NEGATIVE SPATIAL AUTOCORRELATION AND ITS IMPACTS ON GEOREFERENCED DATA ANALYSES: WITH CASE STUDIES OF CANCER INCIDENCES

by

Lan Hu

APPROVED BY SUPERVISORY COMMITTEE:

_____

Yongwan Chun, Chair

_____

Daniel A. Griffith

_____

Dohyeong Kim

_____

Fang Qiu

I dedicate this dissertation to my parents.

.

NEGATIVE SPATIAL AUTOCORRELATION AND ITS IMPACTS ON GEOREFERENCED DATA ANALYSES: WITH CASE STUDIES OF CANCER INCIDENCES

by

LAN HU, BS, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

GEOSPATIAL INFORMATION SCIENCES

THE UNIVERSITY OF TEXAS AT DALLAS

May 2020

ACKNOWLEDGMENTS

From the bottom of my heart I would like to say a big thank you to my supervisor, Dr. Yongwan Chun, who has continuously supported and assisted me throughout all my research projects. I would also like to thank Dr. Daniel A. Griffith, a coauthor of my research papers, for providing advice on research and guiding analyses and manuscript developments. Finally, many thanks to Dr. Dohyeong Kim and Dr. Fang Qiu for serving as committee members and for their comments and suggestions on this dissertation.

March 2020

NEGATIVE SPATIAL AUTOCORRELATION AND ITS IMPACTS ON GEOREFERENCED DATA ANALYSES: WITH CASE STUDIES OF CANCER INCIDENCES

Lan Hu, PhD
The University of Texas at Dallas, 2020

Supervising Professor: Yongwan Chun

Spatial autocorrelation has been a popular research topic in spatial analysis for decades, mainly attributable to its frequent detection in georeferenced phenomenon. In addition, the presence of spatial autocorrelation complicates statistical analysis, because it violates the independence assumption in conventional statistics. However, most research, to date, focus on positive spatial autocorrelation while works about negative spatial autocorrelation relatively are scant. Negative spatial autocorrelation has long been neglected in literature, largely because of its rare observation in empirical data.

This dissertation aims to contribute to the understanding of negative spatial autocorrelation with two major goals. One goal is to examine the impacts of spatial autocorrelation on statistical random variables with both positive and negative spatial autocorrelation being assessed and contrasted with each other. The literature is replete with acknowledgments that positive spatial autocorrelation inflates the variance of a random variable, and it also may alter other random variable distributional properties. Moreover, due to different quantifications of negative and positive spatial autocorrelation, their impacts on random variables are expected to differ. The other goal is to explore simultaneous materialization of negative spatial autocorrelation with positive spatial autocorrelation in empirical data, and a potential treatment of spatial autocorrelation mixture in spatial statistical analysis. Moran scatterplot and local Moran statistics can furnish efficient methods to uncover spatial autocorrelation mixture patterns. Other statistical

methodologies are also employed to identify and capture negative spatial autocorrelation, including a spatial autoregressive model with two-spatial autocorrelation-parameters, the mixed regressive spatial autoregressive moving average model, and Moran eigenvector spatial filtering method.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Spatial autocorrelation (SA) has been a popular research topic in spatial analysis for decades, mainly attributable to its frequent detection in georeferenced phenomenon. In addition, the presence of SA complicates statistical analysis, because it violates the independence assumption in conventional statistics (Griffith 1987). The literature (e.g., Griffith 2011) argues that SA distorts the distribution of a random variable (RV). One noticeable impact is that it dramatically inflates variance of RV. An improper treatment of a spatial correlated pattern would lead to underestimated error sum of squares, which increases the probability of rejecting the null hypothesis in statistical analysis (Griffith 1987).

However, most research, to date, focus on positive SA while works about negative SA (NSA) relatively are scant. NSA has long been neglected in literature, largely because of its rare observation in empirical data. NSA refers to phenomenon that the values of a variable tend to be dissimilar when they are geographically proximate (Griffith 2019). By nature, NSA materializes with a spatial competitive process. For example, NSA has been discovered in federal grants competition among local governments (Boarnet and Glazer 2002), in forest competition for light (Montgomery and Chazdon 2001), and in research activities where researchers engaged in competition (Elhorst and Zigova 2014).

NSA also is found in georeferenced data, in a mixed form with PSA regardless of a global SA pattern (that is, essentially is positive, negative, or appears to be absent. For example, Griffith and Arbia (2010) use a two-SA-parameter SAR specification to model ratios of actual municipality areas to Thiessen polygon areas of Puerto Rico, which exhibits significant and moderate global NSA. This model successfully uncovers hidden PSA that counterbalance a NSA component. Jacob et al. (2011) posit an MESF model specification to capture hidden PSA and NSA components in their georeferenced data. Their results reveal that PSA and NSA ESFs are of importance equally, and the data exhibit a trivial near-zero global SA. Griffith (2006)

summarizes three examples of hidden NSA, and suggests that a local Moran's *I* plot can be used to uncover mixture patterns, where significant high-low/low-high local Moran's *I* are detected along with high-high/low-low local Moran's *I* values.

This dissertation aims to contribute to understanding of negative SA with two major goals. One goal is to examine the impacts of SA on statistical RVs with both positive and negative SA being assessed and contrasted with each other. The literature is replete with acknowledgments that positive SA inflates the variance of a RV, and it also may alter other RV distributional properties. Moreover, due to different quantifications of negative and positive SA, their impacts on RVs are expected to differ. The other goal is to explore simultaneous materialization of negative SA with positive SA in empirical data, and a potential treatment of SA mixture in spatial statistical analysis. Moran scatterplot and local Moran statistic can furnish efficient methods to uncover SA mixture patterns (e.g., Griffith 2006). In the literature, other statistical methodologies are also employed to identify and capture negative SA, including a spatial autoregressive model with two-SA-parameters (e.g., Griffith and Arbia, 2010), the mixed spatial autoregressive moving average (SARMA) model (Kao 2016; Kao and Bera 2016), and Moran eigenvector spatial filtering (MESF) method (e.g., Jacob et al., 2011). Specifically, the below three research topics about negative SA are investigated in this dissertation.

First, this dissertation examines the impact of SA on the beta and multinomial RVs. Its influence has been investigated for three popular distributions in geospatial data analysis: normal, Poisson, and binomial distributions. In contrast, much less is known about its effects on the two RVs that are utilized in GIScience research, i.e., the beta and the multinomial distributions. The beta distribution—which is considered to be very flexible because it can mimic a uniform, exponential, sinusoidal, and normal RV—can be utilized to analyze the radiance of a remotely sensed image, for example. The multinomial distribution, a generalization of the binomial distribution, has been widely used for land use classification in order to describe land use change. This dissertation extends the investigation about the effects of SA to beta and multinomial RVs. As it is indicated in the literature that RV impacts of negative SA may differ from those of positive SA, at least, for some RVs including Poisson RV (e.g., Chun and Griffith

2018), this dissertation evaluates and contrasts the impacts of both positive SA and negative SA through simulation experiments.

Second, this dissertation investigates model specifications that can accommodate positive and negative SA simultaneously. In regression analysis, the presence of SA violates the principle assumption, i.e., independence of observations, which leads to biased model estimation if SA is not appropriately addressed. For example, in a spatial cancer data analysis, regression techniques frequently are utilized to investigate associations between cancer incidence rates and potential risk factors, a Poisson or negative binomial model is preferred over a linear regression model because the former can incorporate heterogeneous population sizes. However, conventional statistical model specifications for a Poisson or negative binomial model are unable to accommodate SA, and their popular extensions to model cancer rates (e.g., spatial autoregressive model) often fail to fully capture SA with a single SA parameter, especially when a SA mixture exists. This dissertation applies MESF methodology to investigate breast cancer incidences in Broward County, Florida. While the cancer rates are globally positively autocorrelated, the proposed spatial model results reveal that the model specification including both positive (filtering positive SA) and negative (filtering negative SA) eigenvectors yields the best model performance. This indicates the presence of a mixture SA pattern in the breast cancer data.

Third, this dissertation assesses a SA mixture pattern with a random effect (RE) model and MESF methodology. A RE model furnishes an efficient alternative for a space-time analysis, and RE can capture unexplained SA in a model specification when relevant covariates that contains considerable SA are unavailable. This dissertation investigates geographic variation in Florida lung cancer incidence for the time period 2000-2011 using RE models. In doing so, a MESF technique also is utilized, which can allow a decomposition of RE terms into spatially structured (SSRE) and spatially unstructured (SURE) components. The analysis results confirm that RE models capture a substantial amount of variation in the cancer data. Furthermore, the results suggest that spatial pattern in the cancer data displays a mixture of positive and negative SA, although the global map pattern of the RE term may appear random.

