique that is

used for spatial smoothing of relative risk. Typically, this method borrows strength from neighboring areas and helps in smoothing the local disease rates towards the local, neighboring disease rates thereby reducing the variance in the estimates (Venkatesan et al 2012). This is particularly useful when estimating disease risk in small areas as the direct estimates are likely to yield large variances because of the small sample sizes in these areas. For example, the standardized mortality ratio which is a common measure of the relative risk (calculated by dividing the observed deaths by the expected deaths) tends to have large variances in small areas. This is because the expected value is small in regions with small population size. On the other hand the variance in the SMR is small where the expected value is large due to larger population size. This makes the interpretation of the SMR ambiguous. One approach to overcome this problem is to employ the CAR model with random effects as it may help reduce the noise. By adding random effects in the model we are inducing a connection among the local relative risks, which helps in smoothing the risk estimates, especially where the rates are available in counts in small areas.

Geographic data is often aggregated to present the results of a study in a more useful context and to examine if potential spatial associations or relationships exist. The way the data has been aggregated has an influence on the results obtained from the data. This brings us to the phenomenon of modifiable areal unit problem (MAUP) that introduces a potential source of error ,and affects the results from spatial studies when aggregate data is used. The MAUP has important implications because, if the arbitrarily defined units of analysis change shape then the findings on these units change as well.

Chapter 3 discusses the MAUP further and explains how it can be taken into consideration in order to prevent its effects from biasing the results.

The MAUP has two main aspects - the scale effect and the zone effect. This thesis addresses the scale change issue (change of support problem) where the results change based on the data that are anal zed at higher or lower levels of aggregation. For our study to address the scale change issue we used the Georgia Oral Cancer data set that consists of aggregate data at two levels- the Public health (PH) district level and the county level. The PH level data is aggregated over county level data ,and each PH district contains a unique set of counties. It is possible that there might be some grouping effect observed for the counties that lie within a given PH district, thus a spatial correlation exists between these two levels. To evaluate the two scale effects, we fitted five variations of the Poisson model that included correlated and uncorrelated random effects. The five variations of the models differ in the way their relative risks have been aggregated for the two scales. This approach is an extension of previous approaches used to model Multiscale data in small areas. All our analysis were carried out via the BRugs package, which is one of the statistical software that integrates R and OpenBUGS ,and comes under the umbrella of the BUGS project.

The BUGS (Bayesian inference Using Gibbs Sampling) project provides flexible software for Bayesian analysis of complex statistical models using MCMC methods. WinBUGS and OpenBUGS are the two main versions of BUGS that provide easy access to fit a range of hierarchical or multilevel models for spatial data. WinBUGS also includes GeoBUGS, which allows mapping of the fitted parameters. These two BUGS packages can also be integrated with the R package that is freely available and has the

functionality of interacting with MCMC programs. BRugs which is a group of R routines that calls OpenBUGS from R has been recently used in several simulation studies (Cocchi, 2010; Calogine, et al. 2012, Latouche, et al. 2007). The features of BRugs allow users to analyze graphical models and it can be easily implemented through R. Thus, using the capabilities of these statistical and computational software packages, we can set up simulation studies to develop models that best describe our study data.

The overall chapter structure of this thesis is as follows:

Chapter 2 presents a general introduction on Bayesian spatial models and focuses on Conditional Autoregressive (CAR) models that are used for risk estimation. In Chapter 3 we describe the scale change issue that mainly arises in spatial analysis when data are observed at different levels of aggregation. We describe our methodology and computational approaches that include the scenarios for the simulation, along with application to the Georgia data set, in Chapter 4. The comparison of the simulation results and some discussion are detailed in Chapter 5. Figures and tables are presented in following sections ,and finally the appendix contains the code used for simulations.

# CHAPTER 2

## BACKGROUND ON BAYESIAN SPATIAL MODELING

Bayesian methods offer a flexible and robust approach to disease mapping, spatial analysis ,and decision making. The Bayesian approach starts with the formulation of a model that we hope is adequate to describe the situation of interest (Radford 1998). The fundamental idea involving this approach comes from the likelihood, which is a data level model that is dependent on parameters. These parameters have prior distributions and are hence allowed to be stochastic. This leads to a natural parameter hierarchy. The likelihood for data, conditional on parameters, can be defined as

$$L(y|\theta) = \prod_{i=1}^{m} f(y_i|\theta) \quad (1)$$

where $\theta$ is a $p$ length vector $\theta$: $\{\theta_1, \theta_2, ..... \theta_p\}$ and $f(.|.)$ is a probability density function and is a function of the observed sample $\{y_i\}$. The data are assumed to be conditionally independent, (i.e., the sample observed values of $y$, given the parameters, are independent), which makes it possible to take the product of individual contributions. This conditional independence is one of the important assumptions of Bayesian disease mapping (Lawson 2013).

In Bayesian analysis all unknown parameters are considered to be random variables. Hence *prior distributions* must be defined initially for parameters, as they

express the information available to the investigator before any data are involved in the analysis. The prior information defines a random variable that can take on a set of values with specified probability.

After analyzing some data we then apply Bayes rule to get a posterior distribution for the unknown parameters, which accounts for both the prior and the data. This posterior distribution helps in predicting distributions for future observations. The Bayesian method thus proves to be advantageous because it allows us to incorporate the prior distributions that account for the uncertainties related to the models and the parameter values. It also allows the specification of spatial dependence structures within prior distributions and hence simplifies spatial analysis.

One of the main goals of model-based Bayesian inference is to calculate the posterior distribution, $f(\theta|y,)$ for the parameter vector $\theta$ given the observed data $y$. This posterior distribution can be given as:

$$f(\theta|y)=\frac{f(y|\theta)f(\theta)}{f(y)} \propto f(y|\theta)\, f(\theta) \qquad (2)$$

The posterior distribution includes both the prior $f(\theta)$ and the observed data $f(y|\theta)$. The likelihood for $f(y|\theta)$ can be given as:

$$f(y|\theta) = \prod_{i=1}^{m} f(y_i|\theta) \qquad (3)$$

Usually a prior mean and variance is specified in the prior distribution. The prior mean provides a prior point estimate and the prior variance gives us an idea of the uncertainty around that estimate. The prior variance is set at a higher value if the investigator assumes that the point estimate may not be accurate and sets it at a lower value when the estimate is accurate.

Prior information in some cases may be available either from previous studies or from literature ,and it helps in making important inferences about the unknown parameters. It is hence advantageous to include the prior information in the form of prior distribution if it is available to the investigator. In order to complete the specification of a Bayesian model, one must specify both the prior distribution and the likelihood. When the two components have been specified, we can analyze the posterior distribution using density plots and descriptive measures. On deriving the posterior distribution it is important to examine how the form of the posterior distribution needs to be evaluated (Lawson 2013). Examining the posterior distribution via posterior sampling gives us an idea about a variety of features of the posterior distribution. Statistical inferences can be made by examining the different characteristics of this distribution (e.g., the posterior mean, median or mode can give us an estimate of $\theta$). The variance of the distribution can tell us about the uncertainty in our estimates. Thus the Posterior distribution summarizes the current state of knowledge about all the uncertain quantities in Bayesian analysis.

Due to complexity of spatial posterior distributions, it is often necessary to use numerical approximation algorithms to evaluate parameters . In order to obtain more precise estimates of the posterior distribution many analyses make use of simulation methods that include sampling algorithms (Chen 2009). Since the data are often available

at multiple levels in disease mapping, it is convenient to have a flexible posterior sampling procedure that allows us to examine a variety of complex models. One of the commonly used simulation methods is Markov chain Monte Carlo (MCMC), where samples are taken from a distribution and each sampled value usually depends on the previous one, it thus forms an iterative simulation from a Markov chain. The MCMC methods are popular because they provide an accurate estimate of the posterior parameter quantities. Many statistical software packages like SAS, R, SPSS, etc. now incorporate MCMC methods to analyze probability models. WinBUGS is another general purpose software that uses simulation based MCMC methods to fit Bayesian statistical models (Lunn, et al 2000). It examines the models as hierarchies with parameter nodes, and each parameter is represented as a stochastic node which has a distribution, a constant node, or a logical node (Lawson 2013).

WinBUGS uses information about the likelihood and prior to sample from the posterior distribution. On obtaining large enough samples we might be able to get a very good approximation of the posterior distribution. The program in WinBUGS requires three sections: stating the likelihood and priors, reading in the observed data ,and entering a set of initial values for the parameters that gives the MCMC algorithm a set of starting values for the parameters. When estimating the posterior distribution we can use a burn-in period of initial samples that we can then discard when estimating the posterior distribution. This allows the MCMC sampling procedure to stabilize. Before collecting the samples for the posterior distribution it is important to check convergence. Convergence helps us examine if our samples are from the correct distribution. Once convergence has been achieved samples resemble a random scatter about a stable mean

value (Figure 2.1). If we are running more than one chain simultaneously, the trace and the history plots show chains in different colors (Figure 2.2), and if all chains appear to be overlapping one another then we can be somewhat certain that convergence has been achieved. We can also look at the Brooks-Gelman-Rubin (BGR) diagnostic that is based on the ratio of between-within chain variances. On convergence the ratio of between-within chain variances converges to 1.

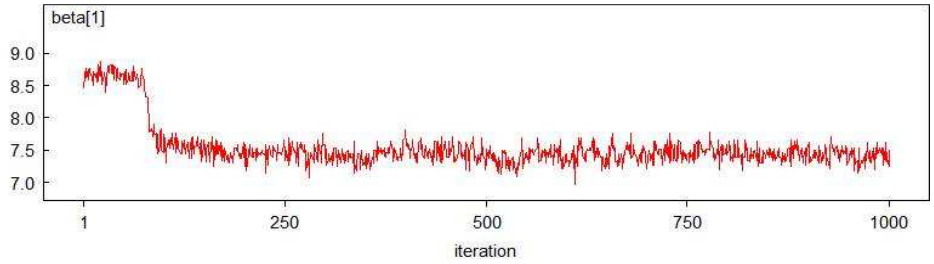


Figure 2.1. Convergence plot for one chain

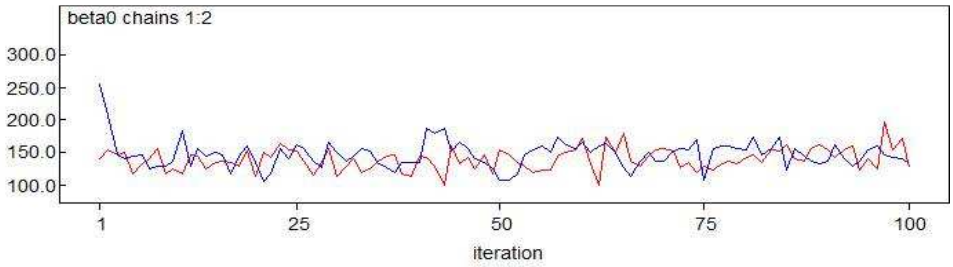


Figure 2.2. Convergence plot for two chains

At times convergence is not easily achieved; some of the reasons could be that the model may not be fitting well, there could be errors in programming like syntax mistakes, or the starting distribution or values are causing slow convergence. In order to improve the convergence process, we can standardize all the variables by subtracting them from their sample means and then dividing by their standard errors, this may speed up the convergence by decreasing the posterior correlation between parameters.

Often if the initial values are near the posterior nodes convergence occurs more quickly, so we may want to select good initial values. Often the models just take a long time to converge so before we implement any other technique to get better convergence we can just let the model run for a longer time. After convergence is achieved, further iterations are required to obtain samples for posterior inference. Since MCMC chains are dependent samples, the autocorrelation will influence how many iterations are needed. Typically the mean or the median is reported for each parameter of interest as the point estimate. The 2.5% and 97.5% percentiles of the posterior sample for each parameter give the 95% posterior credible interval. Based on the summaries of the posterior distribution, we can address questions by investigating the sample.

## 2.1 Bayesian Spatial Models

The conditional nature of geographic data in disease mapping makes it suitable to a Bayesian hierarchical model, with parameter estimation accomplished via MCMC methods. The spatial correlation relates to the idea that areas close to each other in space

will experience similar outcomes; this correlation must be accounted for in spatial analyses. A simple way to incorporate the correlated/uncorrelated random effects in the analysis is to include them in the model. The random effects are commonly represented as $u_i$ and $v_i$ in spatial models. $u_i$ represents the spatial random effect that captures the spatial dependence between areas ,and $v_i$ is the independent random effect that captures geographically unstructured heterogeneity among areas (You, Zhou 2011). A flexible way for the inclusion of these terms is to include a log-linear term with additive random effects (equation 4) (Lawson 2013).

$$\exp\{x_i^{`}\beta + u_i + v_i\} \quad (4)$$

where $x_i^{`}\beta$ is the covariate component ,and $u_i$ and $v_i$ are correlated and uncorrelated heterogeneity, respectively. Both of these random effects have separate prior distributions. The spatial random effect $u_i$ usually has an intrinsic Gaussian (CAR) distribution or a fully specified multivariate normal prior distribution whereas the independent random effect $v_i$ has a gamma or beta prior.

Often there is unobserved confounding present within a study area ,and it may not always be possible to obtain prior information about these unobserved effects. Thus, it is necessary to include both of the correlated and uncorrelated random effects in the model. The sum of these effects is an important component because we are interested in the total effect of the unobserved confounding. A CAR model fits spatially correlated effects well. It's also possible to include uncorrelated effects in the same CAR model, such a model is called the *convolution* model. These *convolution* models employ a common disease mapping technique that is used to obtain smooth relative risk estimates. This method

helps in reducing the variance in the estimates by borrowing information from neighboring areas and aids in producing more stable estimates. This is particularly useful for data from small geographical areas where the risk estimates have larger standard errors due to the small sample sizes as compared to the larger areas.

In WinBUGS the CAR model is fitted as the car.normal( ) distribution. It requires an adjacency vector and a list of the number of neighbors for each region. Equation 5 gives the common syntax for the CAR distribution in WinBUGS, where *u* is the spatial

$$u[1:m] \sim car.normal(adjc[], weic[], numc[], tauU) \quad (5)$$

random effect and *m* is the number of counties or Public Health districts. Once the model has been specified and fitted, we can output parameter estimates and statistics derived from the chains such as quantiles. We can also obtain the Deviance information criteria (DIC) for the models which we can later use as a comparison tool for the goodness-of-fit for the models.

Using the simulation capabilities of BUGS packages with R, we can evaluate the statistical properties of different models under different scenarios. These scenarios include using risk estimates (or other parameter estimates) in the model that are different than the ones observed. This would give us an idea of how the statistical properties of models may change when put to test under various scenarios.

# CHAPTER 3

# SCALE CHANGE ISSUE

The modifiable areal unit problem (MAUP) that arises in spatial analysis uses units of analysis at aggregations higher than incident level. The MAUP often arises when aggregate units of analysis are arbitrarily produced by the study investigator. A special case of the MAUP is the concept of ecologic fallacy, where the issue arises of making inferences at an individual level from aggregate data. In addition to the inferences based on such data, the MAUP also affects the analysis, including correlation and regression (by sometimes inflating the coefficients (Armheim 1995)).

The MAUP can be broken up into two major effects- *the scale effect* and *the zoning effect*. The term was first introduced by Openshaw & Taylor (1979) to describe the effects of these two problems. The *scale effect* is concerned with how the scale of any given data impacts the way in which they are analyzed and interpreted. The *zonal problem* involves keeping the same scale of research but changing the actual shape and size of the areas of research (Jones 2011). Our study deals with the *scale change issue*, which is known as the *change of support problem* in Geostatistics. The study focus is primarily on multiple scale analysis for disease data using the Georgia Oral Cancer data set that includes 18 Public health districts and 159 counties, as an example.

## 3.1 Background on Scale Change Issue

The *Change of support problem* or the *scale change problem* is concerned with different statistical inferences and estimates that are generated by the same data when it is aggregated into different spatial resolutions, especially when small areas are aggregated into a larger unit. Some studies have suggested methods to incorporate the *scale effects* within a Bayesian framework (Louie and Kolaczyk 2006; Lee et al., 2009).

In a study conducted by Gelfand et. al. (2005), a Bayesian Kriging approach was used for the *change of support problem* using spatio-temporal data. They investigated the relationship between ozone exposure and pediatric asthma. Their data set consisted of ozone levels in Atlanta, Georgia metropolitan areas. The ozone measures were available from 8 to 10 monitoring sites ,and the relevant health outcome data was available only at the zip level. In order to assess the association between pediatric asthma and ozone exposure, the issue of the mismatch for the support of these two variables (zip level scale for asthma and ozone level scale for ozone exposure) had to be taken into account. They used noninformative priors within the Bayesian Gaussian process so that their results would resemble those of a likelihood analysis. They also proposed a simulation approach that would allow prediction from points to points, points to blocks, blocks to points and blocks to blocks data (Gelfand 2000) using simulation based models. Overall their methods included the Kriging approach within the Bayesian framework.

In addition to the Kriging procedure another common approach to the *Change of support problem* in linear spatial statistics is to model the spatial autocorrelation of the variable observed at different spatial scales. Lee et. al. (2009) presented a novel approach

for the *Change of support problem* using nonlinear Bayesian Maximum Entropy (BME) approach which is an extension of linear spatial statistics ,and that provides a nonlinear integration of data reported at different scales. They applied their approach to the problem of mapping childhood asthma across North Carolina that included prevalence data aggregated over counties together with the data obtained at the school district levels. Their work provides a methodology that complements the area-to-point Kriging method which involves integration of disease prevalence data collected at different scales. The data observed at large scales (at county level) was modeled as soft data ,and the survey data collected as a part of North Carolina School Asthma Survey (NCSAS) was modeled as hard data.

The BME approach they presented included two stage procedures: the prior and the posterior stages. The prior stage includes the incorporation of general knowledge which may include summary statistics, scientific laws, etc., and are usually computed from any prior information available to the investigator (like the prior distribution). The posterior stage includes incorporating specific knowledge that contains the on-site measurement (including the hard and soft data). One of the three methods of analysis they used accounted for the combination of the hard and soft data ,and were analyzed in a linear spatial Gaussian model. They found that this method led to more precise estimation of the spatial distribution of childhood asthma prevalence as compared to the other methods they used which did not account for these scale effects (of hard and soft data) together.

An aspect of their work that can be investigated is the small number problem or small area problem that leads to noisy spatial distribution of observed disease rates. Their

study does not address this issue because the prevalence of asthma among children in North Carolina is comparatively higher than prevalence of other diseases and hence is not classified as a rare disease. The small area problem has been discussed widely in many studies and most of the results point towards a Bayesian framework to smooth out the noise due to small numbers. Using a similar approach as above, with Bayesian hierarchical modeling we can account for the different scale effects for disease counts in small areas. The following section describes the problem formulation for my thesis and the potential approach to address this problem based on previous methods.

## 3.2 Problem formulation

Changing the support of a variable typically creates a new variable by averaging or aggregating over an existing one. The new variable is related to the old one but has different spatial and statistical properties that need to be accounted for. The concept of *scale change* plays an important role in analysis and inferences, especially when multiple scales are present in the data. It is also important to account for the linkage or the relation between these scales in the analysis. Many Public Health policy decisions are based on statistical associations obtained from the analysis of spatial data available at multiple scales, it is therefore important to account for the these effects in analysis. For a given data set there could be multiple aggregation levels ,and generally these levels are nested in a hierarchical way. Multiscale analysis is thus needed to quantify the useful information that may be present in the data at more than just one level of aggregation.

There has been moderate consideration in previous studies and literature regarding the roles that aggregation and scale effects play in spatial analysis. Several studies have used the Kriging and Cokriging approach to address this issue. An alternative approach, which is an extension and variation of these methods, is to use Bayesian Hierarchical models for Multiscale data. This method allows for the spatial dependence as well as accommodates covariates related to these in the models. It also helps us deal with the problem of overdispersion in modeling count data for small areas by allowing us to include random effects. These random effects also account for any unobserved confounding in the data.

Louie and Kolaczyk (2006) introduced a Bayesian framework for count data that helps in generating informative disease maps across multiple scales for a given data set and it also helps in describing the distribution of disease risk for these hierarchical scales. They used an extension of the Poisson-based multiscale spatial process models for count data in order to create disease maps. They developed two multiscale SMR estimation strategies that were distinguished by whether the hyperparameters are random or fixed. Using empirical Bayes estimation, they considered fixed hyperparameters, and using a fully Bayesian inference, they examined the use of random hyperparameters. They also computed relative risks under a hierarchical Bayes Poisson lognormal model where $Y_k \sim Poisson(\mu_k)$ and the log spatial mean is expressed as $(log\ \mu_k = log\ e_k + \alpha + u_k + v_k)$ where $\alpha$ is the overall log relative risk and $u_k$ (with Gaussian prior) and $v_k$ (with CAR prior) are the random effects that capture the unstructured heterogeneity and extra-Poisson variation, respectively. They used the IMSE criteria (integrated mean square error) to compare their three SMR estimators. Their results indicated that the empirical

Bayes estimator for SMR is better than the fully Bayesian estimator where the risk is elevated. They also noted that when the spatial patterns of risk elevation were nested the multiscale estimators are better than the log normal model but when the spatial patterns were not nested the log normal model performed better.