The next three chapters are individual papers to investigate the three research topics. Chapter 2 is published in *Geographical Analysis (Hu, Griffith, and Chun 2019)*. Chapter 3 has been accepted for publication in *Journal of Geographical Systems (Hu, Chun, and Griffith 2020)*. Chapter 4 is published in *International Journal of Environmental Research and Public Health (Hu, Griffith, and Chun 2018)*. Chapter 5 presents anticipated implications, limitations, and conclusions of this dissertation as well as a summary of the three chapters.

CHAPTER 2

IMPACTS OF SPATIAL AUTOCORRELATION IN GEOREFERENCED BETA AND MULTINOMIAL RANDOM VARIABLES

Authors – Lan Hu, Daniel A. Griffith, Yongwan Chun

Geospatial Information Sciences Program, GR 31

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

Reproduced with permission from John Wiley and Sons (4787190601394).

Hu, L., Griffith, D. A., & Chun, Y. (2019). Impacts of Spatial Autocorrelation in Georeferenced Beta and Multinomial Random Variables. *Geographical Analysis*. 1-21

ABSTRACT

The literature is replete with acknowledgments that spatial autocorrelation (SA) inflates the variance of a random variable (RV), and that it also may alter other RV distributional properties. In most studies, impacts of SA have been examined only for the three most commonly used distributions: the normal, Poisson (and its negative binomial counterpart), and binomial distributions; much less is known about its effects on two other RVs that are utilized in GIScience research: the beta and the multinomial. The beta distribution— which is considered to be very flexible because it can mimic a uniform, exponential, sinusoidal, and normal RV—can be utilized to analyze the radiance of a remotely sensed image, for example. The multinomial distribution, a generalization of the binomial distribution, has been widely used for land use classification, and to describe land use change. The literature also suggests that RV impacts of negative SA, a neglected topic in spatial analysis, may differ from those of positive SA, at least for some RVs (e.g., the Poisson RV). The purpose of this article is to extend the investigation of effects of SA to beta and multinomial RVs, with both positive SA and negative SA assessed and contrasted with each other, using simulation experiments. The simulation experiments are designed to support this assessment. One of the major discoveries is that impacts of positive SA and negative SA behave similarly when a RV conforms to a normal distribution; however, maximum negative SA is unable to materialize for asymmetric RV, whereas positive SA always converges upon its maximum.

2.1 Introduction

In a statistical analysis, the identification of an underlying data distribution often is necessary and critical. Statisticians generally are interested in computing the first four statistical moments (the mean, variance, skewness, and kurtosis) because these moments help summarize the frequency distribution of data generated by a given random variable (RV). One complication is that spatial autocorrelation (SA) frequently is detected in georeferenced data, which is known to contribute to an inflated variance for normal RVs, and is a source of overdispersion for Poisson and binomial RVs. Impacts of SA on histograms of normal, Poisson and binomial RVs have been investigated systematically by Griffith (2011), and Chun and Griffith (2018). They observe that the mean tends not to be impacted by SA; however, SA inflates the variance, and may distort the skewness and/or kurtosis of a histogram.

To date, impacts of SA on other RVs have not yet been explored in detail. The beta and multinomial distributions are two statistical distributions that have increasingly appeared in the literature of a number of disciplines, including ecology, sociology, epidemiology, and GIScience. For example, the beta distribution is applied to measure spatial heterogeneity of ecological objects (e.g., Shiyomi et al., 2000; Chen et al., 2008). It also is incorporated into a Bayesian setting to examine if time and geographic distance influence the evolution of disease (Branscum et al., 2007). The multinomial distribution commonly is utilized to model land cover/soil classes (e.g., Dendoncker et al., 2007; Debella-Gilo and Etzelmüller 2009). And this RV also is a popular descriptor of spatial pattern and the distribution of disease (e.g., Kazembe and Namangale 2007; Cordeiro et al., 2011).

This paper summarizes research extending the investigation of SA impacts to beta and multinomial RVs with simulation experiments. This investigation was conducted for negative SA as well as positive SA, particularly because differences between them have been discussed in the literature. For example, in many cases, the range of the extreme Moran coefficient value for negative SA is between −0.5 and −1, whereas it often goes beyond 1 for positive SA (Griffith 2017). Because negative SA is expected to have a different effect on a RV distribution, the

impacts of both positive SA and negative SA were investigated and are compared. In addition, this research also investigated impacts of SA on beta and multinomial RVs with different shapes; Chun and Griffith (2018) discuss that a distribution has an increasing ability to capture maximum SA as it increasingly better conforms to a normal distribution. The simulation experiments were designed employing a regular surface partitioning.

2.2 Literature review

SA measures correlation between values of a variable attributable to their relative geographical proximity (Griffith 1987). SA has drawn research attention for decades for three major reasons. First, SA describes the spatial pattern of a geographic phenomenon, such as hot spot analysis (e.g., Myers et al., 2000). Second, SA supports spatial prediction of unknown georeferenced values; accounting for SA also yields more reliable spatial prediction outcomes (Cressie 1991). For example, kriging, considered to be one of the most robust interpolation methods, is built upon the notion of SA. Third, the presence of SA violates the classical statistical assumption about independence of residuals (or observations in general). A major impact of this violation is that the error sum of squares is underestimated, therefore inflating standard errors, which increases the probability of rejecting the null hypothesis in a geospatial statistical analysis (Griffith 1987).

To date, attention to SA has focused on positive SA, mainly because it is frequently observed in empirical georeferenced data. Positive SA has been detected in various geographic phenomena, including social variables such as crime (e.g., Murray et al., 2001), population migration (e.g., Gorman and Speer 2001; LeSage and Pace 2008), economic activities such as house prices, household income, and employment status (e.g., Can 1990; Conley and Topa 2002; Cohen and Paul 2004), and diseases (e.g., Jones et al., 2008). Griffith (1987) proposes that socio-economic phenomena tend to exhibit moderate positive SA due to the way they are geographically distributed and aggregated. Positive SA also is observed in remotely sensed data, in which the degree of SA tends to be very strong (Griffith and Chun 2016).

Because of its rare detection in empirical data, negative SA essentially has been neglected in spatial analysis work for a long time. The scant literature about negative SA is well summarized in Chun and Griffith (2018). Negative SA naturally materializes with a competitive spatial process (e.g., Montgomery and Chazdon 2001; Boarnet and Glazer 2002; Elhorst and Zigova 2014); it also is found in georeferenced data, mixed with positive SA when global SA essentially is weakly positive, or negative, or appears to be absent. For example, Griffith (2006) summarizes three examples of hidden negative SA, and suggests a local Moran's I plot can be used to uncover the mixture pattern, where the presence of significant high-low/low-high local Moran's I coincides with high-high/low-low local Moran's I values. Griffith and Arbia (2010) use a two-SA-parameter spatial simultaneous autoregressive model (SAR) specification to model ratios of actual municipality and Thiessen polygon areas of Puerto Rico that exhibit significant and moderate global negative SA; this model successfully uncovers a SA mixture in which a positive SA counterbalances a negative SA component. Jacob et al. (2011) posit a Moran eigenvector spatial filtering (MESF) model specification to capture a mixture of positive SA and negative SA components in their georeferenced data; their results reveal that positive SA and negative SA spatial filters are of equal importance, with their data exhibiting trivial near-zero global SA.

Beta regression has been applied to describe many socio-economic phenomena, such as the poverty (e.g., Do et al., 2013) and migration rates (e.g., Kalhori and Mohammadzadeh 2016), which are continuous variables and constrained to the interval [0, 1]. This list can be extended by rescaling any limited variable to the interval (0, 1). More specifically, a variable, Z, within an interval [a,b], can be transformed as $Y=(Z-a)/(b-a)$ to the range [0, 1], and then treated as a beta RV (Cepeda-Cuervo and Núñez-Antón 2013). Griffith (2011) argues that the beta RV also is a good choice to analyze the radiance of a remotely sensed image. A beta RV can be mixed with other RVs (e.g., binomial and beta RVs)—either as a finite combination, or a parametric distribution—to model the underlying spatial process of an overdispersed event. For example, Kaiser et al. (2002) propose a spatial beta-binomial mixture to model the number of affected trees in forest-health monitoring. Griffith and Chun (2016) utilize a beta-beta mixture to model uncertainty of the SA parameter ρ, which is transformed to the range $[0, (1+\rho\_max)/2]$; this

transformed parameter exhibits some properties of an overdispersed beta RV that can be easily defined by the two shape parameters of the beta distribution.

The multinomial distribution has been utilized extensively in epidemiology (e.g., Hedeker 2003; Blazer and Wu 2009), tourism (e.g., Lee et al., 2002), transportation (e.g., Bolduc 1999), and in land use change evaluation in geography (e.g., Verburg et al., 2004; Millington et al., 2007). Spatial multinomial models also have been applied to account for spatial effects. For example, Kavousi et al., (2011) introduce an auto-multinomial model to analyze multivariate lattice discrete data; their results suggest that the auto-model outperforms an aspatial model. Sinha (2017) uses a MESF multinomial specification to model land use change in Collin County, TX.

2.3 Methodology

The research summarized in this paper utilized a MESF to account for SA latent in beta and multinomial RVs. Simulation experiments were designed to assess how SA affects beta and multinomial RVs' histograms; these experiments included 1,000 replications in order to exploit the Law of Large Numbers.

2.3.1. The beta and multinomial distributions

The beta RV is a family of continuous distributions bounded by the interval [0, 1]; it has two positive parameters that control the shape of its specific distribution. If these two parameters are equal and large, then it can be approximated by a univariate normal distribution. It is similar to a continuous version of a binomial distribution. The multinomial distribution is a generalization of the binomial distribution, it can be approximated by the multivariate normal distribution. Their probability density/mass function, and first four statistical moments appear in Table 2.1.

Table 2.1. The beta and multinomial distributions

	The beta distribution	The multinomial distribution
Probability function	$f(x;\ \alpha,\beta) = \frac{\Gamma(\alpha+\ \beta)}{\Gamma(\alpha)\Gamma(\beta)}\ x^{\alpha-1}(1-x)^{\beta-1},$ $0 < x < 1,\ \alpha > 0,\ \beta > 0$	$f(x_1,\cdots,x_k;\ n;\ \pi_1,\cdots\pi_k) = \frac{n!}{x_1!,\cdots,x_k!}\pi_1^{x_1}\cdots\pi_k^{x_k},\ \sum_{j=1}^{k} x_j = n;\ \sum_{j=1}^{k}\pi_j = 1$
The first two moments	$E(x) = \frac{\alpha}{\alpha+\ \beta}$	$E(x_j) = n\pi_j$
	$var(x) = \frac{\alpha\beta}{(\alpha+\ \beta)^2\ (\alpha+\ \beta+1)}$	$ar(x_j) = n\pi_j(1-\pi_j)$
The third and fourth standardized moments	$skewness(x) = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$	$skewness(x_j) = \frac{1-\pi_j}{\sqrt{n\pi_j(1-\pi_j)}}$
	$excess\ kurtosis(x)$ $= \frac{6[(\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$	$excess.kurtosis(x_j) = \frac{1-6\pi_j(1-\pi_j)}{n\pi_j(1-\pi_j)}$

α and β are the shape parameters, $\Gamma(\cdot)$ is the gamma function. n denotes the total number of cases, x_j denotes the number of cases with outcome j, π_j denotes the probability for outcome j.

The beta distribution is very flexible, and can mimic a uniform, exponential, sinusoidal, normal, and skewed RV by combining different values of α and β (Figure 2.1). When $\alpha = \beta$, and $\alpha > 1$, $\beta > 1$, the beta distribution conforms to a symmetric distribution. When $\alpha > \beta$, the beta distribution is negatively skewed, and converges to a negative exponential RV when $\alpha = 1,\ \beta > 1$. When $\alpha < \beta$, the beta distribution is positively skewed. When $\alpha = \beta = 1$, the beta distribution reduces to a continuous uniform distribution. When $\alpha < 1$ and $\beta < 1$, the beta distribution becomes a sinusoidal type distribution.

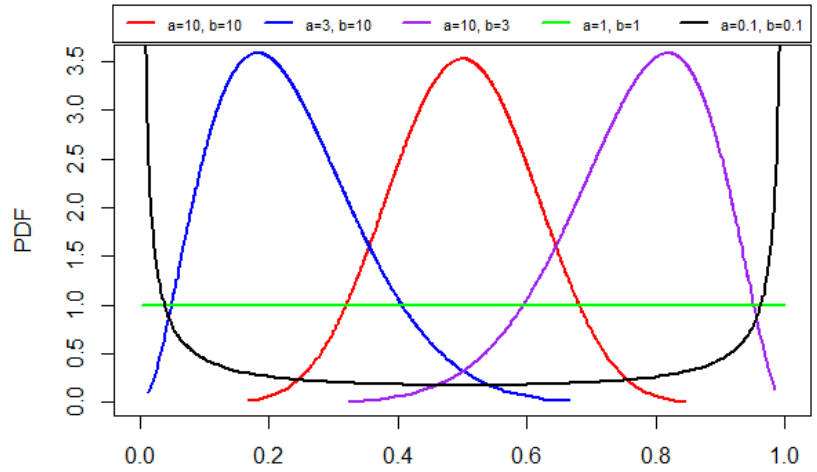


Figure 2.1. Probability densities of beta RVs with selected parameter values. (a=b=1: a continuous uniform distribution; a=b>1: a symmetric distribution; a=b<1: a sinusoidal distribution; a<b: a positively skewed distribution; a >b>1: a negatively skewed distribution)

In statistical modeling, statisticians are interested in positing a connection between covariates and the mean of the beta RV (Ferrari and Gribari-Neto 2004). If a variable is assumed to follow a beta distribution with mean μ, then the mean can be written as

$$g(\mu_i) = \sum_{i=1}^{n} \boldsymbol{x}_{ij}\, \boldsymbol{\gamma}_j, \tag{1}$$

where $\boldsymbol{x}_{ij}$ denotes covariates, $\boldsymbol{\gamma}_j$ denotes unknown covariate coefficients, and $g(\cdot)$ is a monotonic link function. Because the beta RV is restricted to the interval [0, 1], the logit link function: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, is popularly used; hence the mean can be rewritten as

$$\mu_i = \frac{e^{x_{ij}\gamma_j}}{1 + e^{x_{ij}\gamma_j}}. \tag{2}$$

SA can be accounted for in the linear combination appearing in the exponent of *e*, similar to the auto-binomial specification.

Like binary logistic regression, multinomial logistic regression is used to predict the probabilities of categories based upon multiple covariates (Greene 2003 pp. 720-723). Typically, one of the outcomes is picked as a reference, log-odds for all other outcomes are computed based on the selected reference, and then the log-odds are linked to covariates with a linear function such that

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\gamma}_j, j = 2,3,\cdots k; \eta_{i1} = 0, \quad (3)$$

The following probabilities for outcomes can be derived from eq (3):

$$\begin{aligned} \pi_{i1} &= \frac{1}{1+ e^{\eta_{i2}+\cdots+e^{\eta_{ik}}}}, \\ \pi_{i2} &= \frac{e^{\eta_{i2}}}{1+ e^{\eta_{i2}+\cdots+e^{\eta_{ik}}}}, \\ &\vdots \\ \pi_{ik} &= \frac{e^{\eta_{ik}}}{1+ e^{\eta_{i2}+\cdots+e^{\eta_{ik}}}}. \end{aligned} \quad (4)$$

Again, SA can be accounted for in the linear combination appearing in the exponent of *e*, similar to the auto-binomial specification.

2.3.2. Moran eigenvector spatial filtering

MESF is a spatial statistical methodology that introduces a set of proxy variables, which are eigenvectors extracted from a transformed *n*-by-*n* spatial weights matrix $\mathbf{C}$, into a regression model specification to capture SA and transfer it from residuals to the mean response term. The transformed spatial weight matrix is expressed as

$$\mathbf{MCM} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n), \quad (5)$$

where $\mathbf{I}$ is an *n*-by-*n* identity matrix, $\mathbf{1}$ is a *n*-by-1 vector of ones, n is the number of areal units, and T is the matrix transpose operator. MESF is sufficiently flexible to account for positive SA, negative SA, or a mixture of both. The n eigenvectors represent distinct underlying map patterns, and their corresponding eigenvalues represent their levels of SA. These eigenvectors are mutually orthogonal and uncorrelated; the first eigenvector, E_1, is the set of real numbers that has the maximum possible positive SA; the j^{th}, E_j, is the set of real numbers that has the j^{th} largest MC value of any vector that is uncorrelated and orthogonal with all of its $j-1$ preceding eigenvectors; and, the n^{th} eigenvector, E_n, is the set of real numbers that has the largest negative SA (Griffith 2003).

Eigenvectors are included as covariates in a regression model specification to account for SA. A linear MESF model without covariates is specified as

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{E}_k \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad (6)$$

where $\mathbf{Y}$ denotes the *n*-by-*1* response variable vector, $\mathbf{E}_k$ denotes an *n*-by-*k* matrix containing a set of selected eigenvectors, β_0 and *n*-by-*k* vector $\boldsymbol{\gamma}$ denote regression coefficients, $\mathbf{E}_k\boldsymbol{\gamma}$ denotes a constructed eigenvector spatial filter (ESF) that captures SA, and $\boldsymbol{\epsilon}$ is a *n*-by-*1* vector of non-spatial random errors. In addition, because the means of the eigenvectors ($\mathbf{E}_k$) are zero, the addition of the eigenvectors does not have an impact on mean responses. This specification filters SA out of the conventional residuals ($\mathbf{E}_k\boldsymbol{\gamma} + \boldsymbol{\epsilon}$), and adds it to the mean response ($\beta_0\mathbf{1} + \mathbf{E}_k\boldsymbol{\gamma}$); in other words, the model specification retains latent SA while constructing stochastic residuals that mimic independent ones. The k eigenvectors can be identified from a candidate eigenvector set with a stepwise selection procedure, which selects considerably fewer eigenvectors than n (Chun et al., 2016). MESF also is adopted by econometricians in their research (e.g., Eckey et al., 2006; Crespo and Feldkircher 2013; Pace et al., 2013), and furthermore, Paez (2018) discusses that MESF furnishes an effective approach to identify omitted but potentially substantive covariates.