In another study conducted by Louie and Kolaczyk (2006), they presented a framework to model multiscale incidence data that is an extension of their above study of multiscale disease mapping. The basic idea of this study was to compare Poisson models within a collection of hypotheses to find patterns in localized disease variation. They suggested that this could be accomplished computationally using specified values for the hyperparameters and prior probabilities given by the investigator for hypotheses. They used two set of tests: one aimed at detecting arbitrary deviations from uniformity and the other test was aimed at detecting local elevations in risk within the Bayesian framework. The multiscale testing approach they use is analogous to the Bayesian wavelet shrinkage method that is based in coefficient by coefficient testing in the space of wavelet coefficients of observed data. Their methods focused on detecting potential anomalies in aggregate disease incidence data to find deviation from the uniformity in relative risk taking into account the expected counts. The method they propose can also be used to identify the locations and scales of isolated disease clusters within a spatial region.

The framework they present on Bayesian disease mapping can be further extended for estimating relative risk in Poisson based models for count data. Our study is motivated by their use of Poisson log normal models in estimating the relative risk for data available at multiple scales. Many studies have looked at multiscale modeling of

spatial data within the Bayesian framework, but few of them address the multiscale issue in small areas (Louie and Kolaczyk 2006; Lee et al., 2009; Louie and Kolaczyk 2004; You and Zhou 2011). Also, the methods proposed by many of the previous studies involve complex statistical and computational techniques that may not always be easily accomplished. Taking into account the need for simpler and more user friendly methods, we propose the implementation of simple log linear Poisson models to fit multilevel data ,and to obtain smoother risk estimates for small areas or counts.

This thesis provides a framework to address the effects of scale on the analysis of spatial data in particular the Georgia Oral Cancer data ,and also presents potential ways of describing the distribution of relative risk across a hierarchy of multiple scales. Within the Bayesian framework and with the help of simulation studies we explore a methodology that allows us to estimate and compare measures of relative risk in Poisson based convolution models.

# CHAPTER 4

## SIMULATION STUDY

### 4.1 Motivation

We conducted a simulation study in BRugs to illustrate the potential of our multiscale models in integrating the two scale levels for the Georgia data set. The simulations were conducted under two scenarios of relative risks to examine how the fitted models perform under these specified conditions. One of the objectives of the analysis is to model the relative risk in our study area by linking the two scale levels in the model, and also to analyze the two scales separately in a model. We also included correlated and uncorrelated random effects observed at these aggregation levels. The addition of these components captures the unobserved confounding effects and the uncertainty related to the models.

Some of the standard methods used to estimate relative risk rely on smoothing techniques that involve additional assumptions or include additional model components like random effects. In a multiscale models study by Louie et. al. (2004), a framework was introduced that allows one to derive an interrelated sequence of informative disease and confidence maps across a hierarchy of multiple spatial scales. They illustrated their methodology using tract count data on Gastric cancer in Tuscany. They used the multiscale extensions approach of the canonical SMR statistic to estimate measures of

relative risk in Poisson-based models. Most of the previous efforts in multiscale analysis have been directed towards incorporating some sense of spatial distribution so as to control for the spatial variation in the 'at-risk' population. Our current work is an extension to previous approaches where we fit two separate models for the two different levels of the data and also fit joint models with contextual effects.

## 4.2 Georgia Oral Cancer Data

The Georgia data set contains the counts of oral cancer deaths as well as the expected rates for the 18 PH districts and 159 counties from 2004. Each of the 159 counties falls uniquely and completely within a given PH district due to which we might observe some grouping effect for the counties that lie within a district. These PH districts are the administrative units that provide health services. The Oral cancer deaths are the outcome of interest. Figure 4.1 shows the 18 PH districts and the 159 counties of the state of Georgia in the US. The data can also be found online from OASIS which gives access to datasets from the Georgia Department of Public Health (http://oasis.state.ga.us/). The expected rates for this data were calculated from the statewide rates and were applied to the counties and PH districts. The expected rates for the counties range from 0.089 to 38.19 with a mean of 2.59 ,and the expected rates for the districts range from 6.65 to 40.28 with a mean of 23.

This spatial data is observed over a regularly spaced set of points, i.e., it has a regular lattice structure ,and has a reasonably large set of counties. These counties have relatively small areas and the disease counts are measured in these areas. Since the

counties have similar spatial shapes we can consider diverse designs for the spatial clusters, thus making this data a good test bed to evaluate the multiple scale effects.

The data is reported at two aggregation levels: the county level and the PH district level, with $y_i^c$, i= 1,….,159 and $y_j^{PH}$, j=1,….,18. In this case $y_j^{PH} = \sum_{i \epsilon j} y_i^c$ where $y_i^c$ is the county level count of disease and $y_j^{PH}$ is the PH district level count of disease that is aggregated over the counties. There are several potential ways of linking these two scales in the models. In our study we analyze five of these variations that are described in the following sections.

## 4.3 Models Considered

We used the Bayesian modeling approach and compared models using DIC. It is commonly assumed that counts of a disease within small areas have Poisson distribution with mean $e_i\theta_i$ . So $y_i^C$ has a Poisson distribution with expected value $e_i\theta_i$ , i.e., $y_i^C \sim$ Pois($e_i\theta_i$) where $e_i^c$ is the expected county level count of the disease and $\theta_i^c$ is the relative risk for the counties. The counts have a joint probability of arising based on the likelihood $L(y, \theta)$, which are the Poisson probabilities for each of the regions. It also gives information on how likely the data are given the expected rates $e_i\theta_i$ ,and gives us the most likely values for $\theta$ for the observed data set.

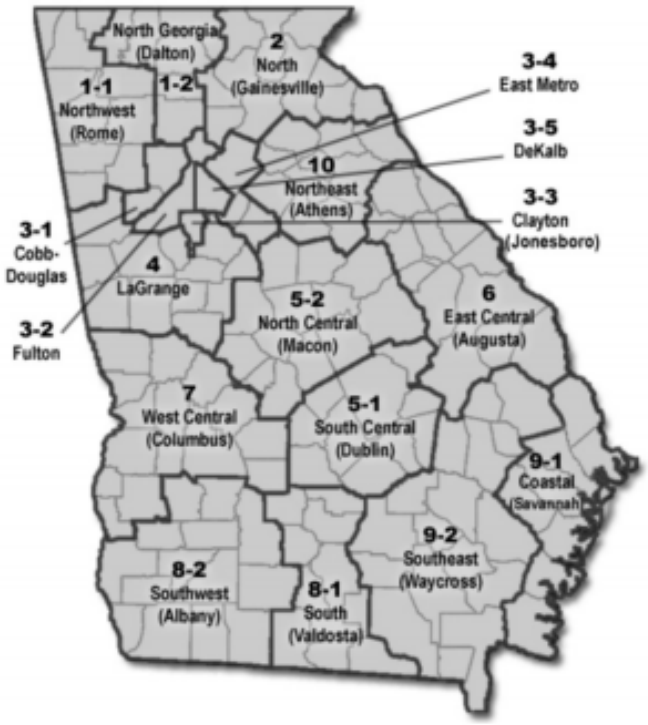


Figure 4.1 State of Georgia, US: Public Health district and country boundary map (Lawson 2013)

A useful property of the Poisson distribution is that the sum of independent Poisson variables also has a Poisson distribution. In the case of this data, since the $y_j^{PH}$ are aggregated over the $y_i^C$, the $y_j^{PH}$ disease counts also have a Poisson distribution. The Georgia data is defined at two scale levels, with $y_i^c$ the county level and $y_j^{PH}$ the PH district level. The general model could thus be written as $y_i^C \sim f_1(\mu_i^C;\ l_1)$ where $l_1$ is the county level, or $y_j^{PH} \sim f_2(\mu_j^{PH};\ l_2)$ where $l_2$ is the PH district level. In order to ensure that the county level and the PH district levels are linked, we need to model the joint

behavior of $\mu_i^C$ and $\mu_j^{PH}$. Also since there is a dependence between the mean levels of the two scales the joint effect of their latent factors must also be considered (Lawson 2013). To model this dependence we used five variants of the Poisson log-linear model (described in Table 4.1).

The log spatial mean, $\mu_i^C$ and $\mu_j^{PH}$ for all the five models, is expressed as $\log \mu_i^C = \log e_i^C + \alpha_0^C + u_i^C + v_i^C$ for the county level and $\log \mu_j^{PH} = \log e_j^{PH} + \alpha_0^{PH} + u_j^{PH} + v_j^{PH}$ for the PH district level. Here the intercepts $\alpha_0^C$ and $\alpha_0^{PH}$ capture the overall level of the log relative risk, and the terms *u* + *v* are the convolution of the correlated and uncorrelated heterogeneity random effects, respectively. The log expected counts ($e_j^{PH}$, $e_i^C$) are included as offset term to model the relative risk. The mean $\mu = e\theta$ in used in all our models and represent the expected counts, which typically should be non-negative. We thus model the log of the mean using a linear model. A common method to get a smoother distribution for the risk is to assume that the risk $\theta$ has a distribution (termed as the prior distribution in Bayesian statistics). In case of Poisson models it is common to assume that $\theta$ has a Gamma distribution (Lawson 2013) so we used a Gamma prior with known shape and rate parameters in the five linkages we used.

The spatial random effects ($u_i^C$, $u_j^{PH}$) are based on the neighborhood adjacencies and they capture the spatially correlated effect found in the outcome. They are assumed to have a Gaussian CAR prior distribution in all our models i.e., *$u_i^C$[1:159]~car.normal(adjc[] ,weightsc[], numc[],tau.u)* and *$u_j^{PH}$[1:18] ~ car.normal(adjph[],weightsph[],numph[],tau.uph).* Here we define an adjacency list *adj*

which is a matrix containing lists of neighbours obtained from the mapping tool in WinBUGS, *num* contains the number of neighbours for each region and *weights* which contains the CAR weights which is commonly assumed to be equal to 1. The uncorrelated random effects encapsulate the extra-variation in the model and are included with a zero mean Gaussian prior i.e., $V_i^c \sim N(0, \tau_v)$ and $V_j^{PHc} \sim N(0, \tau_{vph})$ in all five models. These uncorrelated effects capture the unstructured heterogeneity in the geographical areas of interest.

We use the five variations in linkages to analyze potential ways in which we can link the two aggregation scales. Since the two scales have a hierarchical nature, it is important to include the influence the two scales have on each other, whether in the form of shared random effects or aggregated counts. In model 1 we examine the simplest way to incorporate the district effect into the county level effect. The aggregated county expected values are used to estimate the risk in the districts and the risk in counties is estimated using the basic convolution model with uncorrelated and correlated random effects that may capture any extra-variation in the model.

In model 2 we fit two convolution models for each of the two scales. The two levels are joined through the aggregated county expected values that are included in the PH part of the model. This may account for the county effects within a given district.

In model 3 we link the district component to the county component by including the correlated spatial random effect from the districts in the county part of the model. This model for the county takes into account both the correlated spatial effects within the

counties and also the spatial effects from the districts. We call it the shared component model because the two model parts share the spatial random effect component for districts.

In model 4 we make use of the convolution models for the two levels. The levels are linked by the spatial effect from the district that's incorporated in the county component of the model. This may allow us to analyze how the fit statistics are affected when the spatial effects for county are not added in the model but only the spatial effects for the districts that contain these respective counties are taken into account. This model is a variation of model 3.

Model 5 makes use of two separate models for the two scales and no shared component is included in either parts of the model. This model helps us examine how the fit statistics and the variability of $\theta s$ are affected when the influence of the linkage of the two scales is not taken into account. Below is a further detailed description of the five models in terms of the parameters fitted and the different ways in which the two scales have been aggregated (also described in Table 4.1).

1) Model 1

In model 1 we compute the relative risk for the disease through the log of $\theta_i^c$ for the county part. We make use of the basic convolution model without covariates ,and which includes the intercept; the uncorrelated random effects intercept ($V_i^c$) and the correlated spatial random effect term ($U_i^c$). $y_i^c$ is modeled as a Poisson distribution i.e. $y_i^c \sim pois(e_i^c * \theta_i^c)$ in all the five models considered. The expected value of the count of the

disease is considered to be a multiplicative function of the expected count and the relative risk $(e_i^c * \theta_i^c)$. This is the simplest model of the five considered, since the relative risk for the PH district part is simply aggregated over the county estimated effects. It does not estimate the PH level separately.

2) Model 2

This model is similar to model 1 with respect to the county part of risk estimation but is different from model 1 in the PH district part. The relative risk ($\theta_j^{PH}$) for the public health district is estimated separately using the convolution model including the intercept, random effects term ($V_j^{PH}$) and the spatial random effects term ($U_j^{PH}$). So $\theta_j^{PH}$ has been fitted as $\log \theta_j^{PH} = (\alpha_0^{PH} + V_j^{PH} + U_j^{PH}$, and $\mu_j^{PH}$ ) has been aggregated as $\mu_j^{PH} = (\sum e_i{}^c).\theta_j^{PH}$.

3) Model 3

The county component of model 3 also uses the convolution model for risk estimation. It is however different from models 1 and 2 since the county part includes the spatial random effect $U_j^{PH}$ from the PH district part i.e. $\log \theta_i^c = (\alpha_0^c + V_i^c + U_i^c + U_j^{PH})$, where $U_i^{PH}$ is the shared component. For the PH part of the model $\mu_j^{PH}$ has been fitted as the product of $e_i^c$ and $\theta_i^c$ i.e. $\mu_j^{PH} = e_j{}^{PH}.\ \theta_j{}^{PH}$ and $log\ \theta_j^{PH} = (\alpha_0^{PH} + V_j^{PH} + U_j^{PH}$ ) remains the same as model 2. Model 3 is thus a joint model that accounts for the contextual effects of district on county.

4) Model 4

This model is similar to model 3, except that in the county part we have $\log\theta_i^c = (\alpha_0^c + V_i^c + U_i^{PH})$, where $U_i^{PH}$ is the shared component and the term $U_i^c$ has been eliminated from the

model. Thus, only the correlated spatial effects from PH district are accounted for in the county part. The remaining parts of this model are same as model 3.

5) Model 5

This model is comparatively different than the other fitted models. It can be considered as the null model since we have two separate models for the two different scale levels and there is no linkage in these levels for this model. For the county part we have $\theta_i^c$ fitted simply as $\log\theta_i^c=(\alpha_0^c+V_i^c+U_i^c)$ and $\mu_j^c$ fitted as $\log \mu_i^c=log\ e_i^c + log\ \theta_i^c$. For the PH district part $\theta_j^{PH}$ is fitted as $\theta_j^{PH} =exp(\alpha_0^{PH}+V_j^{PH}+ U_j^{PH})$ and $\mu_j^{PH}$ is fitted as $\log \mu_j^{PH}=log\ e_j^{PH}+log\ \theta_j^{PH}$. This model thus has no shared components.

The above five models evaluated differ in the way their $\theta_j^{PH}$ and $\theta_i^c$ have been aggregated and estimated with the contextual effects, along with the mean level of outcome $\mu_j^{PH}$ or $\mu_i^C$. The analysis of these models was carried out through simulations in BRugs.

## 4.4 Simulation Aim

The simulations were set up to evaluate the goodness of fit for the five models representing the five variations in the linkages of the two scales. The goodness of fit in this sense is to compare the proportion of the observed disease counts in the predefined sample to the expected counts from the sample under a specified probability model. This includes testing the effects on the measurement of the fit statistics (in particular the DIC) when different models are assumed for the scale effects. The summarized results of the simulations can also be used to obtain plots to examine how well the two levels fit. Some

other statistics like the variability for the relative risk $\theta$ can also be computed to compare the five model fits.

## 4.5 Simulation Settings

Using the simulation facilities in R we carried out the model fitting procedure for the five models. We assumed in general that $y_i$ ~Pois($\mu_i$) at the first hierarchical level and that $\mu_i$ = ($e_i$ $\theta_i$), where i =counties. At the next hierarchical level we assume that $\theta_i$~ Gamma(a,b), which is the prior distribution given to $\theta$, with shape parameter a and rate parameter b.

For setting up the simulations we used a *True* model and a *fitted* model. We simulated 500 disease counts from the *True* model for the 159 counties. Thus the simulated values (synthetic data called ysim in the code) consisted of a 500 by 159 matrix in R. The *True* model form is assumed to have a Poisson distribution with mean $\mu_i$ ,and $\mu_i = (e_i^c * \theta_i^T)$. The $e_i^c$, which are the expected counts for the counties, are taken from the Georgia data set ,and the $\theta_i^T$ is the *True* $\theta$ (relative risk) that is computed in R using the *rgamma* function. We examined the model performances for two different sets of shape and rate parameters for the gamma distribution of $\theta$. The first set of $\theta_i^T$ was computed using a shape and rate parameter of 1 ,and the second set of $\theta_i^T$ was computed using a shape and rate parameter of 3, through the *rgamma* function. The rgamma function is defined in R as *rgamma(n, shape, rate, scale = 1/rate)* when we do not define the scale parameter, it is assumed to be equal to 1. So when we use a shape and rate parameter of 1, the gamma distribution for the first set of $\theta_i^T$ has a mean of 1 and

variance of 1. However when we use a shape and rate parameter of 3 the gamma distribution for the second set of $\theta_i^T$ has a mean of 1 and a variance of 0.33 (or 1/3). Thus the two set of $\theta_i^T$ have the same mean and different variances.

Thus, two *True* model forms were established to get the simulated data sets. For the second set of $\theta_i^T$ (with shape and rate parameter 3) I however computed 100 disease counts to examine the fitted models. The basic idea for these simulations is that we subject the models to different scenarios to get an idea about how these models account for the scale effects and how well they fit given the data. In this case the 500 by 159 different disease counts together with the two different estimates for *true* risk form the scenarios.

The simulated data sets were then passed through the five models (considered as *fitted* models, described in section 4.3) to get the fitted values for relative risk $\theta^F$. The results from the simulations were summarized over 500 runs in case of the first set of $\theta_i^T$ and over 100 runs in case of the second set. The summarized results included mean DICs (Deviance information criteria), pDs (the effective number of parameters) and the fitted $\theta^F$. The variability for the $\theta$s was calculated using the $\theta_i^T$ and the $\theta^F$. Five hundred $\theta^F$ values were obtained for the first set of simulations and 100 were obtained from the next set of simulations. The $\theta^F$ were computed for the counties and the corresponding $\theta^F$ for the PH districts were calculated from the aggregated county values. For computing the difference between the $\theta^T$ and the $\theta^F$, I obtained an average value for $\theta^F$ for the counties and the average value for $\theta^F$ for the districts. The difference of these values was squared

and averaged over the number of counties and districts. This variability between the $\theta^T$ and the $\theta^F$ gives an idea of how the five models differ in the way they capture the true risk in the study area. The model with lower variability is the preferred model. The posterior average maps were obtained for $\theta$, $u$ and $v$ for these models. These maps highlight the observed variation in the relative risks across the counties and the districts.

## 4.6 Computational approach

All of the above simulation procedures were carried out in BRugs, which is a collection of R functions that allows users to graph models using MCMC methods. Over the years R has proven to be a useful source for statistical computing and visualization. It enables investigators to solve complex and sophisticated problems along with routine analysis. A variety of statistical procedures in the form of packages is freely available and helps us to integrate R with other packages. One such package is BRugs that integrates R and OpenBUGS.

The BRugs function can be split into two groups: those associated with the setting up and simulating the graphical models, and those associated with statistical inferences. The package implements OpenBUGS on R. Each of the processes - model checking, compiling, reading in the initial values and reading in the data - that are used to run a model in OpenBUGS can be called from within R (Kerman and Ligges 2013). A few lines of code in R can set up these functions. R2WinBUGS is another package that uses MCMC algorithm and allows us to run simulation based models. BRugs and R2WinBUGS are similar in their basic structure of coding and running processes through

R. But one of the advantages of BRugs over R2WinBUGS is that it does not essentially have to be run under Windows. Also BRugs uses plain text files for data input whereas R2WinBUGS uses compound documents (Kerman and Ligges 2013). BRugs also carries out the simulations at a faster rate and is able to handle large datasets well as compared to R2WinBUGS.