2.3.3. The simulation experiment design

Simulation experiments are designed to fulfil research purposes. The following three important factors are considered before implementing the simulation experiments for this paper. First, simulation experiments are conducted using a 30-by-30 square tessellation (Figure 2.2), and the rook adjacency connectivity definition to produce the spatial weight matrix $\mathbf{C}$. Second, a response variable with positive SA or negative SA only is generated with a simultaneous autoregressive model (SAR), such that

$$\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}, \quad (7)$$

where ρ is a spatial autocorrelation parameter, $\mathbf{W}$ is a row standardized version of matrix $\mathbf{C}$, $\boldsymbol{\varepsilon}$ is a vector of iid normal random errors MESF is used to approximate the SA component [see eq. (6)]. Third, the SA parameter ρ is set to different values to cover weak, moderate, and strong

positive SA and negative SA: 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, and −0.1, −0.3, −0.5, −0.7, −0.9, −0.95. To stabilize the eigenvectors, the constructed ESF term can be adjusted as follows:

$$ESF^{adj} = 2 * \left[\frac{ESF - ESF_{\min}}{ESF_{\max} - ESF_{\min}}\right]^{\delta} - 1, \tag{8}$$

where $\delta = \ln\left(\frac{1}{2}\right) / \ln\left(\frac{-ESF_{\min}}{ESF_{\max} - ESF_{\min}}\right)$, and $ESF_{\min}$ and $ESF_{\max}$ are the minimum and maximum values in an ESF. This adjustment rescales ESF^{adj} to the range [−1, 1], centering it around 0. As mentioned previously, the mean of beta RVs and the probabilities of multinomial outcomes can be linked to covariates through a link function [see eqs. (2) and (3)].

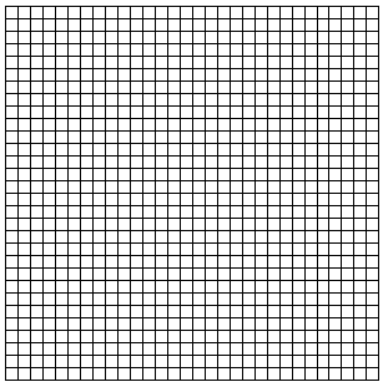


Figure 2.2. A 30-by-30 square tessellation surface

In the MESF specification in eq. (6), the two parameters (α, β) of a beta distribution can be estimated with the method of moments estimator as follows:

$$\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+e^{-b_0-ESF_i^{adj}}}\right)\left[\frac{(n-1)\sum_{i=1}^{n}\frac{e^{-b_0-ESF_i^{adj}}}{1+e^{-b_0-ESF_i^{adj}}}}{n\sum_{i=1}^{n}\left(\frac{1}{1+e^{-b_0-ESF_i^{adj}}}\right)^2 - \left(\sum_{i=1}^{n}\frac{1}{1+e^{-b_0-ESF_i^{adj}}}\right)^2} - 1\right],$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\frac{e^{-b_0-ESF_i^{adj}}}{1+e^{-b_0-ESF_i^{adj}}}}{\sum_{i=1}^{n}\frac{1}{1+e^{-b_0-ESF_i^{adj}}}}\hat{\alpha}.$$

Accordingly, the empirical moments of a beta distribution with the ESF term need to be updated from the standard moments in Table 2.1. Similarly, the empirical moments of a multinomial distribution with embedded SA can be expressed as in Table 2.2.

Table 2.2. The empirical moments of a multinomial RV with SA embedded

Mean of category j	$\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}$
Variance of category j	$\left(\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}\right)\left(1-\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}\right)$
Skewness for category j	$\frac{1-2\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}}{\sqrt{\left(\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}\right)\left(1-\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}\right)}}$
Excess kurtosis for category j	$\frac{1-6\left(\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}\right)\left(1-\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}\right)}{\left(\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}\right)\left(1-\frac{\sum_{i=1}^{n} \frac{1}{1+e^{-b_{0j}-ESF_{ij}^{adj}}}}{n}\right)}$

In a specific analysis, weights are introduced to control the shape of the statistical distribution and the level of SA embedded. For beta RVs, Guolo and Varin (2014) propose a Gaussian copula regression to model serial dependence. MESF is adopted to introduce SA—which is far more complicated because it is two-dimensional and multidirectional—to the beta RVs; in keeping with mathematical statistical generalized linear model theory, essentially the covariates term is substituted with the adjusted ESF term, eq. (2) becomes

$$\mu_i = \frac{ae^{w_0 ESF_i^{adj}}}{ac + ae^{w_0 ESF_i^{adj}}},$$

where a and c denote positive constants that are set to different values to control the shape of a resulting beta RV. The weight, w_0, controls the level of SA embedded in a simulated beta RV. The two shape parameters can be specified as:

$$\alpha_i^{adj} = ae^{w_0 ESF_i^{adj}}, \beta = ac, \quad (9)$$

If c = 1, the map mean of the generated RV is approximately $\frac{1}{2}$ because the adjusted ESF contains both positive and negative values. If c > 1, the generated RV conforms on a positively skewed distribution, and if c < 1, generated RV conforms to a negatively skewed distribution. Multinomial RVs with three mutually exclusive outcomes are examined in this study. When k = 3, and if the probabilities for the first two outcomes are the same (e.g., $\pi_{i1} = \pi_{i2}$), then specifications of probabilities (with the adjusted ESF term incorporated) can be expressed as

$$\pi_{i1} = \frac{1}{2+ e^{w_0 ESF_i^{adj}}}, \pi_{i2} = \frac{1}{2+ e^{w_0 ESF_i^{adj}}}, \pi_{i3} = \frac{e^{w_0 ESF_i^{adj}}}{2+ e^{w_0 ESF_i^{adj}}}. \quad (10)$$

If probabilities vary among outcomes (e.g., 10 $\pi_{i1} = \pi_{i2}$), then they can be specified as:

$$\pi_{i1} = \frac{1}{11+ e^{w_0 ESF_i^{adj}}}, \pi_{i2} = \frac{10}{11+ e^{w_0 ESF_i^{adj}}}, \pi_{i3} = \frac{e^{w_0 ESF_i^{adj}}}{11+ e^{w_0 ESF_i^{adj}}}. \quad (11)$$

Equality or inequality of probabilities controls the frequency of randomly sampled observations occurring in each outcome class.

2.4 Results

For simulated beta RVs, the constant a has a range spanning 0.01 to 100,000 so that all distributions that a beta RV can mimic are covered (see Figure 2.1). For simulated multinomial RVs, the total number of trials, N_{tr}, ranges from 1 to 100,000 to treat the smallest through relatively large population size cases.

2.4.1. Simulation results for beta RVs

Figure 2.1a and 2.3b illustrate that when the mean of a beta RV is 0.5 ($\boldsymbol{c = 1}$), both positive SA and negative SA achieve their maximum levels as $\boldsymbol{a}$ increases. This trajectory occurs because beta RVs converge to a normal RV as $\boldsymbol{a}$ increases, allowing achievement of a full range of SA.

For positively skewed beta RVs (e.g., $\boldsymbol{c = 10}$ and $\boldsymbol{c = 100}$), positive SA starts to converge slightly sooner than for symmetric beta RVs. However, it experiences a delayed convergence for negatively skewed beta RVs (e.g., $\boldsymbol{c = 0.1}$ and $\boldsymbol{c = 0.01}$), and the delay becomes more pronounced as the level of negative skewness gets more severe. For both positively and negatively skewed beta RVs, maximum negative SA fails to materialize. The two sets of slightly skewed beta RVs ($\boldsymbol{c = 10}$ and $\boldsymbol{c = 0.1}$) deviate less from their theoretical maximums, whereas gaps between the observed extremes and the theoretical maximums are larger for the more skewed RVs ($\boldsymbol{c = 100}$ and $\boldsymbol{c = 0.01}$). These two figures suggest that positive SA and negative SA behave differently when beta RVs are skewed, with negative SA no longer converging. This result possibly can be explained by how they deviate from a normal distribution.