All of our simulations were carried out through R in connection with OpenBUGS. Before fitting the models for simulations we performed a preliminary evaluation of the models in WinBUGS (WinBUGS and OpenBUGS provided similar statistics). This was done to check if there are any issues with the model to compile and if the models achieve convergence at a reasonable number of iterations. This procedure is further detailed in the Results section. The BRugsFit function in R was used to specify the number of iterations needed till convergence and number of samples (nburnin) required after convergence to make inferences about the model fit.

In addition to the simulations the posterior average maps were also generated using the map function in R along with GeoBUGS. The aggregated results of these simulations were used to make model based inferences. Chapter 5 presents the comparison and discussion of these results.

Table 4.1 Models Fitted

| Model number | Parameters Included |
|---|---|
| 1 | County level:<br>$\log \theta_i^c = \alpha_0^c + V_i^c + U_i^c$<br>$\mu_i^c = e_i^c \cdot \theta_i^c$<br><br>PH district level:<br>$\theta_j^{PH} = \mu_j^{PH} / e_j^{PH}$<br>$\mu_j^{PH} = \sum e_i^c \cdot \theta_i^c$ |
| 2 | County level:<br>$\log \theta_i^c = \alpha_0^c + V_i^c + U_i^c$<br>$\mu_i^c = e_i^c \cdot \theta_i^c$<br><br>PH district level:<br>$\log \theta_j^{PH} = \alpha_0^{PH} + V_j^{PH} + U_j^{PH}$,<br>$\mu_j^{PH} = (\sum e_i^c) \cdot \theta_j^{PH}$ |
| 3 | County level:<br>$\log \theta_i^c = \alpha_0^c + V_i^c + U_i^c + U_j^{PH}$<br>$\mu_i^c = e_i^c \cdot \theta_i^c$<br><br>PH district level:<br>$\log \theta_j^{PH} = \alpha_0^{PH} + V_j^{PH} + U_j^{PH}$<br>$\mu_j^{PH} = e_j^{PH} \cdot \theta_j^{PH}$ |
| 4 | County level:<br>$\log \theta_i^c = \alpha_0^c + V_i^c + U_j^{PH}$<br>$\mu_i^c = e_i^c \cdot \theta_i^c$<br><br>PH district level:<br>$\log \theta_j^{PH} = \alpha_0^{PH} + V_j^{PH} + U_j^{PH}$<br>$\mu_j^{PH} = e_j^{PH} \cdot \theta_j^{PH}$ |
| 5 | County level:<br>$\theta_i^c = \exp(\alpha_0^c + V_i^c + U_i^c)$<br>$\log \mu_i^c = \log e_i^c + \log \theta_i^c$<br><br>PH district level:<br>$\theta_j^{PH} = \exp(\alpha_0^{PH} + V_j^{PH} + U_j^{PH})$<br>$\log \mu_j^{PH} = \log e_j^{PH} + \log \theta_j^{PH}$ |

**Notations used in the above models:**

i – 1:159 – number of counties (observed count, 159)

j-1:18 – number of PH districts

$y_i^c$ (yc)- county level count of the disease

$y_j^{PH}$ (yph)- PH district level count of the disease

$e_i^c$ (ec)– expected rate/count of disease (population at risk)- county level

$e_j^{PH}$ (eph) – expected rate/count of disease (population at risk) - PH district level

$\theta_i^c$ (thc)– is the relative risk (>0) (term we model for) at county level

$\theta_j^{PH}$ (thph)– is the relative risk (>0) (term we model for) at PH district level

$\alpha^c_0$- intercept for county part- gives the overall rate for the disease

$\alpha^{PH}_0$ -intercept for district part

$\mu_i^c$ – mean level of outcome at the county level

$\mu_j^{PH}$ – mean level of outcome at the PH district level

$U_i^c$ - is the spatial random effect for counties

$V_i^c$ - is the uncorrelated random effect introduced in the model for counties.

$U_j^{PH}$ - is the spatial random effect for PH districts

$V_j^{PH}$ - is the uncorrelated random effect introduced in the model for PH districts

# CHAPTER 5

## RESULTS AND DISCUSSION

### 5.1 Preliminary Analysis

In order to make a preliminary evaluation of the goodness of fit for the five models, I estimated the model fit statistics in WinBUGS for a single run. This evaluation was also done to check if the models compile and if they achieve convergence before they are fitted in BRugs for simulations. Convergence refers to the idea that eventually under regularity conditions the sample converges to the distribution of interest. Thus before we summarize the simulated parameters, it is important to ensure that the chains have converged. There is, however, no formal way of confirming convergence in WinBUGS; we can't determine for sure that the parameter has converged but we can determine if the parameter has not converged.

In order to check for convergence we looked at the trace plots and the BGR statistic plots. The trace plots of sample versus the simulation index are useful in assessing convergence. The trace can tell us whether the chains are mixing well ,and if it has reached to its stationary distribution. Multiple chains can also be examined to check for convergence. We looked at two chains in case of our models and observed their trace plots. This was carried out for the parameters $\alpha_0^c$, $\alpha_0^{PH}$ ,$V_i^c$, $U_i^c$ ,$V_j^{PH}$, $U_j^{PH}$, $\theta_j^{PH}$ and $\theta_i^c$ for all the five models. Since the chains for the trace plots of these parameters appeared

to be reasonably overlapping each other we can assume that convergence has been achieved.

The BGR diagnostic plot for the above parameters was also assessed. The BGR statistic assesses the variability within parallel chains as compared to variability between parallel chains. The model is converged if the ratio of between to within variability is close to 1. The green line in the BGR plots represents the between variability, the blue line represents the within variability ,and the red line represents the ratio. If the red line is close to 1 and the blue and green lines are stable across the width of the plot, then we can reasonably assume that convergence has been reached. The figures below show the trace plots (5.1-5.5) and BGR statistic plots (5.6-5.10) for parameters like the $\alpha_0^c$, $\alpha_0^{PH}$ of the five models (model 1 just has $\alpha_0^c$ as the intercept). These plots show the 5000 samples collected after convergence has been achieved after 10,000 iterations for all models; in some models, more than 10,000 iterations were needed to reach convergence. After convergence further iterations are needed to obtain samples for posterior inference. The more the number of iterations the more precise the posterior estimates. We ran the models for about 5000 iterations after convergence to get the posterior estimates and the model fit statistics since 5000 samples seemed to be a good enough number for making inferences on the parameters.

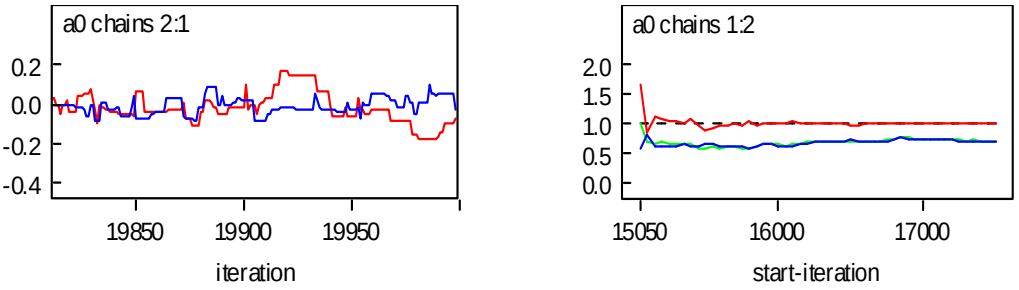


Figure 5.1 Model 1 trace plot and BGR plot

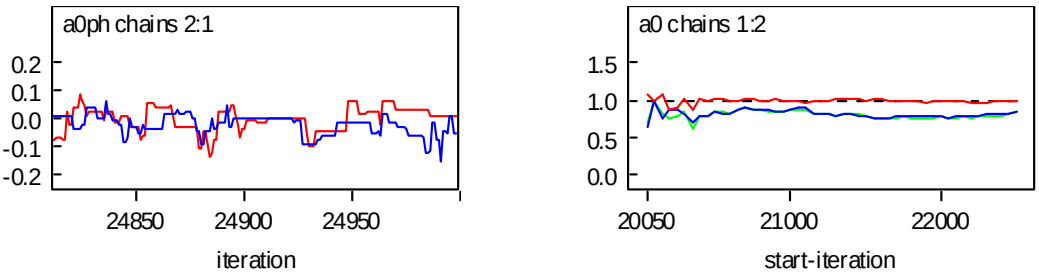


Figure 5.2 Model 2: Trace plot and BGR statistic

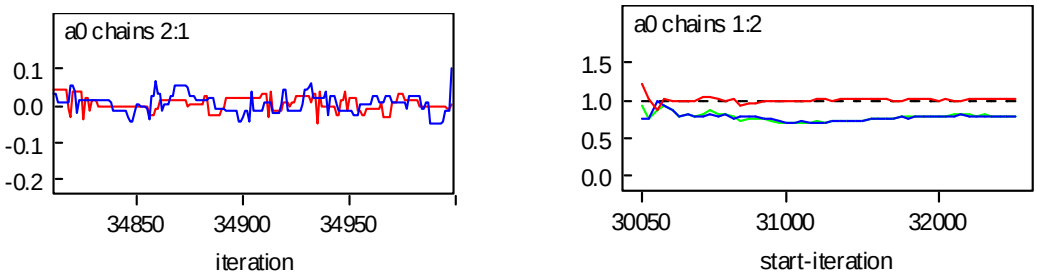


Figure 5.3 Model 3: Trace plot and BGR plot

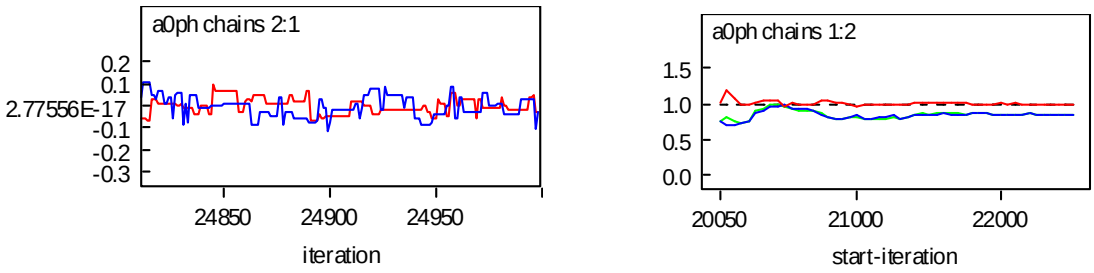


Figure 5.4 Model 4: Trace plot and BGR plot

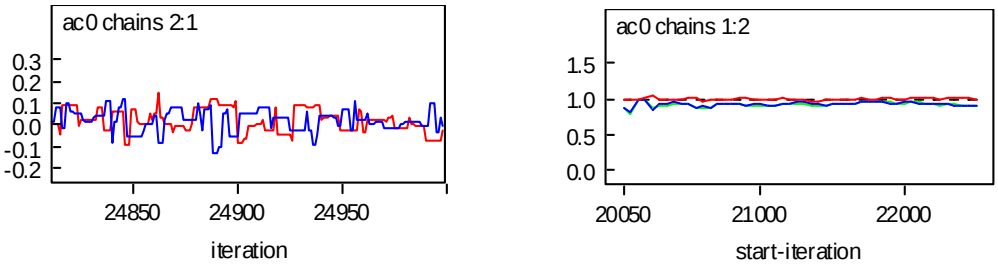


Figure 5.5 Model 5: Trace plot and BGR plot

Table 5.1 contains the fit statistics for the five models from the preliminary evaluation in WinBUGS. It states the DICs and pDs for the Public Health and the County part of the models. Ideally the DIC and the pD should be small. The DIC is defined as

$$DIC = \overline{D} + pD$$

where the $\overline{D}$ is the posterior mean of the deviance and is obtained from the DIC tool in WinBUGS. The pD is the effective number of parameters used and is also obtained from WinBUGS along with the DIC.

The lower the DIC the better the model fit. Based on this, the values from Table 5.1 indicate that model 1 is the best fitting model with the lowest DIC (481.35) for both the PH district and the counties. Models 2 (488.89) and 4 (490.67) have the next lowest DICs for the county part and also Model 4 has the next lowest DIC for the PH district. However, Model 5 has a DIC only two units higher than that of model 4 for the county part suggesting an equivalently good fit. But overall model 1 yields the lowest DIC (10 units lower) suggesting that it is the better fitting model as compared to the others.

## 5.2 Simulation study results

In our simulation study all the Bayesian inferences are based on minimum 10,000 iterations, following a 5000 iteration burn-in period. Using the summarized estimates from the simulations, we can examine the performance of the models under the specified conditions.

### 5.2.1 Simulation results for $\theta^T$ ~*Gamma(1,1)*

The first set of simulations was carried out for $\theta$ with shape and rate parameter 1. The results were summarized over 500 simulations for this set of conditions. In order to examine how many iterations were needed until convergence for each of the models, we performed the preliminary evaluation described earlier. The nBurnin option in BRugs was set to the number of iterations observed in the preliminary evaluations. For example, model 2 was run for nBurnin=20,000 and nIter = 5000. The nBurnin is the number of iterations run till convergence and nIter is the number of iterations used to obtain the posterior estimates after convergence. All of the five models used a nBurnin of 10,000 or more and the nIter of 5000 was the same for all the models.

The results obtained for the fit statistics from the simulations were stored in a matrix along with the parameter estimates for $\alpha_0^c$, $\alpha_0^{PH}$ ,$V_i^c$, $U_i^c$ ,$V_j^{PH}$, $U_j^{PH}$, $\theta_j^{PH}$ ,and $\theta_i^c$. Table 5.2 summarizes the results obtained from these set of simulations. It consists of the mean DIC and mean pDs for the five models and also the values for the $\theta$ variability for the two model components. These statistics were obtained over the average of 500 simulations. The average estimates for DIC indicate that Model 1 is the best fitting model

with the lowest mean DIC of 430.6 for County and 100.40 for the PH district part. The next best fitting model is model 5 with a DIC of 441 for the county part and 117.9 for the district part. The DIC for model 2 for the district component is however lower than that for model 5. Model 3 has the highest DIC of 476.8 for the county component and 123.4 for the district component, suggesting that it does not fit the data as well as the other models. Model 4 has DIC closer to model 3, with 471.4 for the county component and 119.1 for the district component.

The variability for $\theta$s of the county part and the district part were also computed for these models. It was calculated using the formula

$$Variability\ for\ \theta = \frac{\sum_{i=1}^{159}(\theta c^T - \theta c^F)^2}{159}$$

for the counties. And using the formula

$$Variability\ for\ \theta = \frac{\sum_{j=1}^{18}(\theta ph^T - \theta ph^F)^2}{18}$$

for the district part. Here $\theta^T$ is generated using the rgamma function and $\theta^F$ are obtained from the fitted models. In order to get the $\theta ph^F$ and $\theta ph^T$ we aggregated the $\theta c^F$ and $\theta c^T$ for the counties that are located within a given district. Thus $\theta^T$ is common for all five the models and $\theta^F$ are computed from the simulations for the fitted models.

$$MSE = \frac{\sum_{i=1}^{159}(\theta^T - \theta^F)^2}{159}$$ for counties

$$MSE = \frac{\sum_{j=1}^{18}(\theta^T - \theta^F)^2}{18}$$ for PH district

Based on the above formulas, Table 5.2 shows the computed variability for the $\theta$s of the five models. All of our simulations were carried out through R in connection with OpenBUGS. Before fitting the models for simulations we performed a preliminary evaluation of the models in WinBUGS (WinBUGS and OpenBUGS provided similar statistics). This was done to check if there are any issues with the model to compile and if the models achieve convergence at a reasonable number of iterations. This procedure is further detailed in the Results section. The BRugsFit function in R was used to specify the number of iterations needed till convergence and number of samples (nburnin) required after convergence to make inferences about the model fit.

In addition to the simulations the posterior average maps were also generated using the map function in R along with GeoBUGS. The aggregated results of these simulations were used to make model based inferences. Chapter 5 presents the comparison and discussion of these results.

The lower the variability for $\theta$ the better the estimator. The values from table 5.2 indicate that model 2 (0.003) is the best estimator of $\theta^{c}$ (for the county level) and model 4 (3.20) is the best estimator of $\theta^{PH}$ (for the PH district). Model 5 has the next

lowest variability (0.004) for $\theta^c$. The variability for the county component ($\theta^c$) is similar for models 3 and 4 (0.007). Although model 1 has the lowest DIC, it has the highest variability for $\theta^c$ and the third lowest variability for $\theta^{PH}$ which indicates that it may not be a very efficient estimator of the risk. Models 2 and 5 (3.5) have a similar variability for $\theta^c$ , $\theta^{PH}$ and models 3 and 4 (3.2) have a similar variability for $\theta^{PH}$. Although the overall differences in these estimates are marginal they do indicate that each of these models estimates the $\theta s$ (relative risks) differently.

### 5.2.2 Simulation results for $\theta^T$ ~*Gamma(3,3)*

Table 5.3 summarizes the results for this set of conditions. The values of DIC indicate that Model 1 is the best fitting one with the lowest DIC for both the PH component (106.4) and the county component (468.3). Model 2 however has DIC (469.1) one unit higher than that of the model 3 for the county part suggesting that model 2 nearly fits as well as model 1. It also has the next lowest DIC for the PH part (117.9). Model 4 has a DIC (470.8) that is 2 units higher than that for model 1 for counties this difference in the DIC is not reasonably large. Models 3 and 5 have comparatively higher DICs (475.7 and 483, respectively) for both the model components which indicates that they do not fit as well as the other models under these conditions for $\theta^T$.

The variability for $\theta$ indicate that model 2 (0.0012) is the best estimator of $\theta^c$ whereas model 5 (2.53) is the best estimator for $\theta^{PH}$. Model 5 can also be considered as the best estimator because it has the variability 0.0013 which is close to the variability

for $\theta^c$ for model 1. Models 1, 3 and 4 have similar variability for $\theta^c$ and $\theta^{PH}$; their values are higher as compared to models 2 and 5. These different values for variability show similar variation in the relative risk estimator $\theta$ as observed for the previous set of conditions for $\theta^T$ (consistent variation within the five values for $\theta^c$ and $\theta^{PH}$ for the two set of conditions).

## 5.3 DIC plots

The DIC plots were generated for the simulations results from the models. The variation in the shapes of these plots can be used to compare the models and we can examine how well they fit the two scale levels.

### 5.3.1 Plots for $\theta^T$ ~*Gamma(1,1)*

The figures below show the DIC plots for the models 1 through 5. These plots have been computed over the 500 simulations.

The yphDIC curve for model 1 (Figure 5.6) indicates a slight bimodal nature whereas the ycDIC curve is more like a spike indicating lesser variability in the DICs for counties. The DIC plots for model 2 (Figure 5.7) show more variability than the plots for model 1. The ycDIC plot for model 2 shows a bimodal nature towards the end. But no unusual skewness is seen in the yphDIC plot for model 2. The ycDIC and yphDIC plots for model 3 (Figure 5.8) seem to be incorporating the two levels well (i.e. the curves appear closer) ,and the ycDIC curve is relatively smoother as compared to model 2.