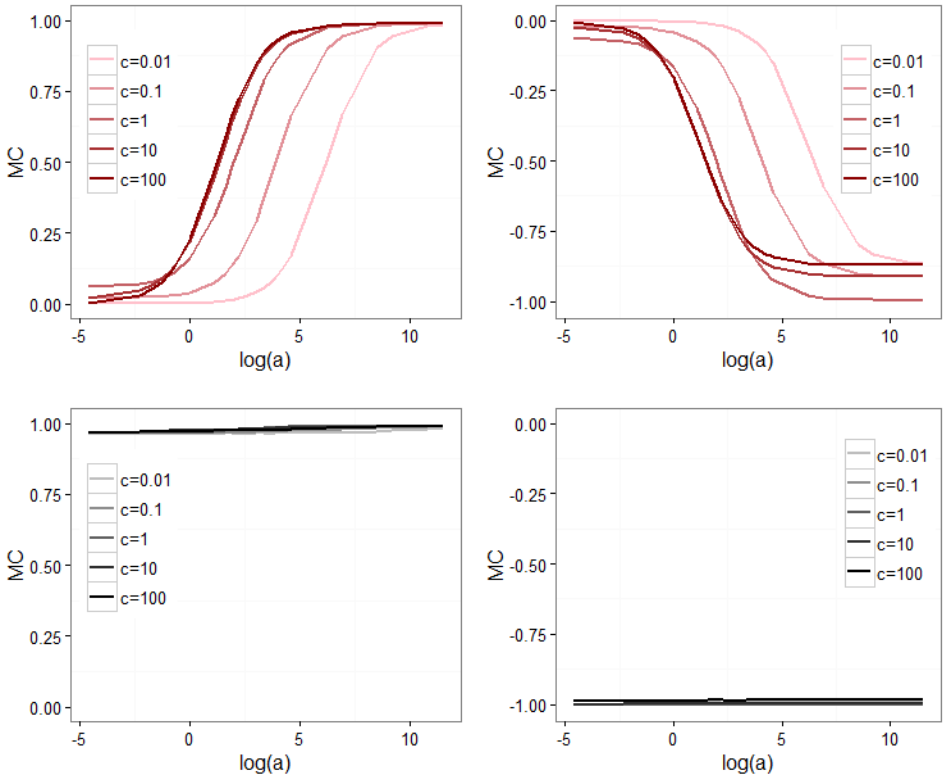


Figure 2.3. The range of SA for beta RVs. Top left (a): PSA. Top right (b): NSA. Middle left (c): maximum PSA. Middle right (d): maximum NSA.

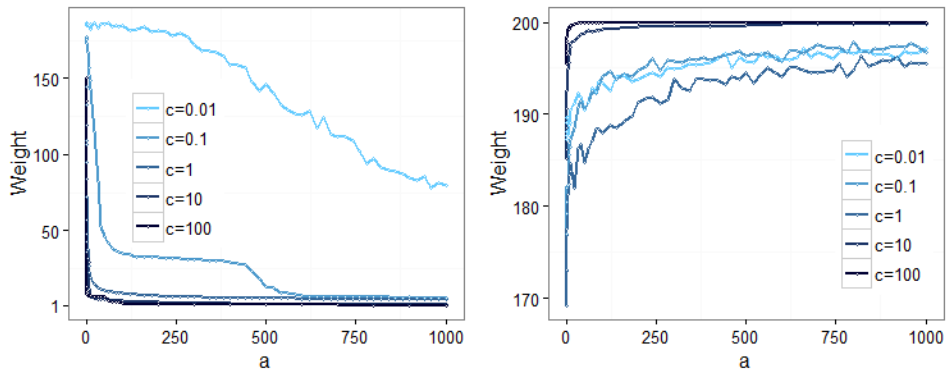


Figure 2.3. (Continued) Bottom left (e): the weights needed to achieve a maximum PSA. Bottom right (f): the weights needed to achieve a maximum NSA.

Figures 2.3c and 2.3d indicate that relatively large weights for a SA component allow both maximum positive SA and negative SA to materialize with a small a, regardless of the shape of a beta RV. Figures 2.3e and 2.3f indicate that both positive SA and negative SA need large weights to converge on their maximums when a is small. However, as a increases, positive SA does not need a larger weight, whereas the weight for negative SA needs to increase in order to achieve its maximum. Symmetric and positively skewed beta RVs experience a rapid drop in their necessary weight levels for positive SA, whereas weights for negatively skewed beta RVs slowly decline, especially for the more skewed RVs. However, for negative SA, symmetric beta RVs display a similar weight growth pattern with negatively skewed RVs, whereas weights for positively skewed RVs dramatically increase first, and then level off.

The impacts of SA upon the four statistical moments also are evaluated for three different scenarios: symmetric ($c = 1$), positively skewed ($c = 5$), and negatively skewed ($c = 0.5$). The plot label numbers appearing in Figure 2.4 are the theoretical values of the statistical moments. Figures 2.4a and 2.4c indicate that the first statistical moment (i.e., the mean) of a beta RV is not affected by SA. However, Figures 2.4b and 2.4d indicate that SA inflates the variance, and this inflation gets increasingly severe as ρ increases. A comparison of Figures 2.4b and 2.4d suggests that the magnitude of inflation is more conspicuous for a larger a. Also, a pronounced difference is observed between positively and negatively autocorrelated beta RVs: a positively autocorrelated beta RV tends to deviate from its theoretical value. In addition, SA introduced in

symmetric beta RVs is more likely to inflate the variance. For example, Figure 2.4d implies that variance for a symmetric RV deviates more from its theoretical value (0.01) than for a negatively skewed RV.

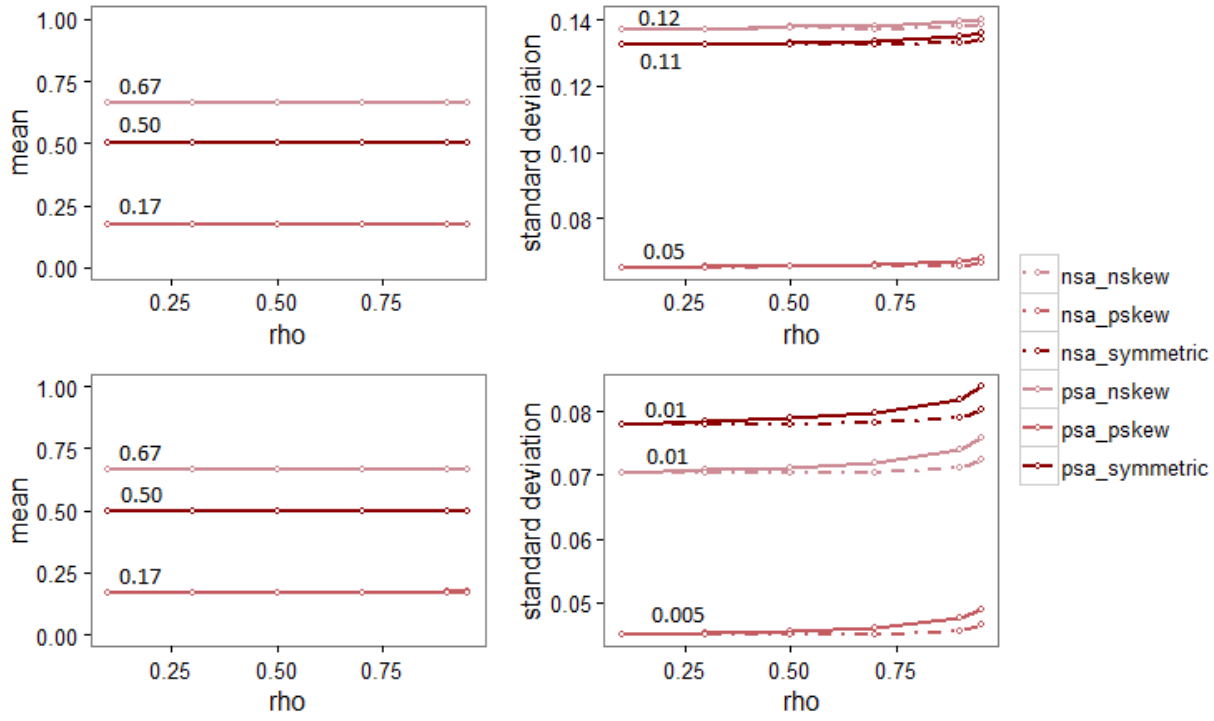


Figure 2.4. Summaries of statistical moments of simulated spatially autocorrelated beta RVs. Left: the mean of sample means [top (a): a = 10, bottom (c): a = 1,000]. Right: the mean of sample variances [top (b): a = 10, bottom (d): a = 1,000]

A different impact upon the skewness is observed for symmetric, positively skewed, and negatively skewed beta RVs (Figures 2.5a and 2.5c): skewness is greater than the theoretical value with positively skewed RVs, and less with symmetric and negatively skewed RVs when the constant a is relatively small. However, increasing a to 1,000 does not affect skewness for symmetric RVs. Rather, deviations from the theoretical values still remain for both positively and negatively skewed RVs, and the degree of deviation increases. Figures 2.5a and 2.5c indicate that although skewness is distorted by SA, it stays constant across different SA levels. Excess kurtosis also experiences some degree of alteration in the presence of SA: it remains unchanged for positively skewed RVs, but is deflated for symmetric RVs, and inflated for negatively skewed RVs when the constant a is relatively small (Figure 2.5b); small changes in this trend

occur across different ρ values. However, increasing a to 1,000 decreases excess kurtosis as the degree of SA increases (Figure 2.5d).