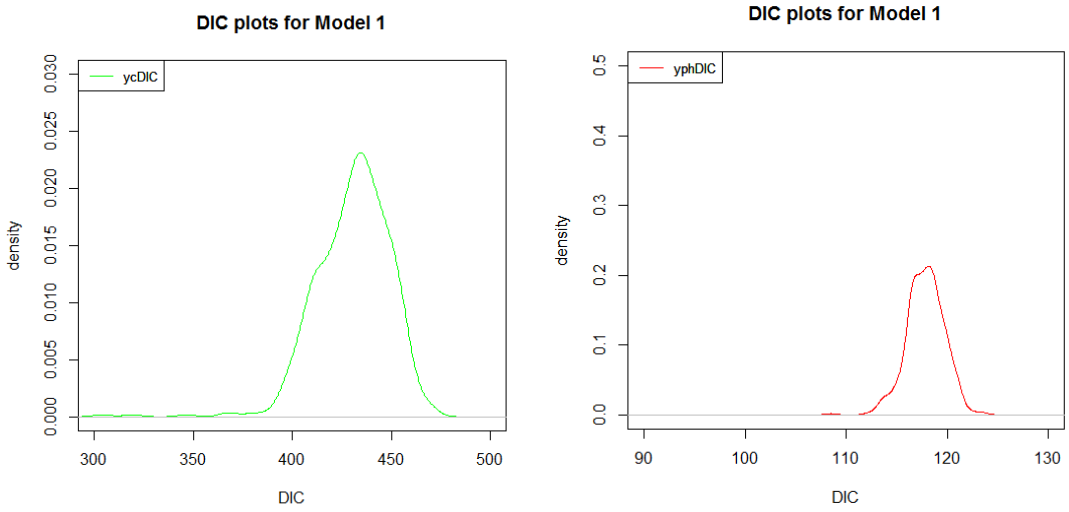


Figure 5.6 DIC plots for Model 1

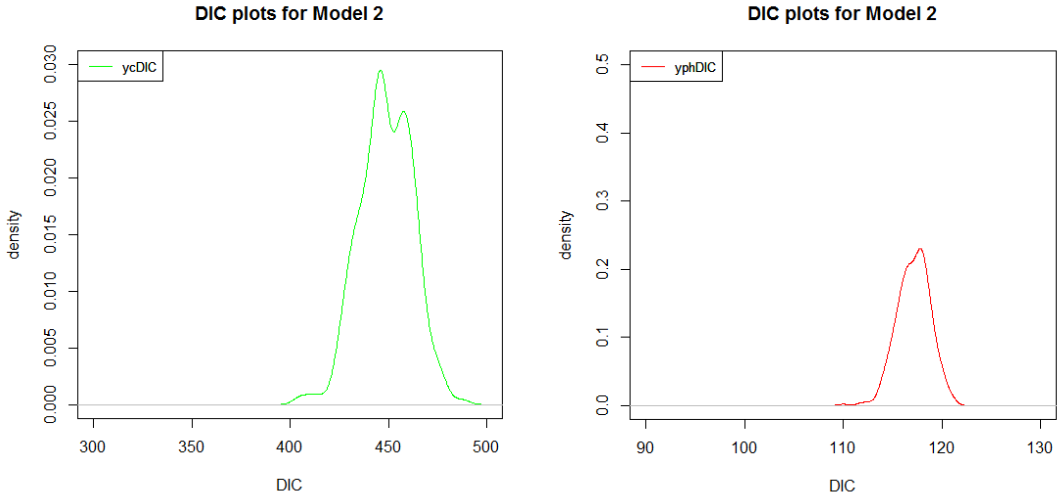


Figure 5.7 DIC plots for Model 2

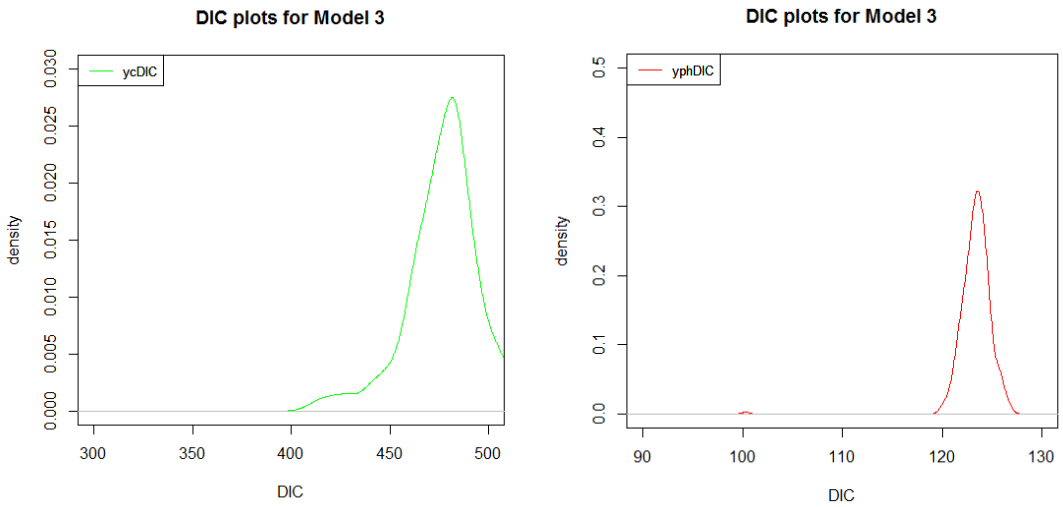


Figure 5.8 DIC plots for Model 3

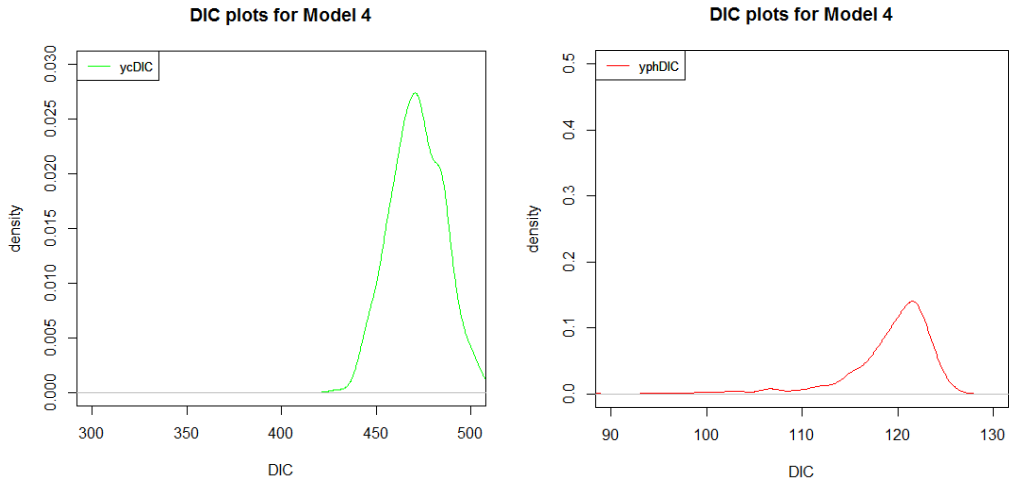


Figure 5.9 DIC plots for Model 4

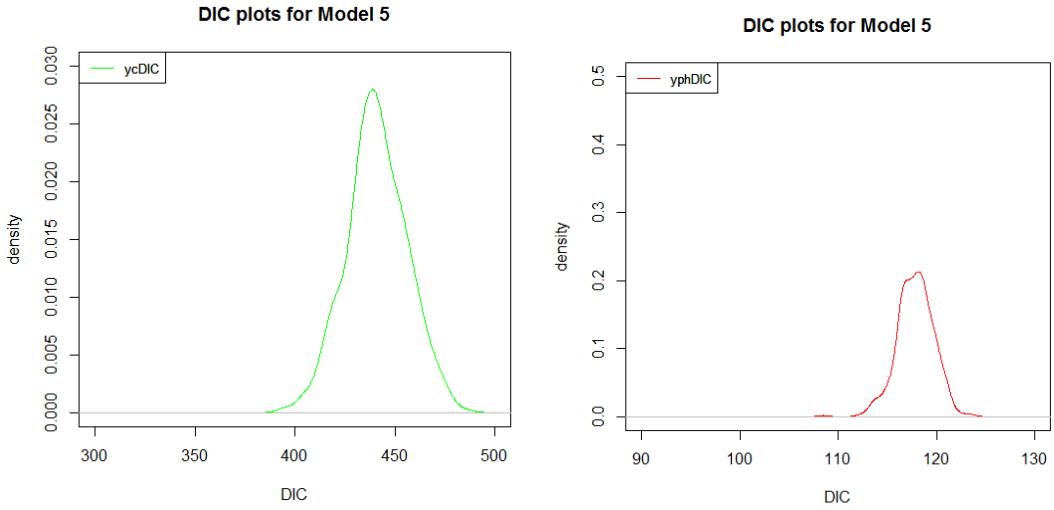


Figure 5.10 DIC plots for Model 5

The ycDIC plots for models 1, 2 and 4 show slight bimodal natures whereas the ycDIC plots for models 3 and 5 appear to be smoother. The model 1 ycDIC plot and model 4 yphDIC plot shows more variance in the DICs as compared to the other plots. Overall these plots for the five models do not indicate any strong skewness or irregular shape for $\theta^T$ *~Gamma(1,1)*.

### 5.3.2 Plots for $\theta^T$ *~Gamma(3,3)*

The DIC plots for models 1 through 5 with shape and rate parameter 3 are included below.

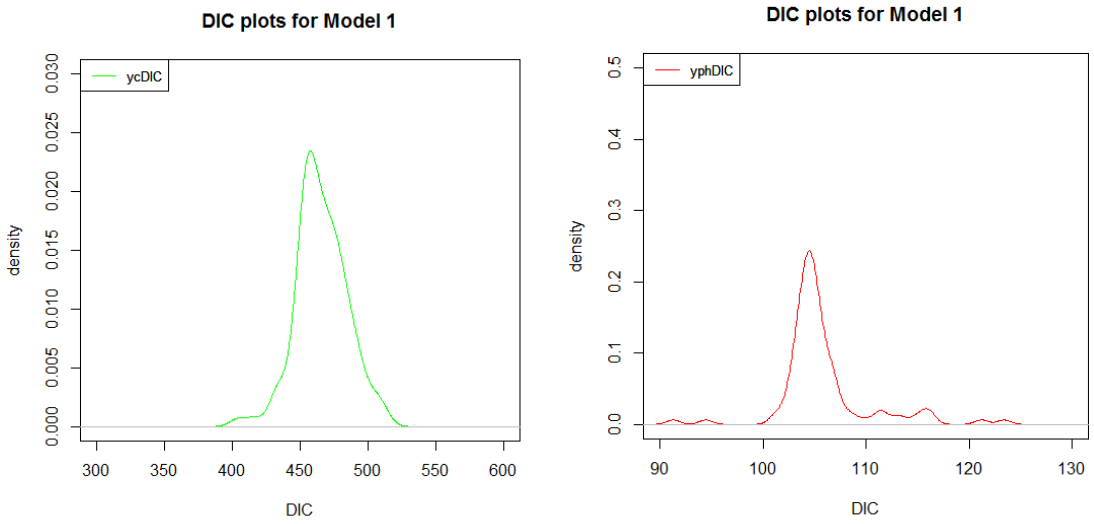


Figure 5.11 DIC plots for Model 1 (from left to right, ycDIC and yphDIC)

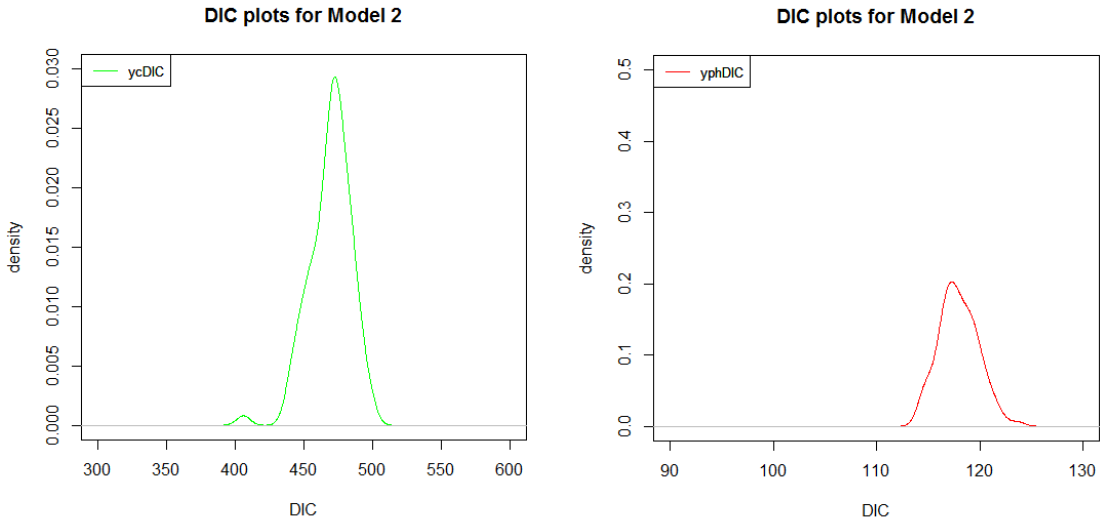


Figure 5.12 DIC plots for Model 2

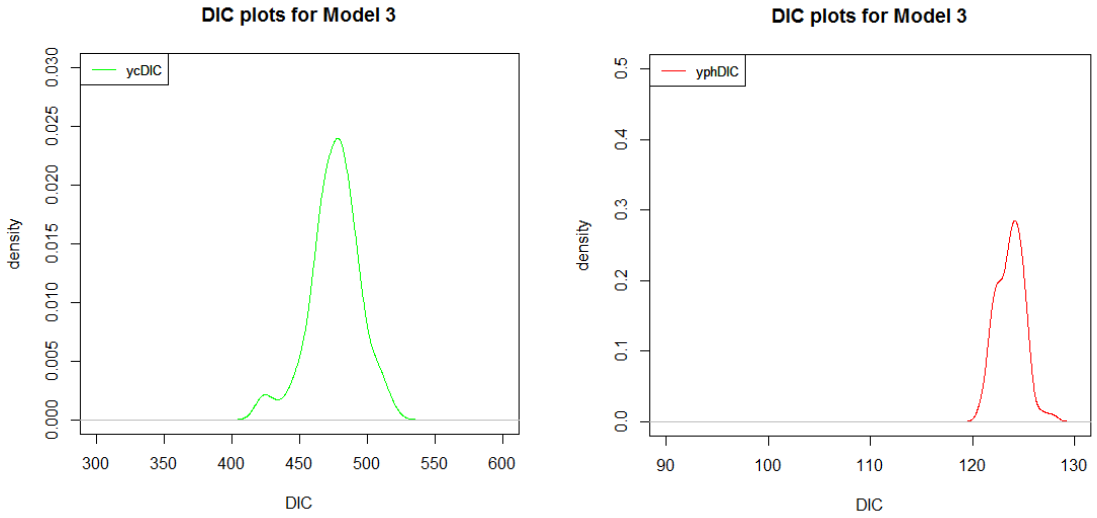


Figure 5.13 DIC plots for Model 3

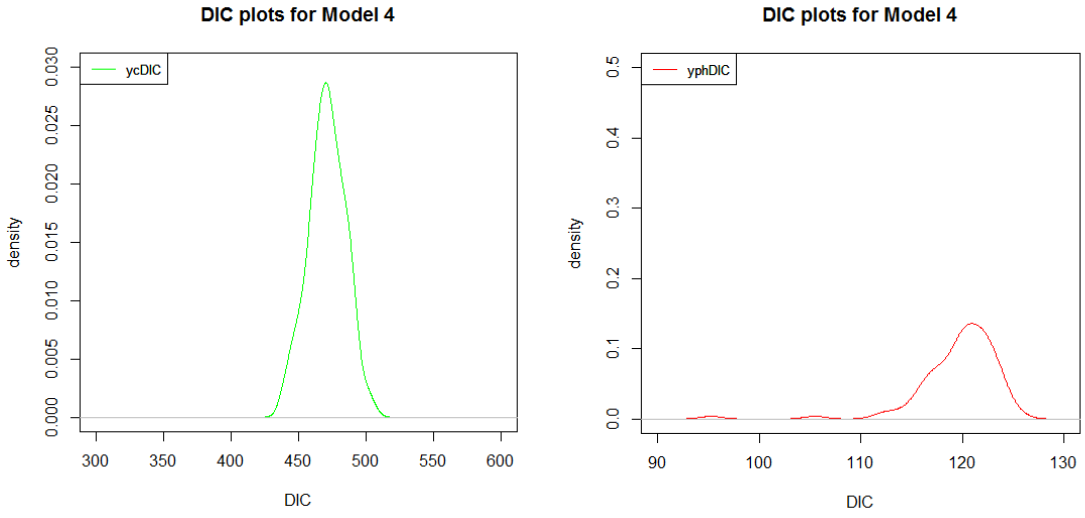


Figure 5.14 DIC plots for Model 4

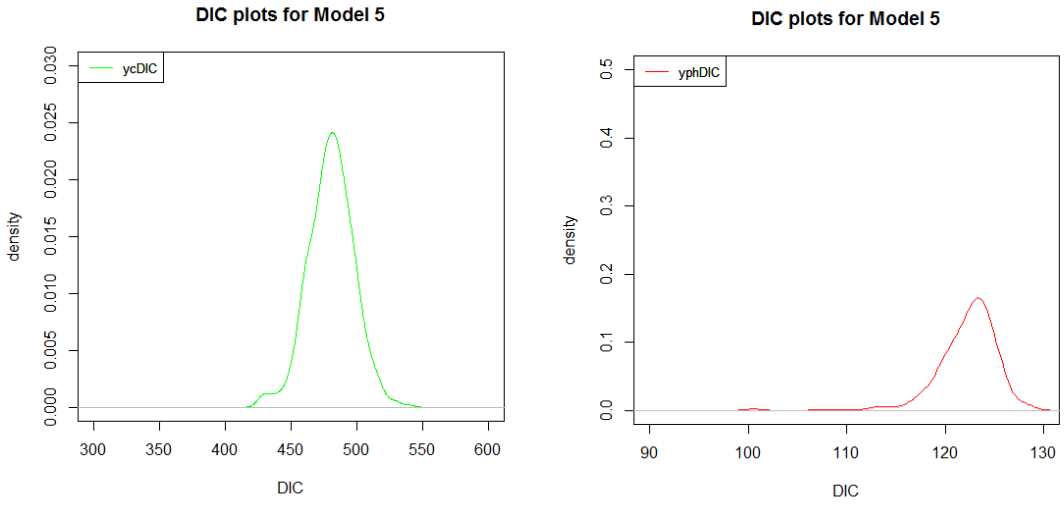


Figure 5.15 DIC plots for Model 5

The yphDIC plot for Model 1 shows less variation than the ycDIC plot but appears to have a smoother curve. Models 4 and 5 also show smoother curves for the plots as compared to other models, although some skewness is seen towards the graph ends for these plots. The plots from this set of simulation too do not show any unusual shape or extreme skewness.

## 5.4 Results from Maps

All five of the models examined include correlated spatial random effects ($U_j^{PH}$ or $U_i^C$) and the uncorrelated random effects ($V_j^{PH}$ or $V_i^C$). We generated maps for these effects with the help of advanced features of WinBUGS with GeoBUGS to analyze the spatial

distribution of the disease outcome (Oral Cancer) in Georgia. The maps for these random effects were generated along with the analysis.

### 5.4.1 Maps from preliminary analysis

From the preliminary evaluation results we generated maps for $V_i^c$ (v), $U_i^c$ (u) ,$V_j^{PH}$ (vph), $U_j^{PH}$ (uph), $\theta_j^{PH}$ (thph) ,and $\theta_i^c$ (thc) from WinBUGS with GeoBUGS. The maps were computed from the 5000 posterior converged samples for the models 1 through 5. The maps for the above parameters have been included in the Map appendix.

The maps for *u* and *uph* generally appear as random patches whereas the maps for *v* and *vph* should appear as random distribution of regions (more like spots as compared to patches) which is expected to be the case with appropriate models. The $\theta_j^{PH}$ (thph) ,and $\theta_i^c$ (thc) parameters have been plotted on the same scale to make them comparable however the same could not be done with the maps for *u* and *v* as some of the effects for some of the regions were not clearly captured when the same scale was used. The maps from the initial run for *v* (from all models) show that the mean values for $V_i^c$ are consistently high for counties like Burke, Grady, Chatham, Troupe and Upson for most of the models. The *u* maps indicate that the mean values for $U_i^c$ are elevated more towards the lower counties (beginning from the mid-counties). The maps for uph and vph also show that the mean values are elevated more for the districts in which the above mentioned counties lie. The maps for thc and thph show the mean level of relative risks in the counties and the districts. Some of the counties (some of them mentioned above) have a high relative risk that is captured by most of the five models. The Albany district

that lies in the southwest, the Augusta district in the east, the LaGrange district in the west and the Savannah district show elevated levels of risk through most of the models. Overall the maps from the five models show consistent results with respect to the elevated risk.

### 5.4.2 Maps for $\theta^T$ ~*Gamma(1,1)*

The maps for the simulation results were obtained through the maps and RColorBrewer libraries in R. It provides us with the Georgia county maps. Through this function I obtained the maps for $U_i^c$, $V_i^c$ and $\theta_i^c$ for the five models. The results for the maps from the simulations are similar to the results obtained from the initial run for the counties. The high risk counties are shown to be spread out through the state. More elevated risk is seen towards the south and the west of Georgia. Higher values for *u* and *v* are seen around the counties with elevated risks.

### 5.4.3 Maps for $\theta^T$ ~*Gamma(3,3)*

The maps from this set of simulations capture the risk in the Northern counties as well as the Southern counties. These results are also similar to the earlier results from maps. Baldwin and Wilkinson counties are consistently seen as counties with elevated risk. Along with these counties, Grady, Thomas and Brooks counties in the south also appear as areas with higher risk. The northern counties like Rabun, Union and Towns are also seen as having higher risk for oral cancer deaths.