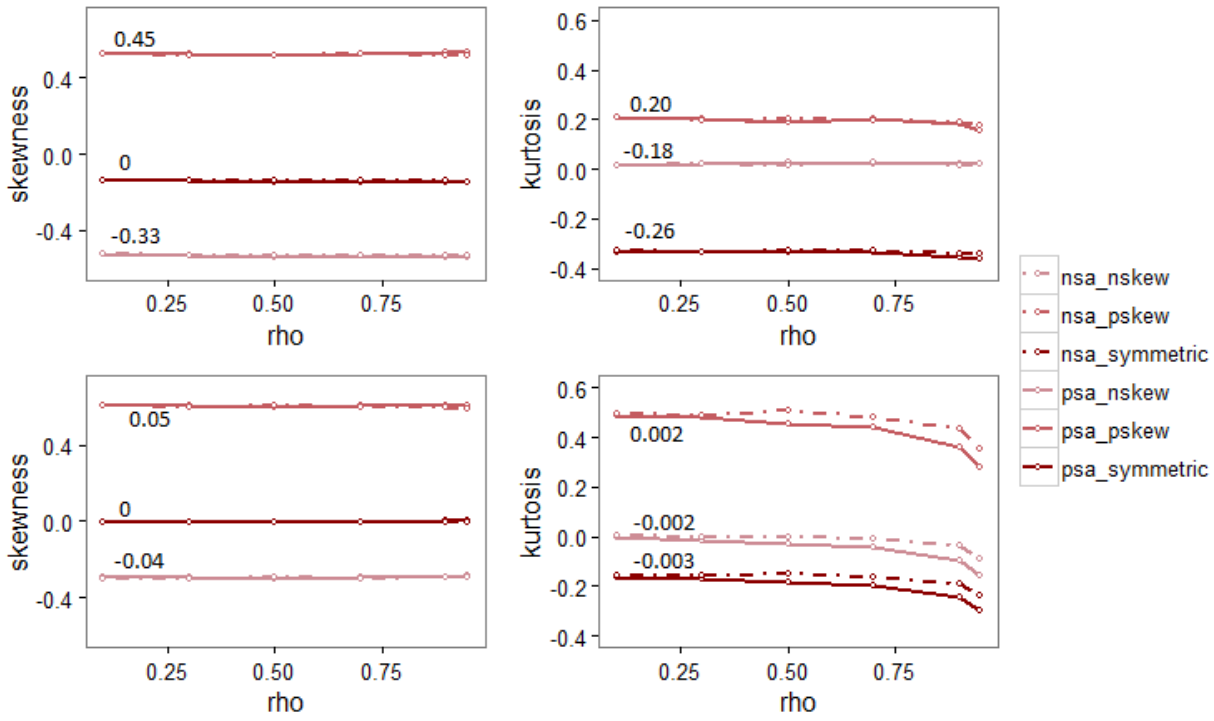


Figure 2.5. Summaries of statistical moments of simulated spatially autocorrelated beta RVs. Left: the mean of sample skewness [top (a): a = 10, bottom (c): a = 1,000]. Right: the mean of sample excess kurtosis [top (b): a = 10, bottom (d): a = 1,000]

2.4.2. Simulation results for multinomial RVs

Figures 2.6a and 2.6b portray the distribution of probabilities simulated with eq. (10) for each outcome across the specimen tessellation surface. The first two outcomes (π_{i1} and π_{i2}) have fundamentally identical distributions, whereas the third outcome (π_{i3}) displays a larger variance. However, the means of the probabilities approximately are the same, 0.33, for all three outcomes. Figures 2.6c and 2.6d illustrate distributions of probabilities generated with eq. (11). π_{i2} yields much larger probabilities. π_{i3} also has a wider range of probabilities. A comparison of Figures 2.6a–2.6d indicates that probability distributions for the asymmetric RVs have substantially more skewness than their counterparts for the symmetric RVs, with π_{i1} and π_{i2} negatively skewed, and π_{i3} positively skewed.

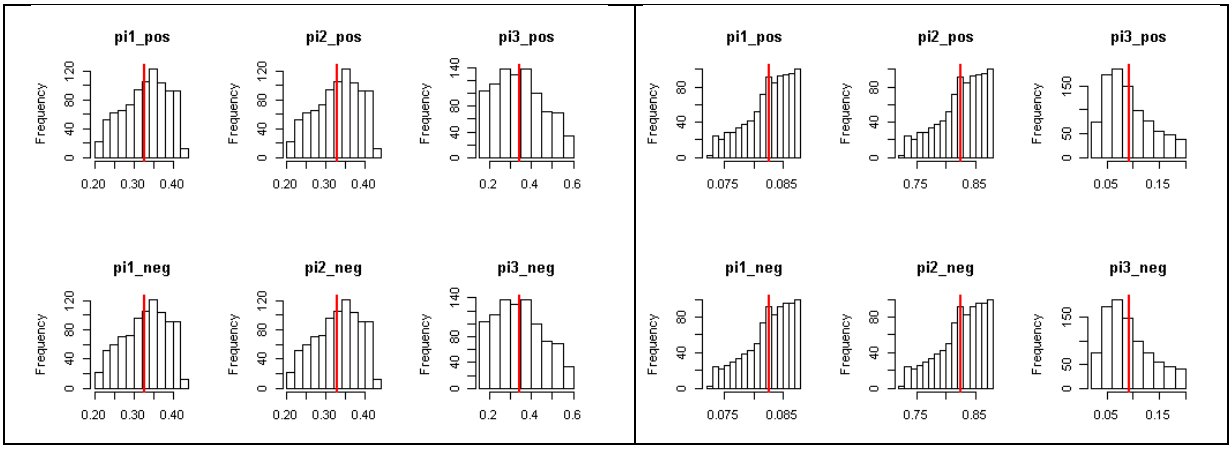


Figure 2.6. Probability distributions of multinomial outcomes. Top left (a): symmetric with PSA embedded. Bottom left (b): symmetric with NSA embedded. Top right (c): asymmetric with PSA embedded. Bottom right (d): asymmetric with NSA embedded.

Figure 2.7 portrays the simulation results for multinomial RVs where π_{i1} and π_{i2} have an identical distribution, hence the lines that represent π_{i1} and π_{i2} are almost identical. Figures 2.7a and 2.7b show that both positive SA and negative SA converge on their maximums as the total population size, N_{tr}, increases. When SA is weighted, both maximum positive SA and negative SA do not materialize for π_{i1} and π_{i2} when N_{tr} is small. But they slowly materialize as N_{tr} gets larger. In contrast, SA converges on its extreme for π_{i3} even for a small N_{tr} (Figures 2.7c and 2.7d). This outcome more than likely is caused by the specification of probabilities [eq. (10)]: π_{i1} and π_{i2} converge to 0, whereas π_{i3} converges to 1, when the SA term is weighted. Figures 2.7e and 2.7f display similar patterns with Figures 2.3e and 2.3f: that is, both positive SA and negative SA need large weights to achieve their potential extremes when N_{tr} is small. However, a large weight becomes unnecessary for positive SA, but remains important for negative SA, as N_{tr} increases.