## 5.5 Discussion

In this thesis we have presented a Bayesian approach to model count data at multiple scales. Multiscale models have been extensively used in spatial epidemiology and other fields of Geostatistics as they help to address the *scale issue* by allowing us to incorporate linkages between the scales. The following sections provide a brief discussion of the results and their implications.

### 5.5.1 Inferences based on Preliminary Analysis and Simulations

The results for the preliminary analysis indicate that model 1 is the best fitting one with the lowest DICs for both model components. Similar results are observed from the two set of simulations for $\theta^T$. In case of the preliminary analysis the DIC for model 1 differs by 7 units from the next lowest DIC for the county component and by 5 units from the district component. This is a significant difference in the DICs as a difference of 2-3 units is considered as significant. Even in case of $\theta^T \sim Gamma(1,1)$ the difference in the DIC of model 1 and other models is greater than 10 units indicating a better fit for model 1. In case of $\theta^T \sim Gamma(3,3)$, however, the DIC for model 1 is not very different from that of model 2 for the county part but the yphDIC differs by more than 10 units. But the overall results suggest that best fitting model is the one that simply aggregates the county estimated effects and does not estimate the PH level separately ,and that we can obtain a better DIC for county data by including a PH district level in the analysis.

The variability for $\theta$ obtained from the simulations however show different results regarding the efficiency of $\theta$ as an estimator in the five models. The results from

$\theta^T$~*Gamma(1,1)* suggest that model 2 is a better estimator of the relative risk for the counties and model 4 is the better estimator for risk in the districts. The variability for $\theta$ for the PH part from this set of simulations are not very different for the five models. Furthermore the variability for $\theta$ for the models 3 and 4 are similar for the counties and model 1 has a higher variability for $\theta$ as compared to the other models. This indicates that model 1 may be not the best estimator of $\theta$, even though it is the better fitting model in this set of simulations.

The results from $\theta^T$ ~*Gamma(3,3)* are similar to the ones obtained from $\theta^T$~*Gamma(1,1)*. Model 2 is still the better estimator for the risk with the lowest variability for risk for the county component and model 5 is the best estimator for the risk in this case for the district component. In addition the variability for $\theta$ for model 5 is close to that of model 2 suggesting that it is an equivalently good estimator for the risk. This result is somewhat unexpected as the mean DIC for model 5 is higher for these simulations. The variability in risk for models 1,3 and 4 are nearly same for both $\theta_j^{PH}$ and $\theta_i^c$. This implies that there is not much variation in the estimator of risk for these three models.

The two sets of simulations have the same mean for $\theta^T$ but different variances. The results for DIC from the two sets of simulations for models 3 and 4 indicate that these models have the same fit for both set of simulations for the county as well as district components (since they have similar DICs). Whereas models 1 and 2 have higher DICs from $\theta^T$~*Gamma(3,3)* as compared to the first set of simulations, and model 5 has a lower DIC for $\theta^T$~*Gamma(3,3)* than the first set of simulations. The variability for the $\theta$ is also

lower for the second set of simulations which is expected as the second condition of $\theta^T$ has a lower variance.

The overall pattern of our results for the model fit statistics indicates that Model 1 is consistently the better fitting model of the five fitted models. However this model may not be the best estimator of the relative risk as it has the highest variability for the relative risks. This could be because the spatial effects which account for the unobserved confounding in the risk for PH districts is not taken into account. Model 2 seems to be the better estimator of risk in the counties and models 3 and 5 are seen to be better estimators for districts.

### 5.5.2 Inferences based on DIC plots and Maps

The DIC plots capture the overall variation obtained for the DICs from the simulations. These plots do not give much information about the best fitting model but indicate how well the levels fit within one another. They also illustrate the extreme values for the DICs using the specified conditions of relative risk. In general the plots seem to have similar shapes for the two set of simulations.

The results obtained from the maps for all three analyses are somewhat consistent, in that they show areas of high risk that appear in most of the maps generated. The map appendix also contains a map for the number of oral cancer deaths in Georgia for the counties from 2004 obtained from Oasis site (Figure 5.31). The maps obtained from the preliminary evaluations (Figures 5.15-5.20) are in some accordance with this map. They indicate that the risk is high in the counties and the surrounding areas where there have

been more number of deaths. Some counties like Richmond, Thomas, Murray are observed to be consistently high in the risk of oral cancer. Based on the report of the State of Oral Health for children in Georgia (2007) a high death rate from oral cancer is seen mostly in rural areas some of which are also indicated in our maps (Grady, Thomas, Rabun counties). Some of the potential reasons could be the lack of access to dental health care facilities and the socioeconomic status in these counties. A more detailed report about the incidence, mortality and trends in oral cancer in Georgia can be found through the National Cancer Institute.

Our thesis presents a Bayesian approach for modeling multiscale data commonly available in environmental Public Health studies. The models proposed in our research allow us to include all data at different levels of aggregation. The aggregation issue within a multiscale Bayesian framework has not been considered in many previous studies. Using the goodness of fit criteria to evaluate the scale effect we found that model 1 provides the best fit for this data even when tested under different conditions. This model however shows more variability in the estimated risk for the study area. The differences in the variability for $\theta$ is however not very large among the five models considered. Thus given these small differences we can apply model 1 to other potential multiscale data sets to get an optimal explanation of each scale. Our overall results thus indicate that the best method to analyze multiscale data includes the use of simple models where we may not need to estimate the risk separately at each level, but we can account for the aggregate effects from the smaller levels in the higher levels.

## 5.6 Further aspects of the study

The expected counts that were used in our models were obtained from the Georgia data. We also examined the models when larger expected counts were used. These counts were computed using the function *ec<-rnorm(159,4.5,1)* in R ,and the remaining parameters were used from the preliminary evaluation models. This procedure was carried out to analyze the model performances when higher expected counts are used. The models were fitted for a single run in WinBUGS. Table 5.4 provides the estimates for DICs and pDs from these models. These estimates indicate that model 1 has the best fit with the lowest DIC which is 12 units lower than the next lowest DIC for model 5. The overall DIC is seen to increase for the five models when higher expected values are used.

The models investigated in our study do not include any covariates other than the random effects. It would be interesting to examine the model fits when more variables are added (It may lower the DICs for the models). Furthermore it could also be of interest to apply these models to larger areas or to get maps for diseases that are not rare or small in counts. These additional avenues of research could be further used to substantiate the scale change effects.

Table 5.1 Results from a single run in WinBUGS:

| Results from a single run in WinBUGS | | | | |
|---|---|---|---|---|
| Model numbers | pD | | DIC | |
| | yc | yph | yc | yph |
| 1 | 41.84 | 5.19 | 481.35 | 104.03 |
| 2 | 26.37 | 12.79 | 488.89 | 115.52 |
| 3 | 3.54 | 3.56 | 496.05 | 113.81 |
| 4 | 10.026 | 6.43 | 490.67 | 109.39 |
| 5 | 11.354 | 7.33 | 492.68 | 113.16 |

Table 5.2 Results from simulations using the condition thetaT<-rgamma(159,1)

| Results from $\theta^T$ ~*Gamma(1,1)* | | | | | | |
|---|---|---|---|---|---|---|
| Model numbers | pD | | DIC | | Variability for θ | |
| | yc | yph | yc | yph | yc | yph |
| 1 | 60 | 7.10 | 430.6 | 100.40 | 0.00819 | 3.34 |
| 2 | 65.50 | 15.24 | 450.1 | 117.3 | 0.0037 | 3.59 |
| 3 | 64.38 | 14.60 | 476.8 | 123.4 | 0.0072 | 3.23 |
| 4 | 31.73 | 12.77 | 471.4 | 119.1 | 0.0075 | 3.20 |
| 5 | 62.31 | 13.77 | 441 | 117.9 | 0.0049 | 3.51 |

Table 5.3 Results from simulations using the condition thetaT<-rgamma(159,3). These were obtained over the average of 100 simulations.

| Results from $\theta^T$ ~*Gamma(3,3)* | | | | | | |
|---|---|---|---|---|---|---|
| Model numbers | pD | | DIC | | Variability for θ | |
| | yc | yph | yc | yph | yc | yph |
| 1 | 24.25 | 4.60 | 468.3 | 106.4 | 0.007 | 3.14 |
| 2 | 41.76 | 12.80 | 469.1 | 117.9 | 0.0012 | 2.93 |
| 3 | 25.27 | 15.49 | 475.7 | 123.7 | 0.0069 | 3.17 |
| 4 | 33.6 | 11.78 | 470.8 | 119.70 | 0.0068 | 3.27 |
| 5 | 20.81 | 14.53 | 483 | 123.6 | 0.0013 | 2.53 |

Table 5.4 Results from a single run in WinBUGS for higher expected values ec<rnorm(159,4.5,1)

| Results from a single run in WinBUGS for ec<-rnorm(159,4.5,1) | | | | |
|---|---|---|---|---|
| Model numbers | pD | | DIC | |
| | yc | yph | yc | yph |
| 1 | 78.24 | 12.98 | 538.41 | 103.94 |
| 2 | 93.76 | 17.48 | 592.74 | 123.25 |
| 3 | 94.54 | 16.32 | 556.70 | 124.12 |
| 4 | 96.43 | 16.33 | 562.70 | 129.95 |
| 5 | 81.57 | 14.27 | 550.98 | 130.21 |

# REFERENCES


Aing, C et al. (2011) A Bayesian hierarchical occupancy model for track surveys conducted in a series of linear, spatially correlated, sites, *Journal of Applies Ecology* 48, 1365-2664

Armhein, C. (1995) Searching for the elusive aggregation effect: Evidence from statistical simulations, *Environment & Planning A* 27 (1), 105-119

Berke, O. (2004) Exploratory disease mapping: kriging the spatial risk function from regional count data, *International Journal of Health Geographics*, 3-18

Best, N., Richardson, S., and Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping, *Statistical Methods in Medical Research* 14, 35-39

Dark, S., and Bram, D. (2007) The modifiable areal unit problem (MAUP) in physical geography, *Progress in Physical Geography* 31, 471-479

Flowerdew, R. (2009), Understanding the modifiable areal unit problem

Gelf,and A., Zhu, L., and Carlin, B. (2000) On the change of support problem for Spatio-temporal data, *Biostatistics* 1, 31-45

Gotway, C., and Young, L. (2002) Combining Incompatible Spatial Data, *Journal of American Statistical Association* 97, 632-648

Jones, R. (2011) The Modifiable Areal Unit Problem in GIS, *Cartographica*

Kerman, J., and Ligges, U. (2013) Package 'BRugs', R interface to the OpenBUGS MCMC software

Knorr-Held, L., and Becker, N. (1999) Bayesian Modeling of Spatial Heterogeneity in Disease Maps with application to German Cancer Mortality data, *Sonderforschungsbereich* 386, 121-140

Latouche, A et al. (2007) Robustness of the BYM model in absence of spatial variation in the residuals, *International Journal of Health Geographics*, 6-39

Lawson, A. (2013) *Bayesian Disease Mapping*: Hierarchical Modeling in Spatial Epidemiology

Lee, D. (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping, *Spatial and Spatio-temporal Epidemiology* 2, 79-89

Lee, S., Yeatts, K., and Serre, M. (2009) A Bayesian Maximum Entropy approach to address the change of support problem in the spatial analysis of childhood asthma prevalence across North Carolina, *Spatial and Spatio-temporal Epidemiology* 1, 49-60.

Louie, M., and Kolaczyk, E. (2006) Multiscale detection of localized anomalous structure in aggregate disease incidence data, *Statistics In Medicine* 25, 787-810

Louie, M., and Kolaczyk, E. (2006) A multiscale method for disease mapping in spatial epidemiology, *Statistics In Medicine* 25, 1287-1306

Lunn, D et al. (2000) WinBUGS- A Bayesian modelling framework: Concepts, structure, and extensibility, *Statistics and Computing* 10, 325-337

Openshaw, S. and Taylor, P. J. (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem, *Statistical Applications in the Spatial Sciences*, 127-144.

Radford, N. (1998) Philosophy of Bayesian Inference

Venkatesan, P., Srinivasan, R., and Dharuman, C. (2012) Bayesian Conditional Auto Regressive model for mapping Tuberculosis prevalence in India, *International Journal of Pharmaceutical Studies and Research* 3, 1-3

Ventrucci, M., Scott, E., and Cocchi, D. (2010) Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation, *Biostatistics* 12, 51-67

You, Y., and Zhou, Q. (2011) Hierarchical Bayes small area estimation under a spatial model with application to health survey data, *Survey Methodology* 37, 25-37

# Appendix A. Map appendix

Maps for a single run (Figures A.1-A.5)

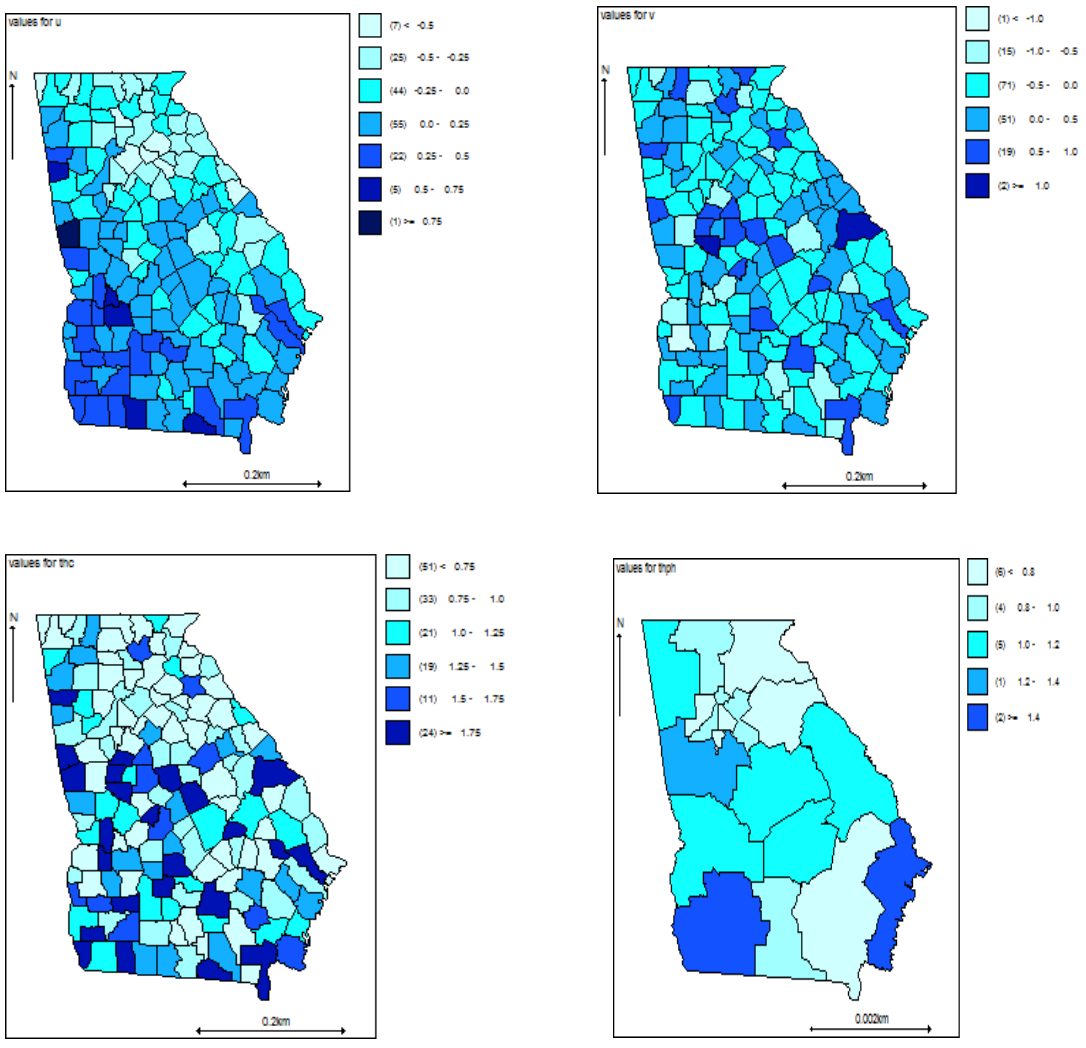


Figure A.1 (left to right) Maps for v, u, thc, thph for model 1

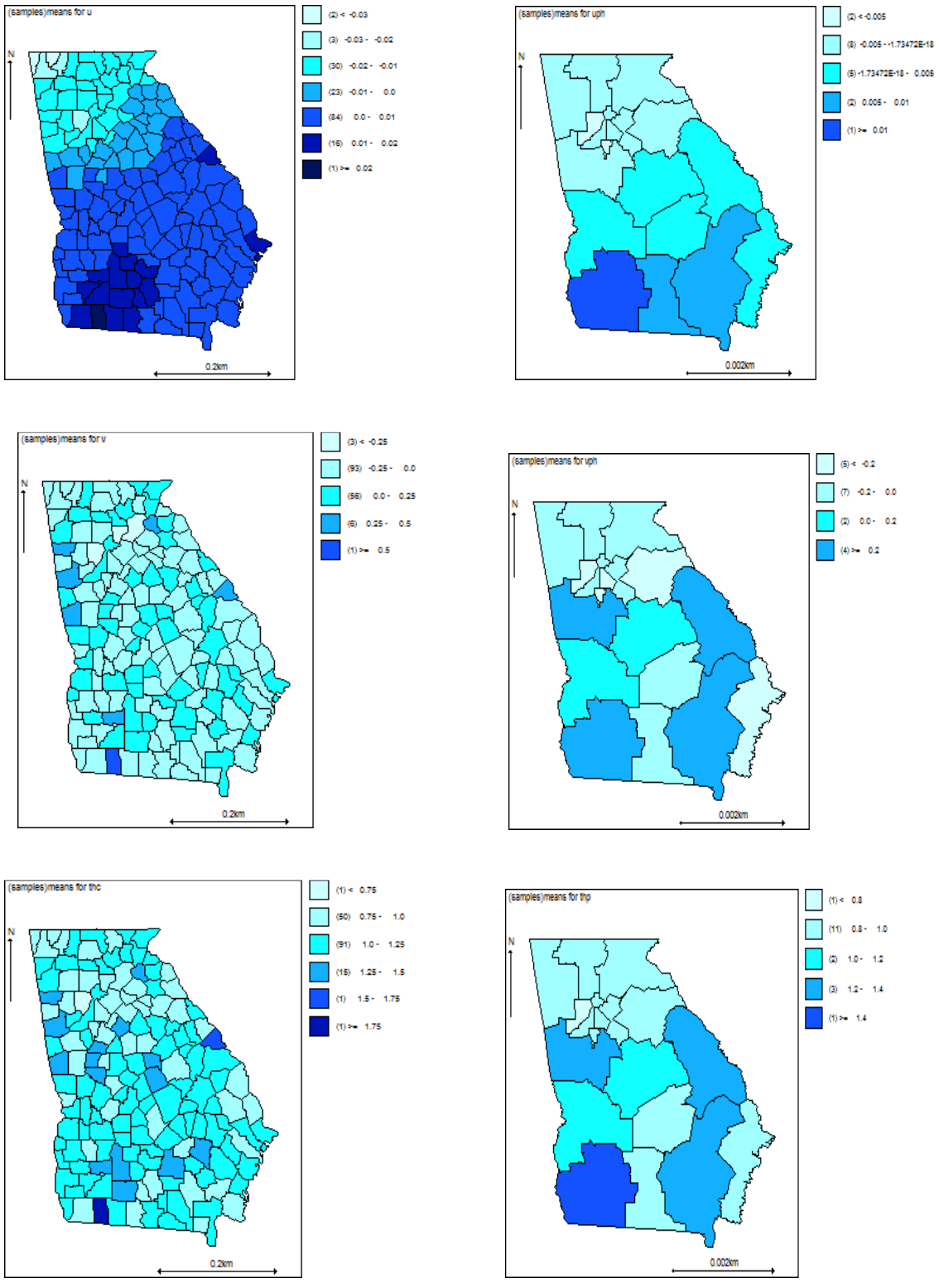


Figure A.2 (left panel are county level maps and right panel are the district level maps; from left to right) Maps for u, uph, v, vph, thc and thph for Model 2

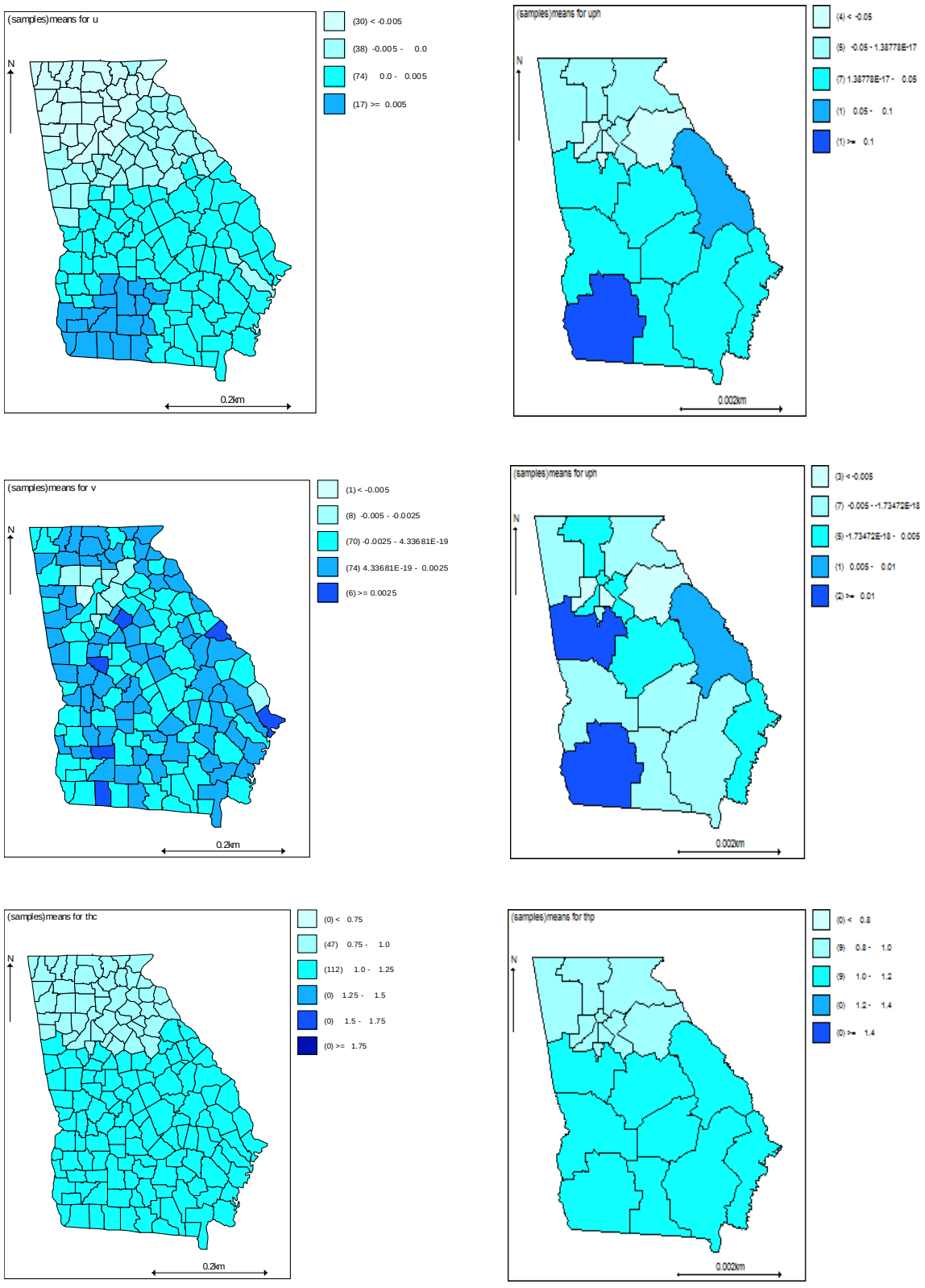


Figure A.3 (left panel are county level maps and right panel are the district level maps; from left to right) Maps for u, uph, v, vph, thc and thph for Model 3

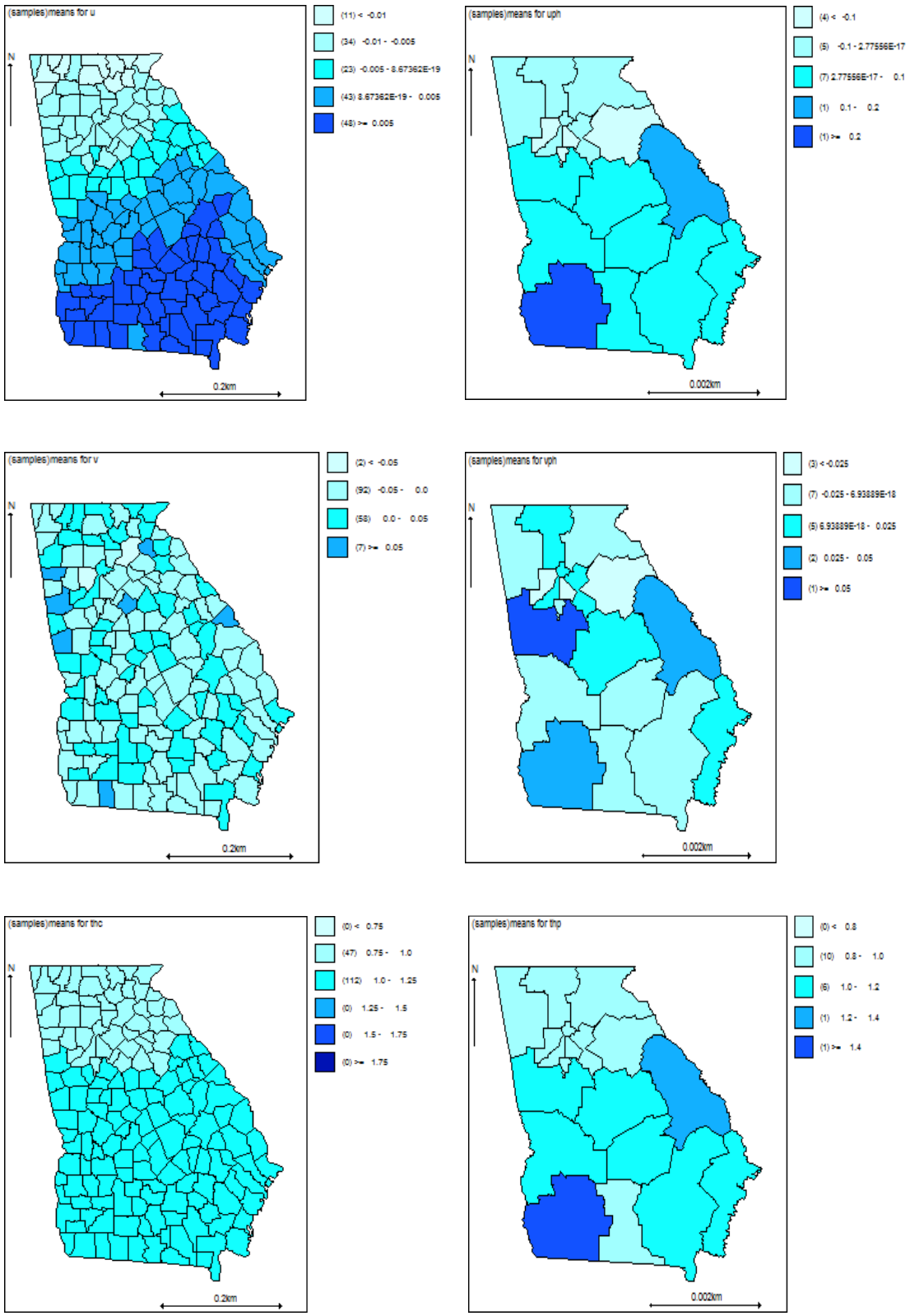


Figure A.4 (left panel are county level maps and right panel are the district level maps; from left to right) Maps for u, uph, v, vph, thc and thph for Model 4

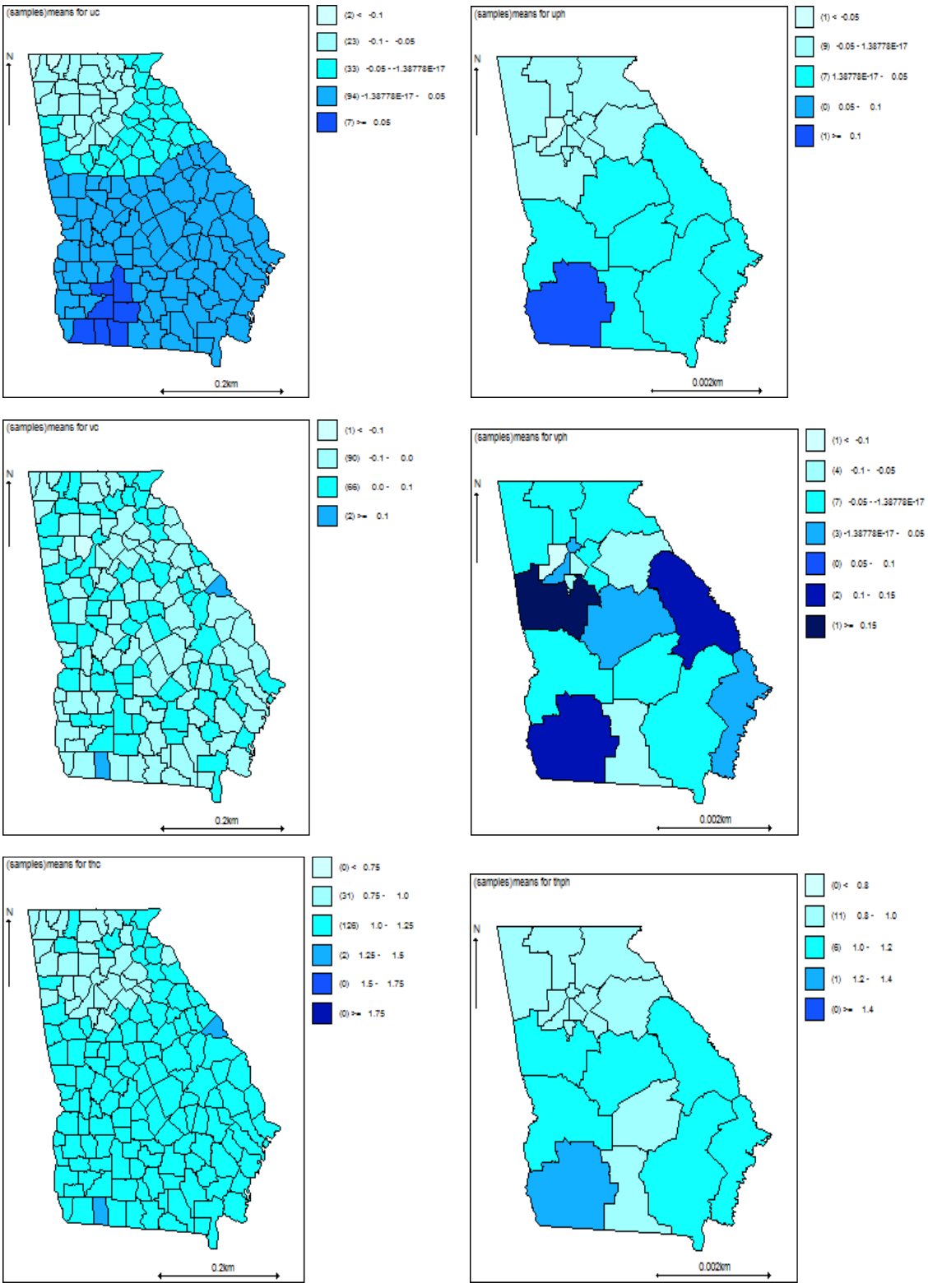


Figure A.5 (left panel are county level maps and right panel are the district level maps; from left to right) Maps for u, uph, v, vph, thc and thph for Model 5

Maps for simulation results from $\theta^T$ *~Gamma(1,1)* (Figures A.6-A.10)