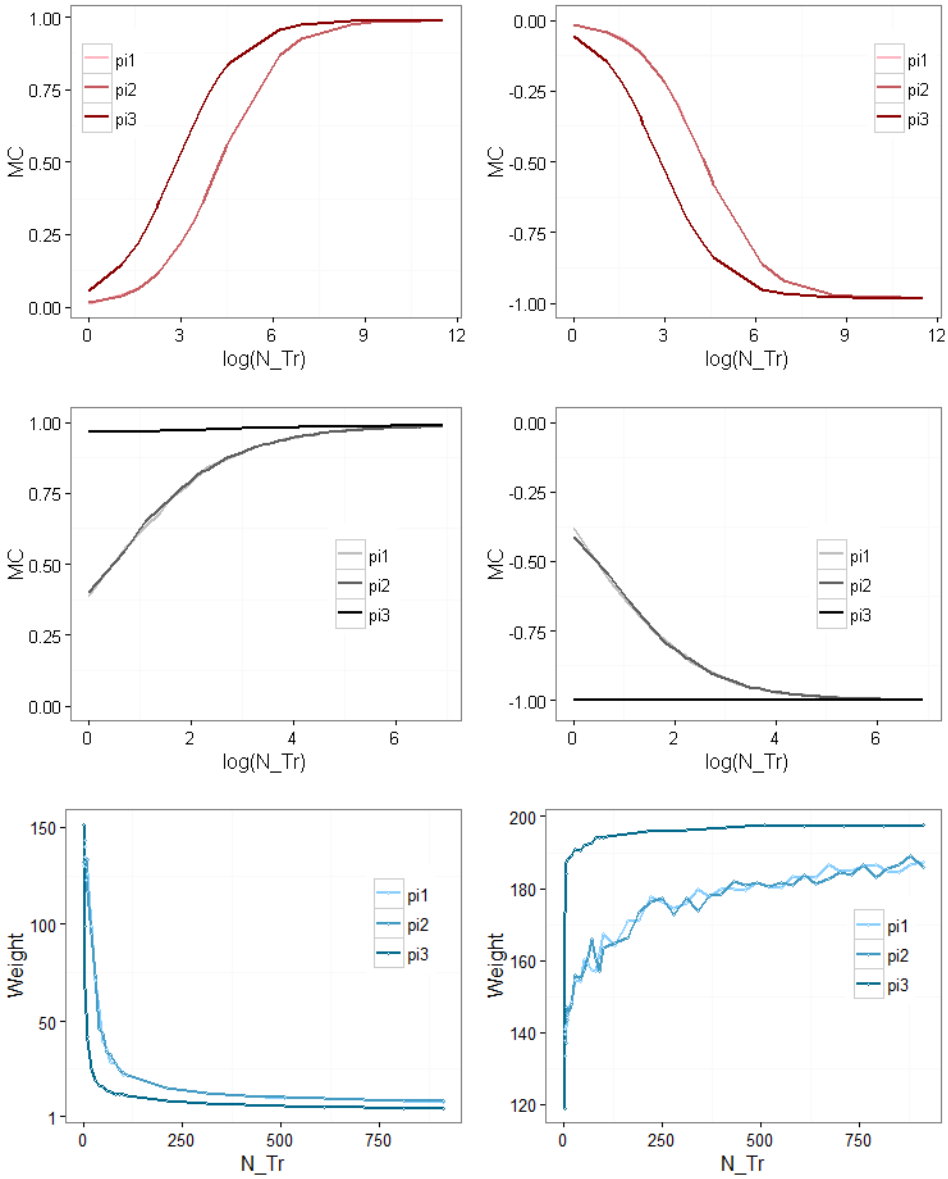


Figure 2.7. The range of SA for symmetric multinomial RVs. Top left (a): PSA. Top right (b): NSA. Middle left (c): maximum PSA. Middle right (d): maximum NSA. Bottom left (e): weights needed to achieve extreme PSA. Bottom right (f): weights needed to achieve extreme NSA.

Figures 2.8a and 2.8b indicate that π_{i1} experiences a delayed convergence for both positive SA and negative SA. They also show that π_{i2} displays a convergence pattern similar to that for π_{i3}. However, a conspicuous gap occurs between the maximum negative SA achieved and the theoretical maximum, and this discrepancy most likely occurs because of the increased skewness of the asymmetric multinomial RVs (Figures 2.6c and 2.6d). Even with weights on a SA

component, the most extreme SA is unable to materialize for π_{i1} and π_{i2}, although it converges as N_{tr} gets larger. π_{i1} starts with an even lower absolute SA level compared with that for Figures 2.7c-2.7d. However, SA converges on its extremes for π_{i3} across a wide range of different N_{tr} values. Figures 2.8e and 2.8f suggest an unstable pattern of the needed weights for π_{i1}, with more instability observed with small N_{tr}. In contrast, π_{i2} and π_{i3} display stable patterns, which indicates that the needed weight decreases for positive SA but increases for negative SA as N_{tr} increases.

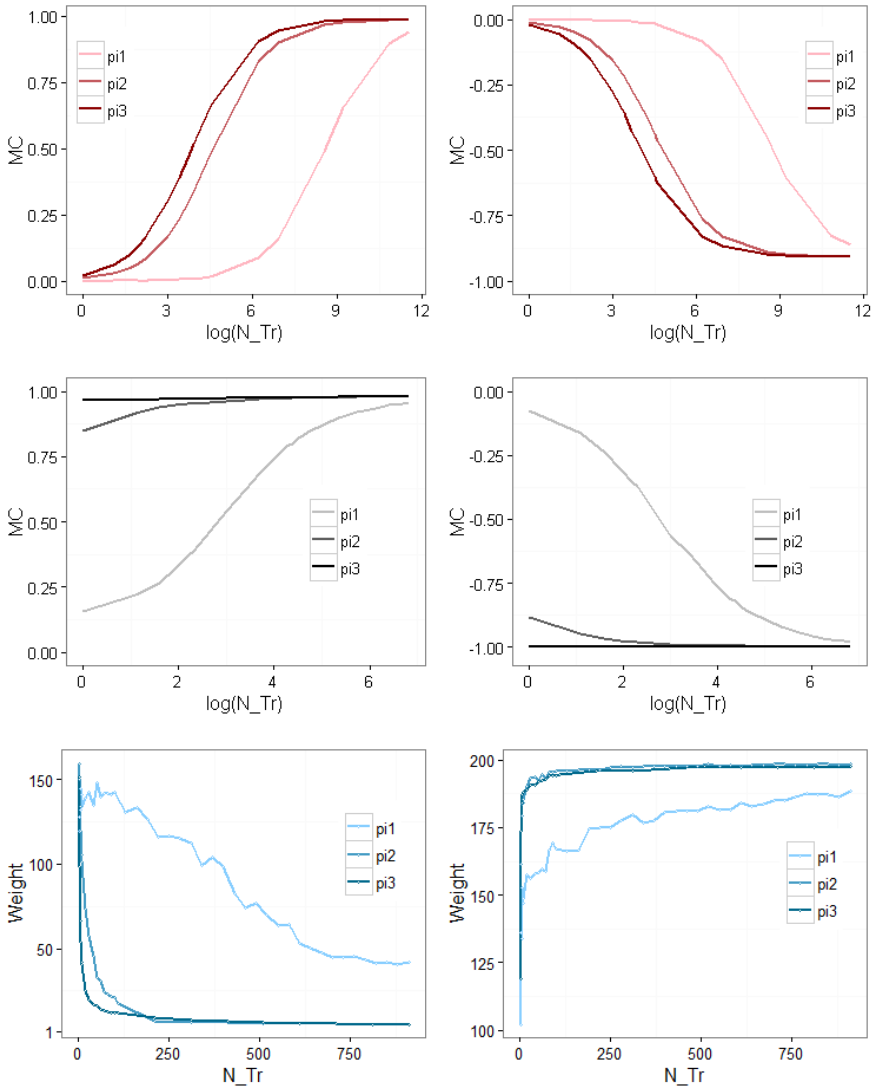


Figure 2.8. The range of SA for asymmetric multinomial RVs. Top left (a): PSA. Top right (b): NSA. Middle left (c): maximum PSA. Middle right (d): maximum NSA. Bottom left (e): weights needed to achieve extreme PSA. Bottom right (f): weights needed to achieve extreme NSA.

The effects of SA on the first four statistical moments of symmetric/asymmetric multinomial RVs also are assessed for the three outcomes, and the results are reported in Figure 2.9. The plot label numbers appearing in Figure 2.9 are the theoretical values of the statistical moments. Similar to Figure 2.7, the two lines representing π_{i1} and π_{i2} are nearly identical in Figures 2.9a and 2.9b, and Figures 2.10a and 2.10b, because of their undistinguishable specifications. Figures 2.9a and 2.9c indicate that the mean of the simulated RVs for each outcome is consistent with its theoretical value, which means that SA does not impact the first statistical moment. Figures 2.9b and 2.9d suggest that the variance has substantial inflation, with more deviation observed for π_{i3}. In addition, positive SA contributes more than negative SA to this inflation, and the difference between positive SA and negative SA becomes more noticeable as ρ increases. Moreover, variance inflation tends to increase as ρ increases for both positive SA and negative SA. A comparison of Figures 2.9b and 2.9d suggests that variance is likely to inflate substantially more for a symmetric than for a skewed RV. For example, the standard deviation inflates from 15 to 70 for π_{i3} for a selected symmetric multinomial RV, whereas it inflates from 9 to 26 for a selected asymmetric multinomial RV.