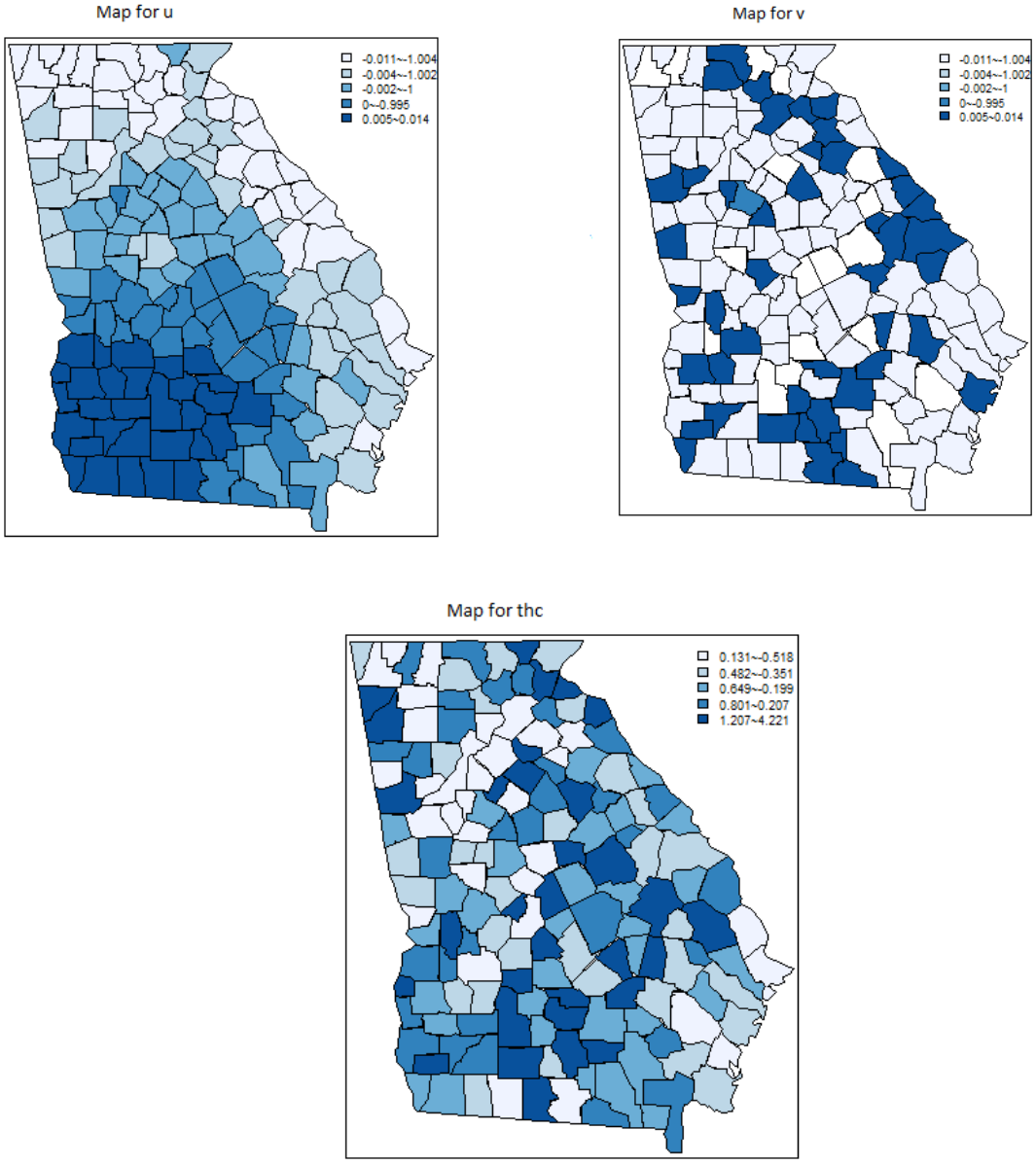


Figure A.6 (from left to right) Maps for u, v, thc for Model 1

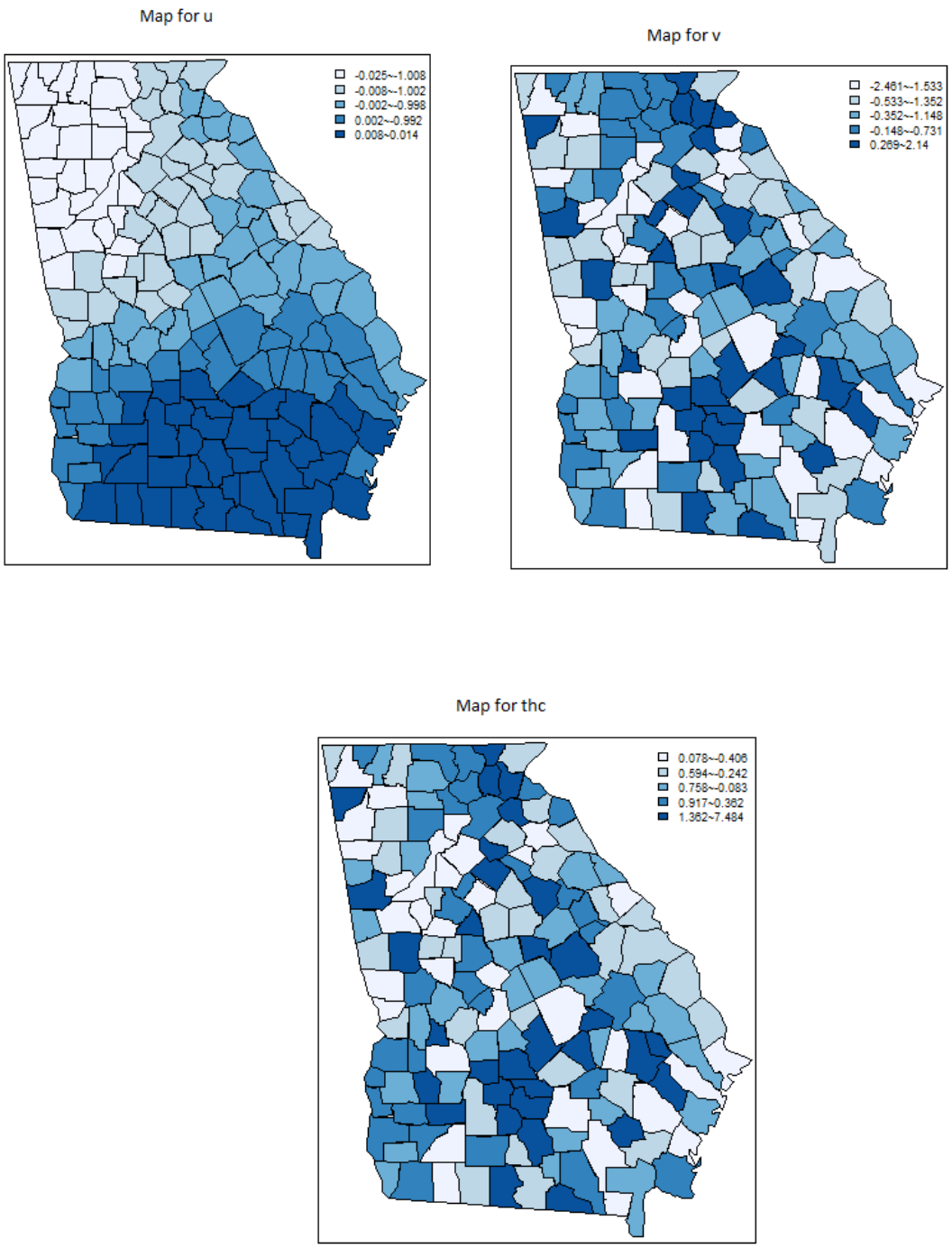


Figure A.7 (from left to right) Maps for u, v, thc for Model 2

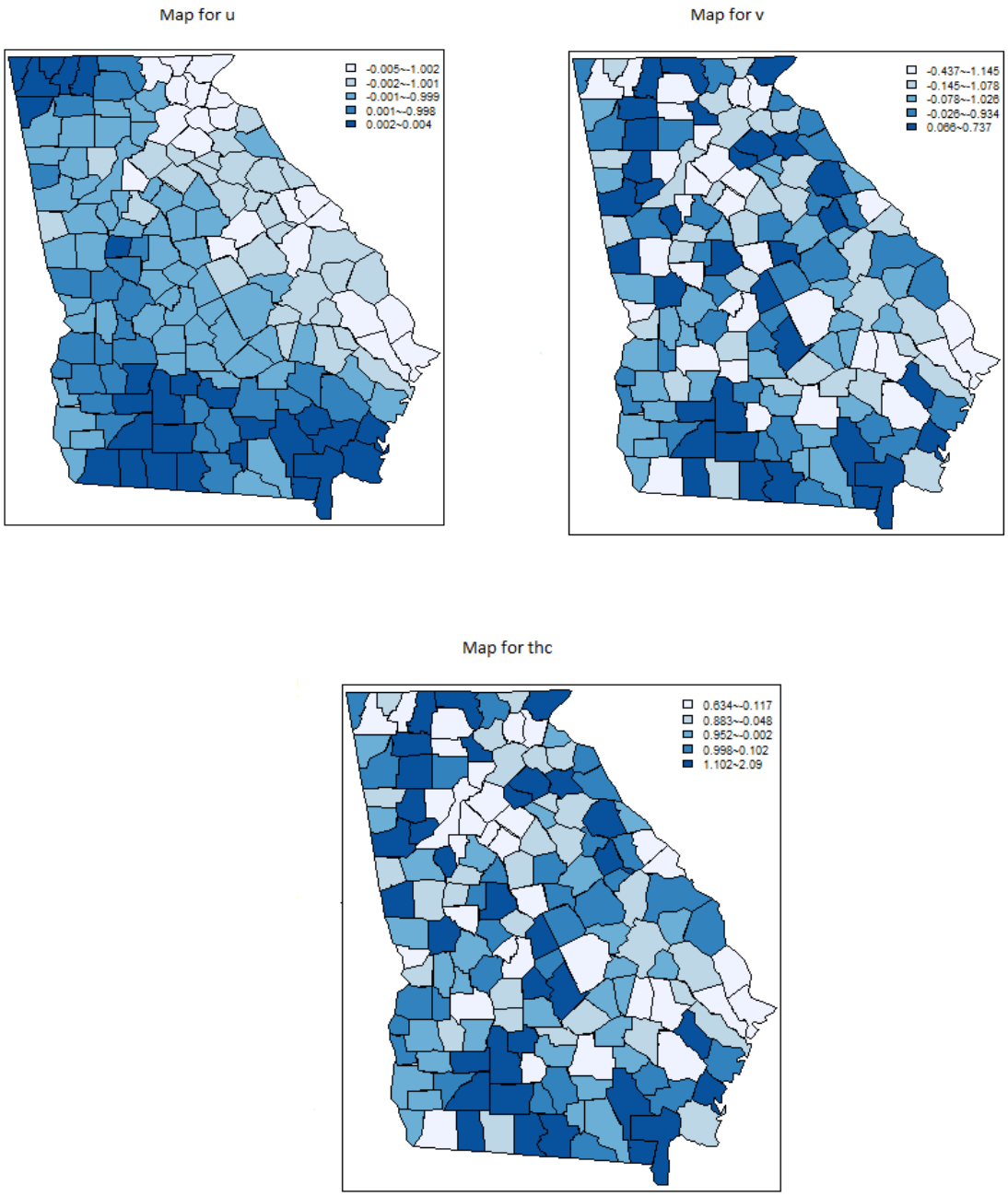


Figure A.8 (from left to right) Maps for u, v, thc for Model 3

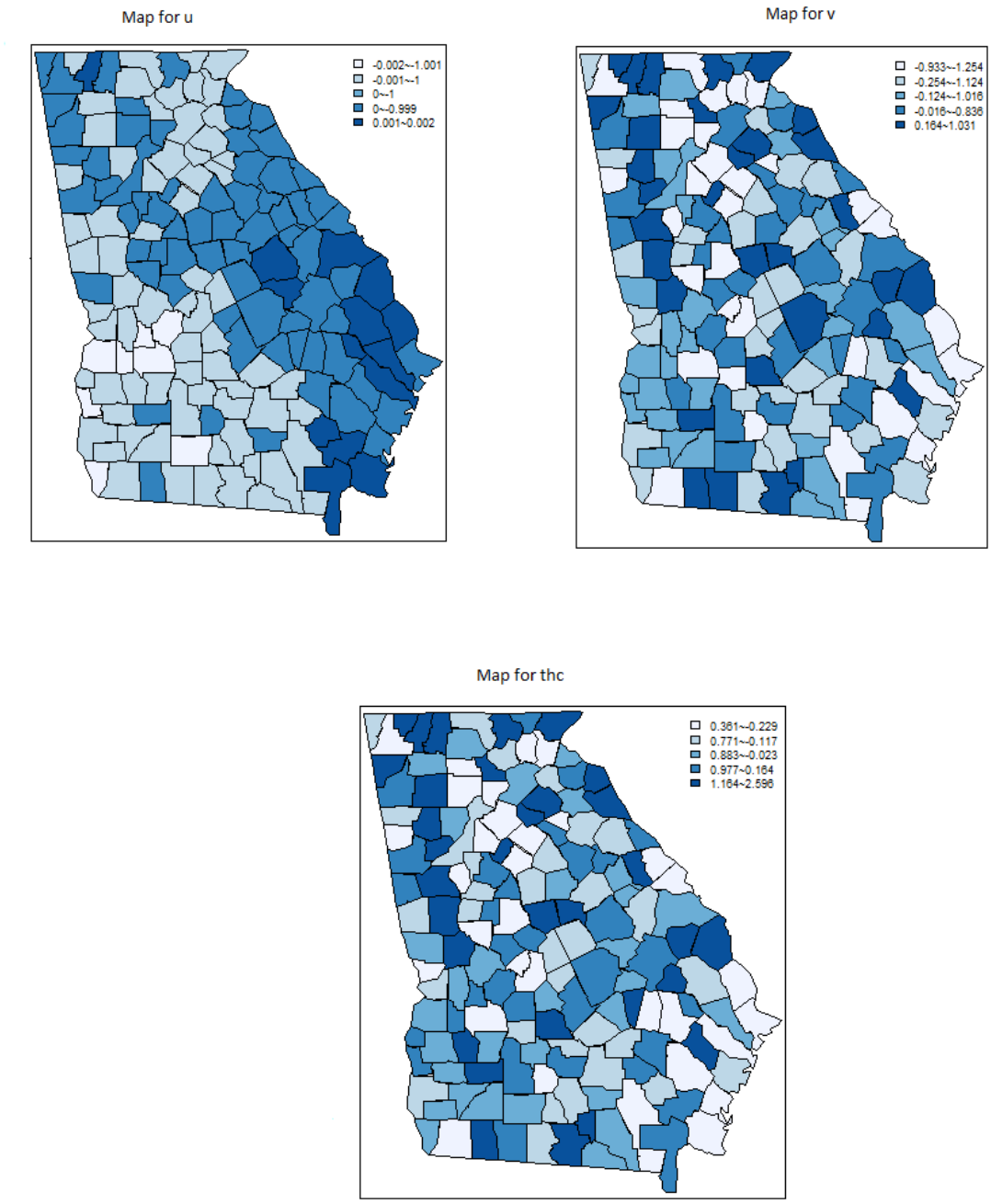


Figure A.9 (from left to right) Maps for u, v, thc for Model 4

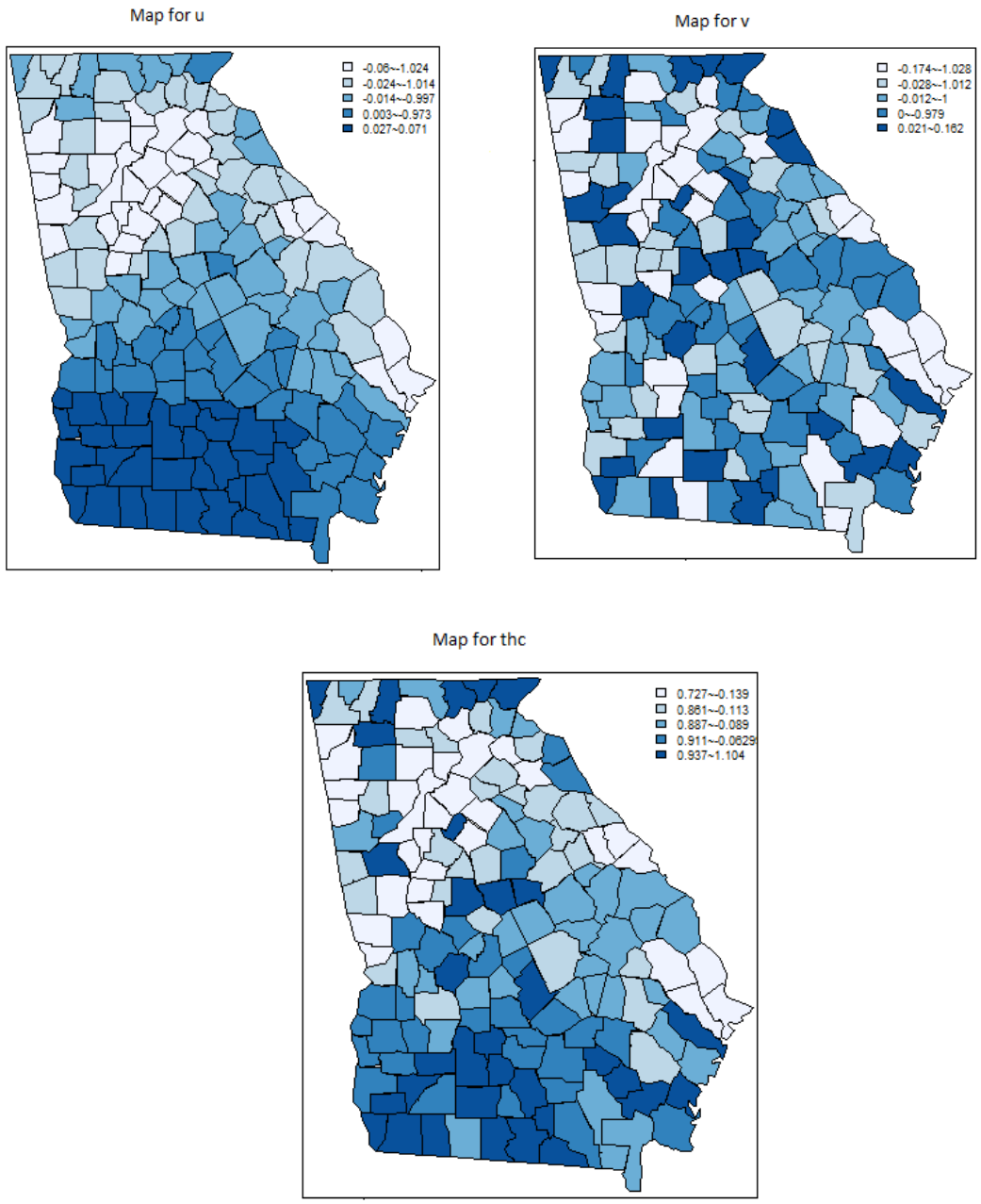


Figure A.10 (from left to right) Maps for u, v, thc for Model 5

Maps for simulation results from $\theta^T$ *~Gamma(3,3)* (Figures A.11-A.15)

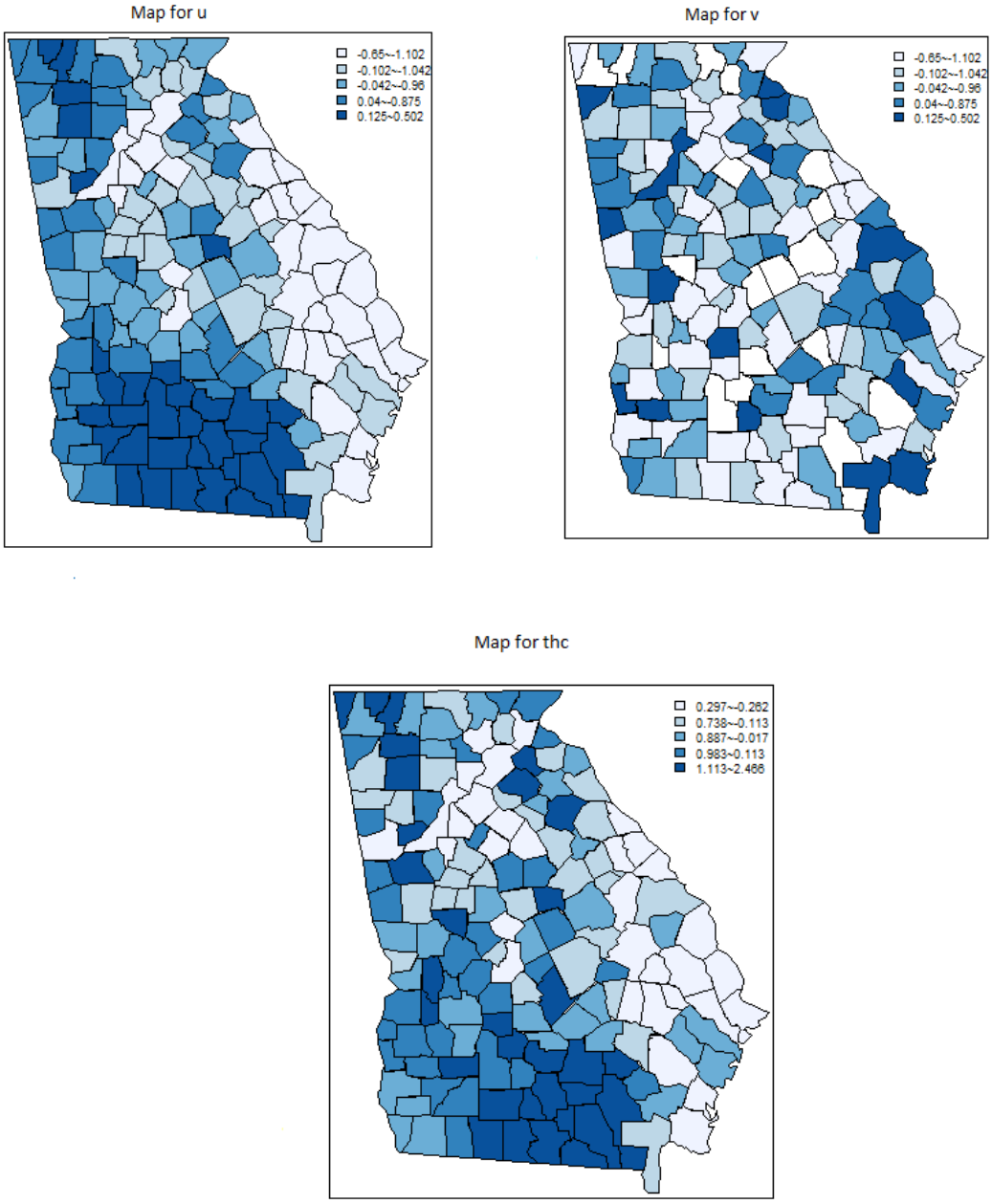


Figure A.11 (from left to right) Maps for u, v, thc for Model 1

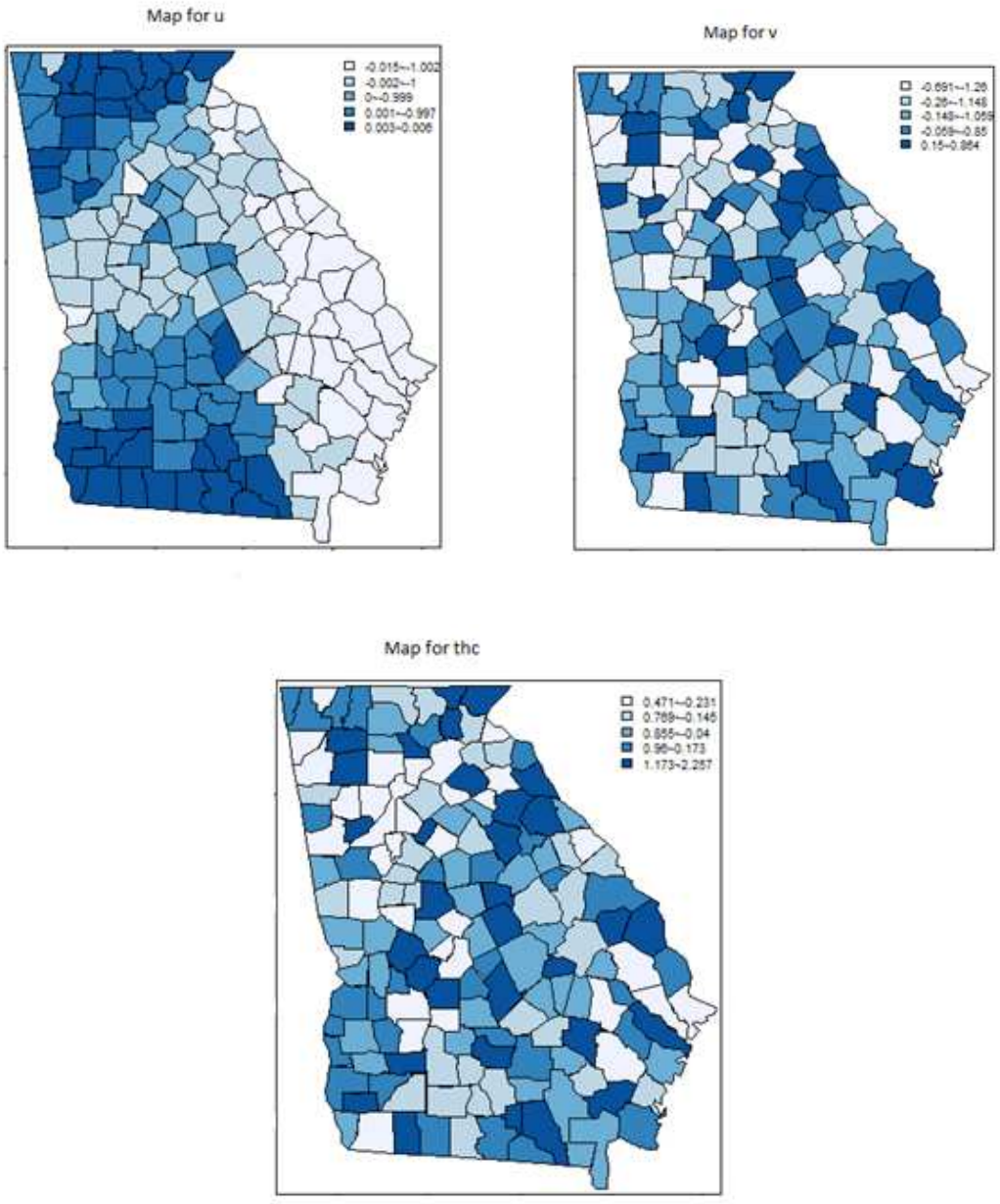


Figure A.12 (from left to right) Maps for u, v, thc for Model 2

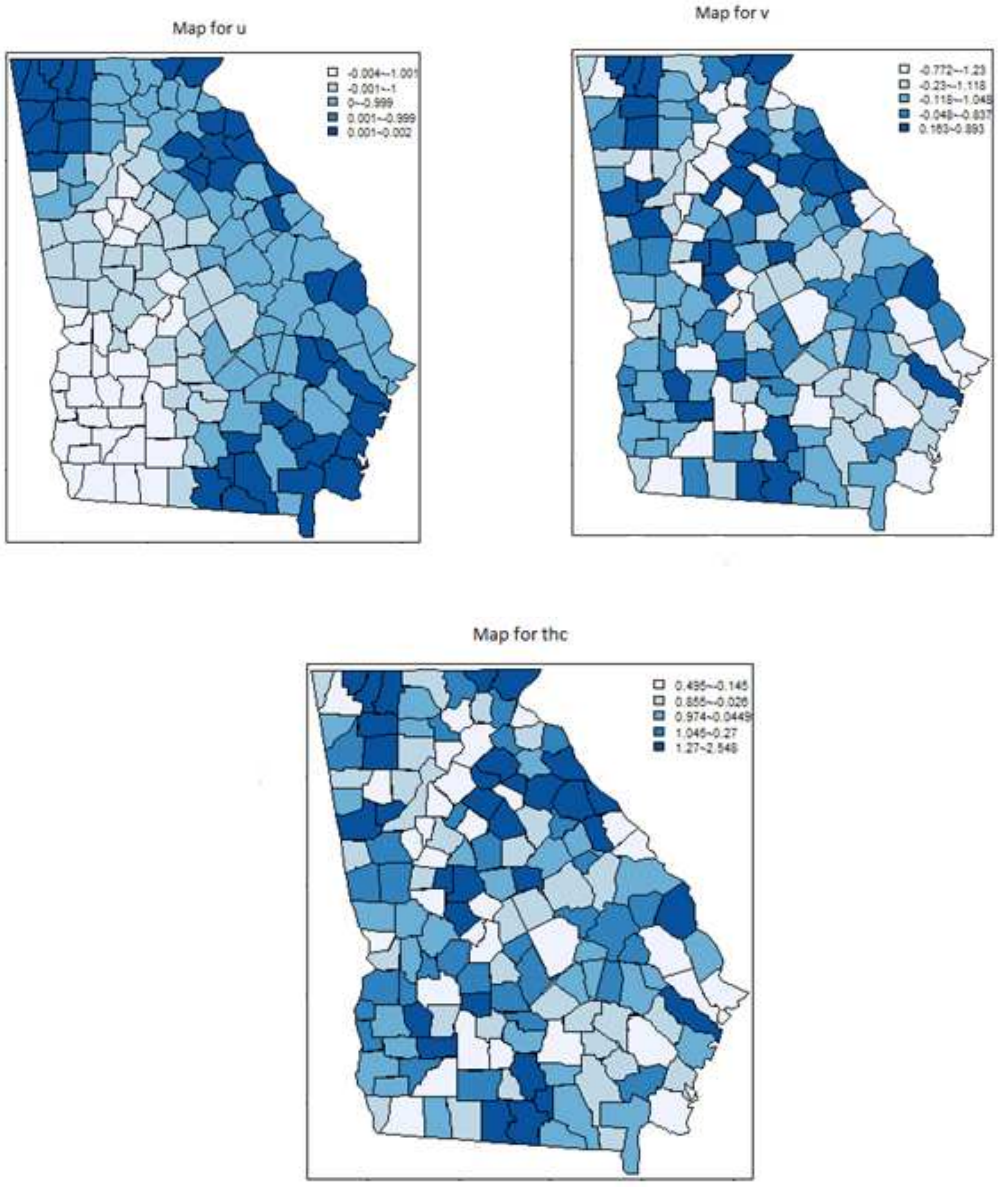


Figure A.13 (from left to right) Maps for u, v, thc for Model 3

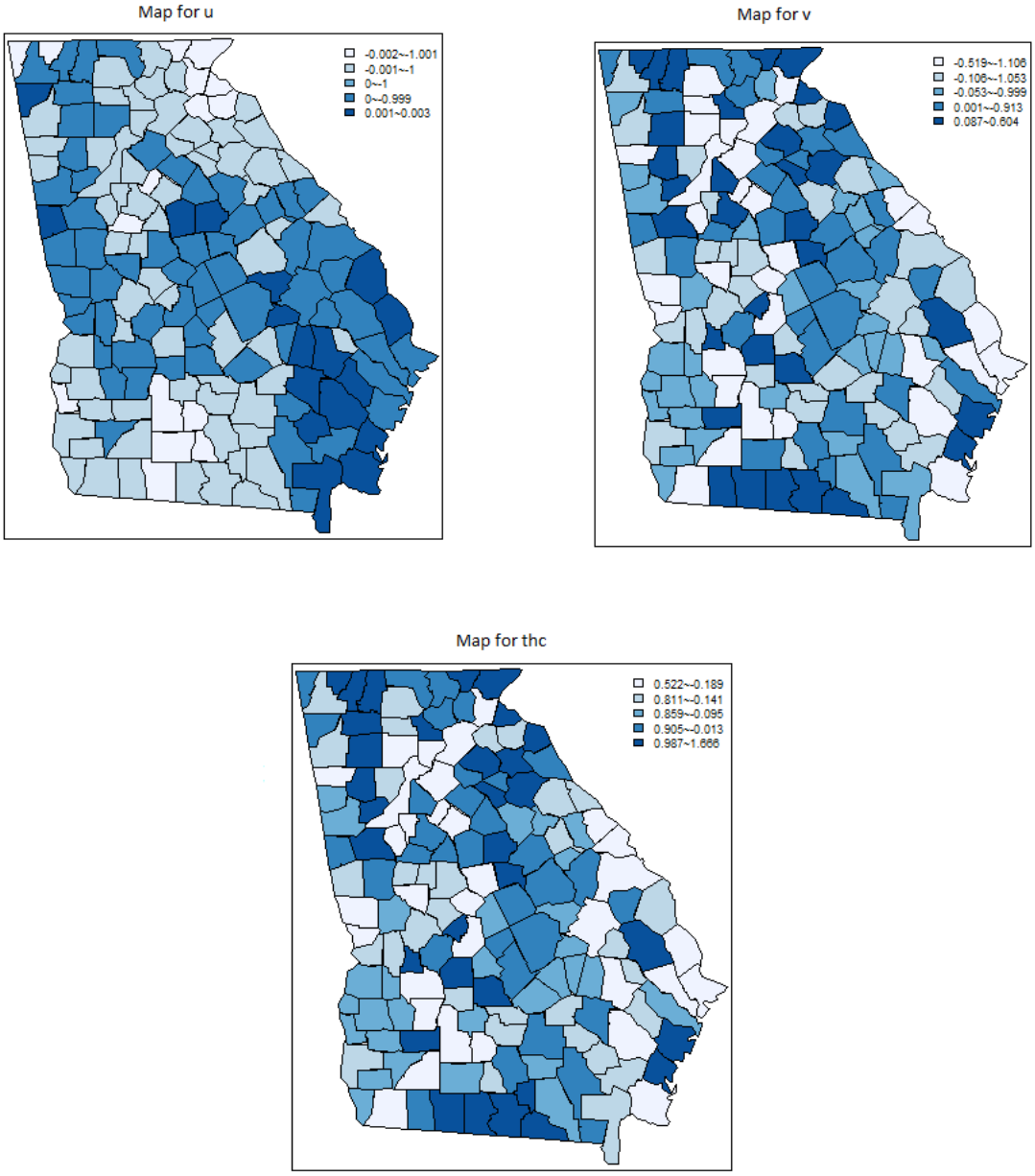


Figure A.14 (from left to right) Maps for u, v, thc for Model 4

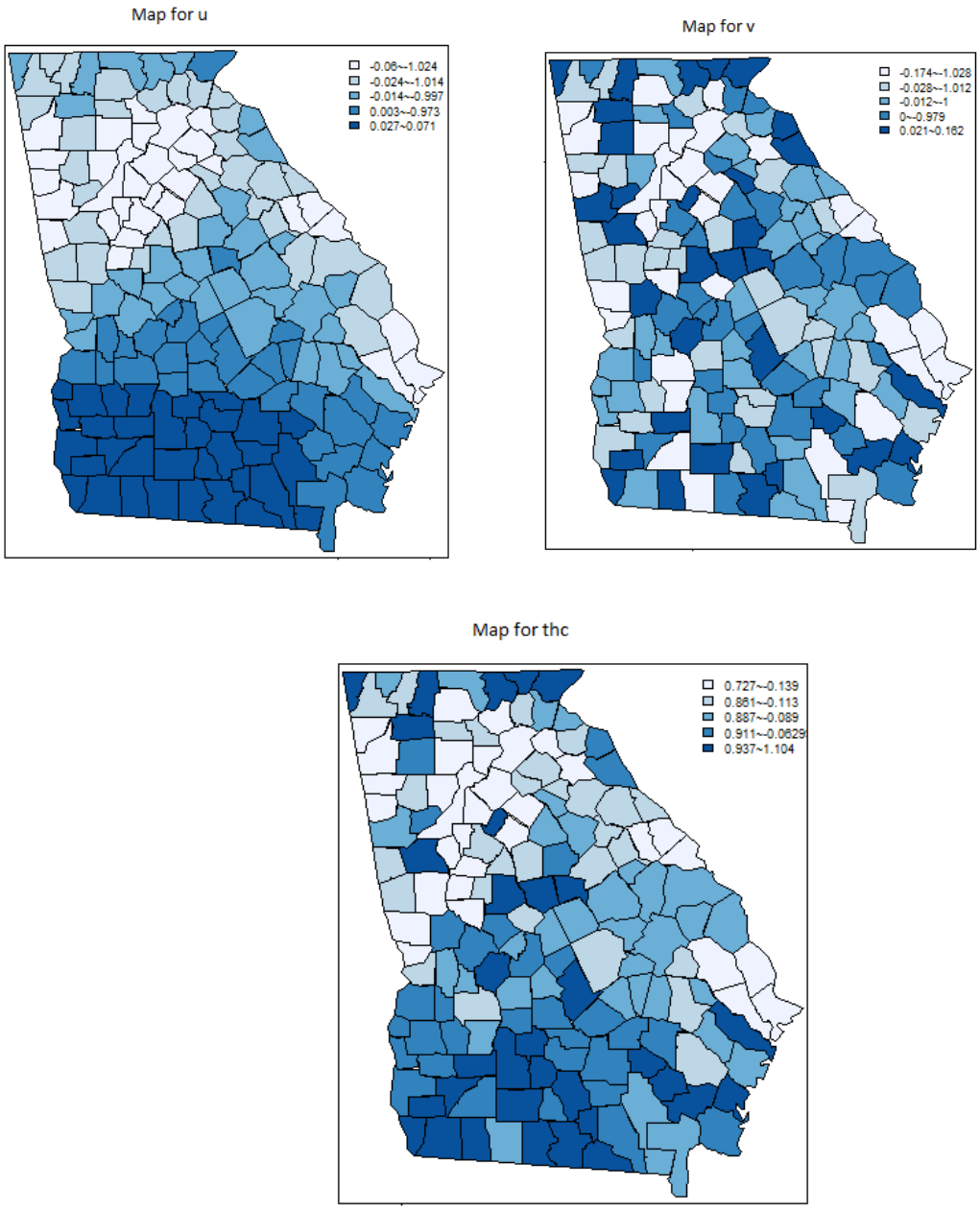


Figure A.15 (from left to right) Maps for u, v, thc for Model 5

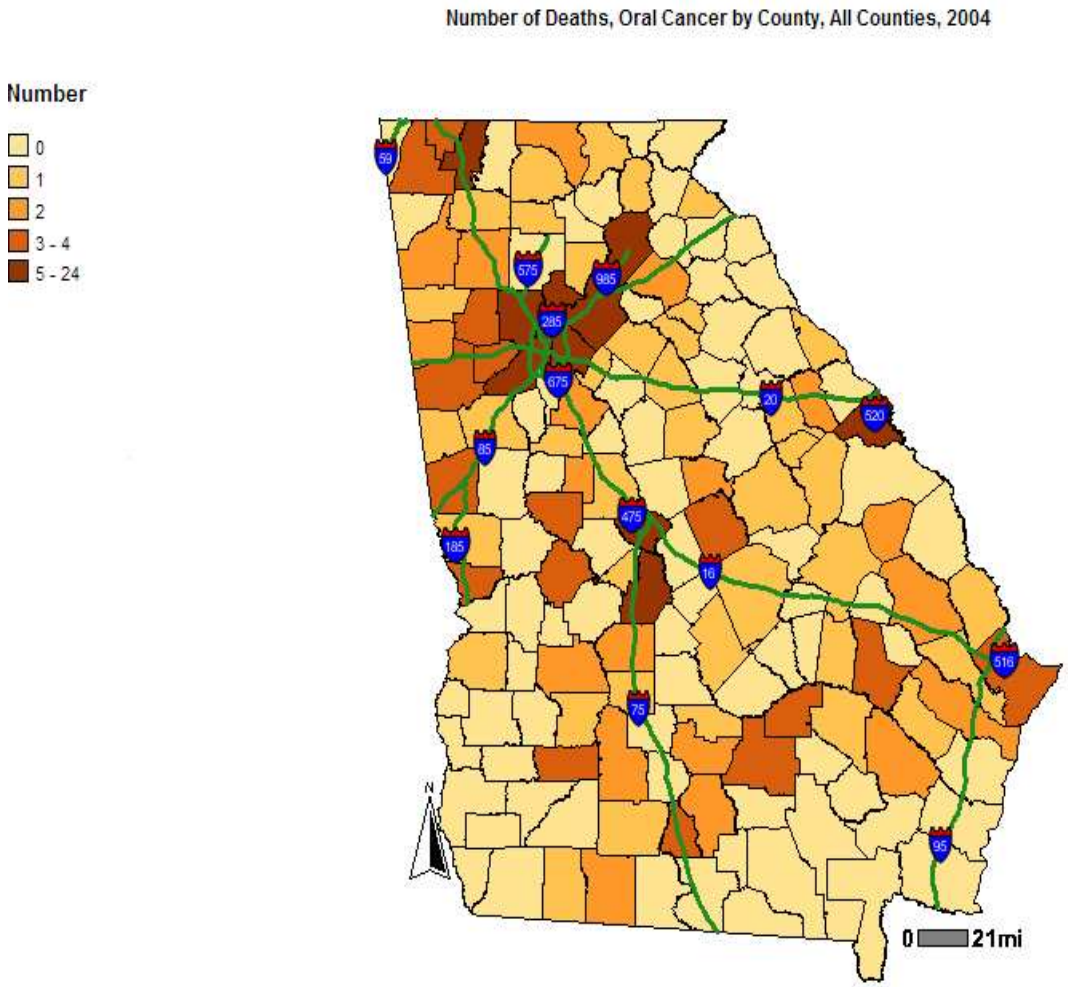


Figure A.16 Map obtained from Oasis Public Health and Public Policy Data analysis (shows the number of Oral Cancer Deaths in Georgia Counties for 2004)

# Appendix B. R and WinBUGS codes

**Code for models 1-5**

```
##########################Model 1#################################
model{
##county  level
for (i in 1:m){
yc[i]~dpois(muc[i])
muc[i]<-ec[i]*thc[i]
log(thc[i])<-a0+v[i]+u[i]
v[i]~dnorm(0,tauV)
resc[i]<-(yc[i]-muc[i])/sqrt(muc[i])
as[i]<-ec[i]*thc[i]}
u[1:m]~car.normal(adjc[],weic[],numc[],tauU)
for( k in 1: nsumc){weic[k]<-1}
a0~dnorm(0,tau0)
tau0~dgamma(4,0.005)
# #PH district level
for(i in 1:m){
phc2[i]<-phc[i]}
for (j in 1:p){
thph[j]<-mph[j]/eph[j]
```

```
yph[j]~dpois(mph[j])
as2[j]<-inprod(Label[,j],as[])
mph[j]<-as2[j]
res[j]<-(yph[j]-mph[j])/sqrt(mph[j])
num23[j]<-numph[j]
eph2[j]<-eph[j]
yph2[j]<-yph[j]}
for (k in 1 :nsumph){adjph2[k]<-adjph[k]}
tauV~dgamma(4,0.005)
tauU~dgamma(4,0.005)}

##########################Model 2###############################
model{
##county  level
for (i in 1:m){
yc[i]~dpois(muc[i])
muc[i]<-ec[i]*thc[i]
log(thc[i])<-a0+v[i]+u[i]
v[i]~dnorm(0,tauV)
resc[i]<-(yc[i]-muc[i])/sqrt(muc[i])
phc2[i]<-phc[i]}
u[1:m]~car.normal(adjc[],weic[],numc[],tauU)
for( k in 1: nsumc){weic[k]<-1}
a0~dnorm(0,tau0)
tau0~dgamma(4,0.05)
```

```
##PH district level
for (j in 1:p){
yph[j]~dpois(mph[j])
eph1[j]<-inprod(ec[],Label[,j])
mph[j]<-eph1[j]*thp[j]
log(thp[j])<-a0ph+vph[j]+uph[j]
res[j]<-(yph[j]-mph[j])/sqrt(mph[j])
vph[j]~dnorm(0,tauVPH)}
for (k in 1 :nsumph){weiph[k]<-1}
uph[1:p]~car.normal(adjph[],weiph[],numph[],tauph)
tauph~dgamma(4,0.05)
tauVPH~dgamma(4,0.05)
tauV~dgamma(4,0.05)
tauU~dgamma(4,0.05)
a0ph~~dgamma(4,0.05)
tau0ph~dgamma(4,0.05)}

##########################Model 3################################
model{
##county level
for (i in 1:m){
yc[i]~dpois(muc[i])
muc[i]<-ec[i]*thc[i]
log(thc[i])<-a0+v[i]+u[i]+uph[phc[i]]
v[i]~dnorm(0,tauV)
```

```
resc[i]<-(yc[i]-muc[i])/sqrt(muc[i])
phc2[i]<-phc[i]}
u[1:m]~car.normal(adjc[],weic[],numc[],tauU)
for( k in 1: nsumc){weic[k]<-1}
a0~dnorm(0,tau0)
tau0~dgamma(2,0.001)
##PH district level
for (j in 1:p){
yph[j]~dpois(mph[j])
label2[j]<-inprod(ec[],Label[,j])
mph[j]<-eph[j]*thp[j]
log(thp[j])<-a0ph+vph[j]+uph[j]
res[j]<-(yph[j]-mph[j])/sqrt(mph[j])
vph[j]~dnorm(0,tauVPH)}
for (k in 1 :nsumph){weiph[k]<-1}
uph[1:18]~car.normal(adjph[],weiph[],numph[],tauph)
tauph~dgamma(2,0.001)
tauVPH~dgamma(2,0.001)
tauU~dgamma(2,0.001)
tauV~dgamma(2,0.001)
a0ph~dnorm(0,tauph0)
tauph0~dgamma(2,0.001)}
##########################Model 4#################################
model{
##county level
```

```
for (i in 1:m){
yc[i]~dpois(muc[i])
muc[i]<-ec[i]*thc[i]
log(thc[i])<-a0+v[i]+uph[phc[i]]
v[i]~dnorm(0,tauV)
resc[i]<-(yc[i]-muc[i])/sqrt(muc[i])
phc2[i]<-phc[i]}
u[1:m]~car.normal(adjc[],weic[],numc[],tauU)
for( k in 1: nsumc){weic[k]<-1}
a0~dnorm(0,tau0)
tau0~dgamma(2,0.01)
##PH district level
for (j in 1:p){
yph[j]~dpois(mph[j])
mph[j]<-eph[j]*thp[j]
log(thp[j])<-a0ph+vph[j]+uph[j]
res[j]<-(yph[j]-mph[j])/sqrt(mph[j])
vph[j]~dnorm(0,tauVPH)}
for (k in 1 :nsumph){weiph[k]<-1}
uph[1:18]~car.normal(adjph[],weiph[],numph[],tauph)
tauph~dgamma(2,0.01)
tauVPH~dgamma(2,0.01)
tauU~dgamma(2,0.01)
tauV~dgamma(2,0.01)
a0ph~dnorm(0,tauph0)
```

```
tauph0~dgamma(2,0.01)}
##########################Model 5#################################
model{
##PH district level
for( i in 1:18){
yph[i]~dpois(muph[i])
log(muph[i])<-log(eph[i])+log(thph[i])
thph[i]<-exp(aph0+uph[i]+vph[i])
vph[i]~dnorm(0,tau.vph) }
##county level
for( j in 1:159){
yc[j]~dpois(muc[j])
log(muc[j])<-log(ec[j])+log(thc[j])
thc[j]<-exp(ac0+uc[j]+vc[j])
vc[j]~dnorm(0,tau.vc)
phc2[j]<-phc[j] }
for (k in 1 :nsumph){weiph[k]<-1}
uph[1:18]~car.normal(adj[],weiph[],num[],tauph)
tau.vph~dgamma(5,0.08)
tauph~dgamma(5,0.08)
aph0~dnorm(0,tauph0)
tauph0~dgamma(5,0.08)
for (o in 1 :nsumc){weic[o]<-1}
uc[1:159]~car.normal(adj2[],weic[],num2[],tauc)
tau.vc~dgamma(5,0.08)
```

```
tauc~dgamma(5,0.08)
ac0~dnorm(0,tauc0)
tauc0~dgamma(5,0.08) }
```

**General code used for all simulations:**

```
library(BRugs)
setwd()
thetaT<-rgamma(159,1)
##for the next condition thetaT values changed to (159,3)
##data entered
attach(data)
store1<-matrix(0, nrow=4, ncol=100) ##stored the DIC and pD statistics in a matrix
# ysim values generated using the code below
ysim<-matrix(nrow=159,ncol=10)
mu<-rep(1,length=159)
for (j in 1:10){
for (i in 1:159){
ysim[i,j]<-rpois(1,mu[i])
mu[i]<-(ec[i]*thetaT[i])}}
####code to set model simulations###
mu<-rep(1,length=159)
ycS<-rep(1,length=159)
yc<-rep(1,length=159)
yph<-rep(1,length=18)
for(i in 1:159){mu[i]<-(ec[i]*thetaT[i])}
```

```
for (j in 1:10){

for (i in 1:159){

# Poisson likelihood for true model

ysim[i,j]<-rpois(1,mu[i])

ycS[i]<-ysim[i,j]

yc[i]<-ycS[i]}

for( k in 1: 18){yph[k]<-ycS[]%*%Label[,k]}

data1$yc<-yc

data1$yph<-yph

#################BRugsFit function#########################

parametersToSave=c("thph","thc","deviance","a0","yc","yph","thc","thph","muc","mph",
"v","u")

asd<-BRugsFit(modelFile="model_1.txt", data=data, inits=NULL, numChains = 2,
parametersToSave,

nBurnin = 20000, nIter = 5000, nThin = 1,DIC = TRUE)

store1[1,j]<-asd$DIC[1,3]

store1[2,j]<-asd$DIC[1,4]

store1[3,j]<-asd$DIC[2,3]

store1[4,j]<-asd$DIC[2,4]}

####storing values for DIC and pD

ycDIC<-store1[1,]

ycPD<-store1[2,]

yphDIC<-store1[3,]

yphPD<-store1[4,]
```