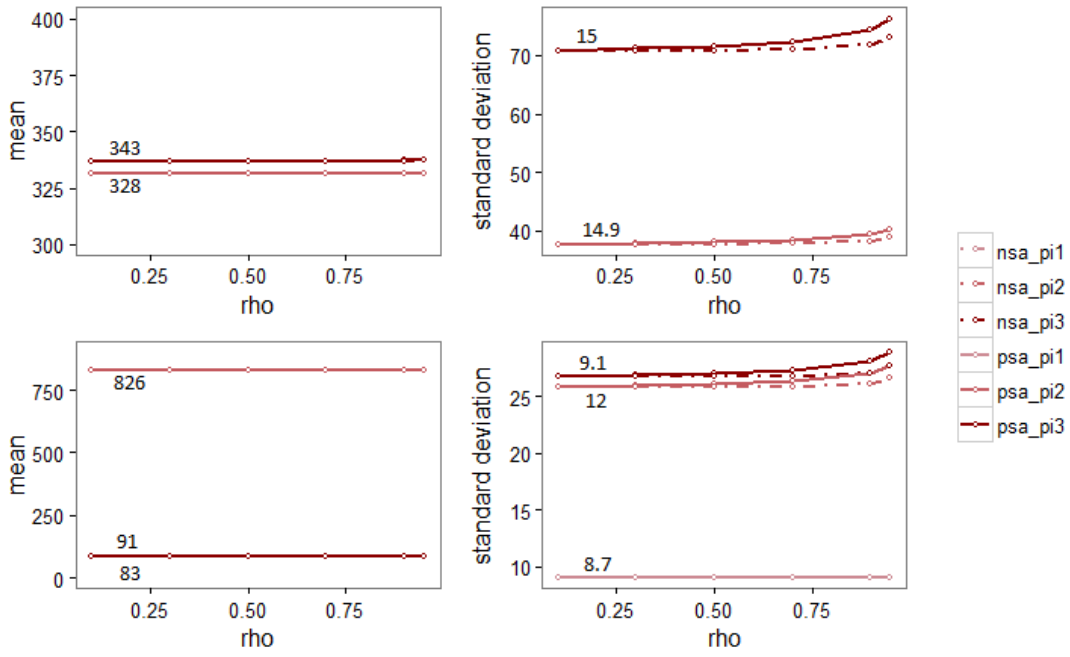


Figure 2.9. Summaries of statistical moments for simulated spatially autocorrelated multinomial RVs. Left: the mean of sample means [top (a): symmetric, bottom (c): asymmetric] Right: the mean of sample variances [top (b): symmetric, bottom (d): asymmetric]

Figures 2.10a and 2.10c indicate different impacts of SA on skewness. Specifically, for symmetric multinomial RVs, skewness is less than its theoretical counterpart for π_{i1} and π_{i2} (negatively skewed), whereas it is greater for π_{i3} (positively skewed). However, for asymmetric multinomial RVs, skewness remains unchanged for π_{i1} , decreases for π_{i2} (negatively skewed), and increases for π_{i3} (positively skewed). Figures 2.10b and 2.10d suggest that SA distorts excess kurtosis of a multinomial RV. For example, it slightly decreases for all outcomes of the simulated symmetric multinomial RVs, and it increasingly deviates from its theoretical counterpart as the degree of SA increases.

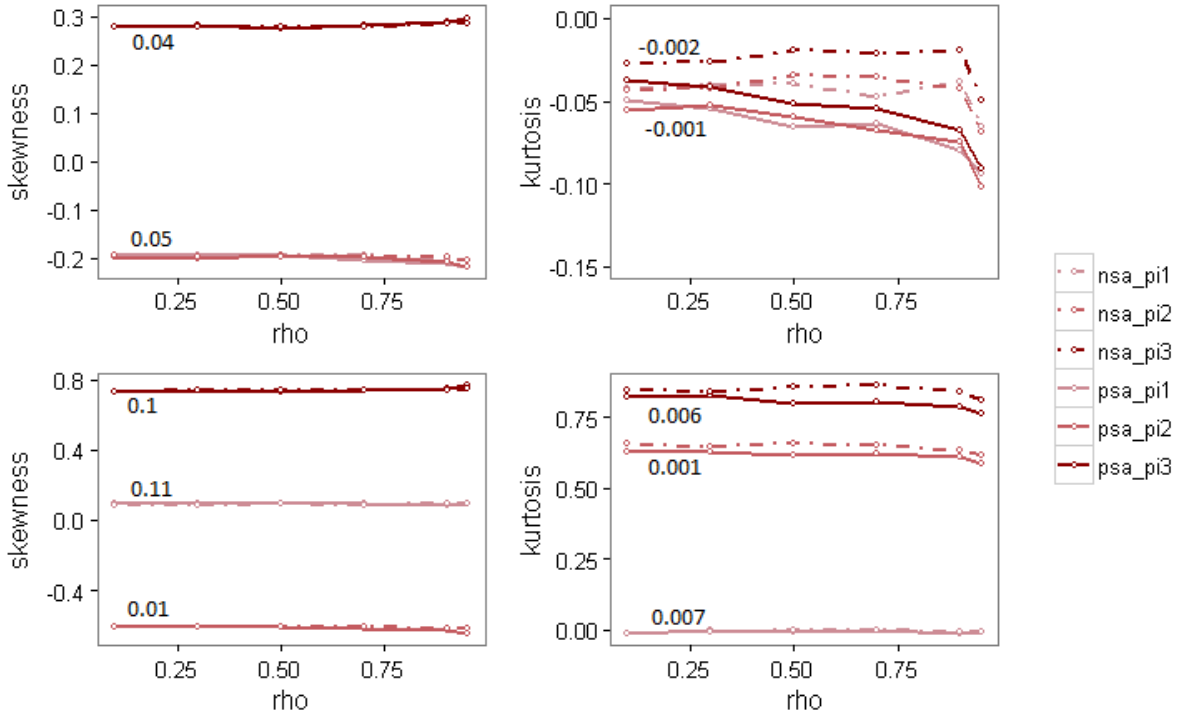


Figure 2.10. Summaries of statistical moments for simulated spatially autocorrelated multinomial RVs. Left: the mean of sample skewness [top (a): symmetric, bottom (c): asymmetric] Right: the mean of sample excess kurtosis [top (b): symmetric, bottom (d): asymmetric]

2.5 Summary and Conclusions

The beta and multinomial distributions are of increasing importance in GIScience, with the multinomial distribution already having play a prominent role for years. However, work has not been done to investigate their distributional properties in the presence of SA. This research

examines the impact of SA upon histograms of beta and multinomial RVs, with regard to both PSA and NSA. The major findings of the research summarized in this paper are the following.

First, PSA and NSA display similar impacts when a RV closely approximates a normal distribution (e.g., symmetric beta and multinomial RVs), which is consistent with the findings by Chun and Griffith (2018) for the binomial RV. However, PSA and NSA impacts behave differently when a RV is skewed (positively or negatively). NSA fails to converge on its maximum, and the gap between the maximum it achieves and the theoretical maximum becomes more conspicuous as skewness increases. In contrast, PSA always converges and is not impacted by skewness. A difference in performance between PSA and NSA also is reported by Chun and Griffith (2018) for Poisson RVs. Second, simulation results imply that a RV mean is unaffected by SA. However, SA inflates variance, with this inflation becoming more pronounced as the SA level increases. Meanwhile, PSA generally creates more inflation than NSA. These results corroborate findings in the literature. Additionally, this simulation output reveals more noticeable variance inflation for symmetric than for skewed beta/multinomial RVs. Third, SA distorts skewness and kurtosis (e.g., Griffith 2011; Chun and Griffith 2018). The simulation results in this paper indicate that skewness is impacted differently by SA, depending on the properties of a RV. Distortion becomes more severe as the degree of skewness increases. However, the degree of skewness remains constant across different SA levels. The fourth moment, excess kurtosis, also is altered by SA, decreasing as the SA level increases.

The research summarized here can be further extended in future work. First, because this research is based on simulated data, an analysis of empirical georeferenced data involving the beta/multinomial distributions should be insightful. Second, a key discovery of this research is different behaviors attributable to PSA and NSA when a spatially autocorrelated RV is skewed, and, furthermore, skewness has different impacts on the second and third statistical moments. However, in-depth studies still need to be undertaken to further explore and understand this outcome. Third, as discussed in the literature, NSA commonly is found in a mixture with PSA in georeferenced data. Simulation experiments need to be extended to investigate the impacts of SA mixtures on a histogram, and to contrast the results with PSA and NSA only. Fourth, simulation

experiments designed for this research were based on a 30-by-30 square tessellation surface. Future research should examine whether or not results are consistent with different sizes and configurations of surface partitioning (e.g., a hexagonal/irregular tessellation). Moreover, because research results may be sensitive to the specification of a spatial weights matrix, SA impacts of beta and multinomial RVs based upon different spatial weight matrices (e.g., one based on a queen contiguity) merit evaluation.

2.6 Appendix A2. Map patterns of RVs

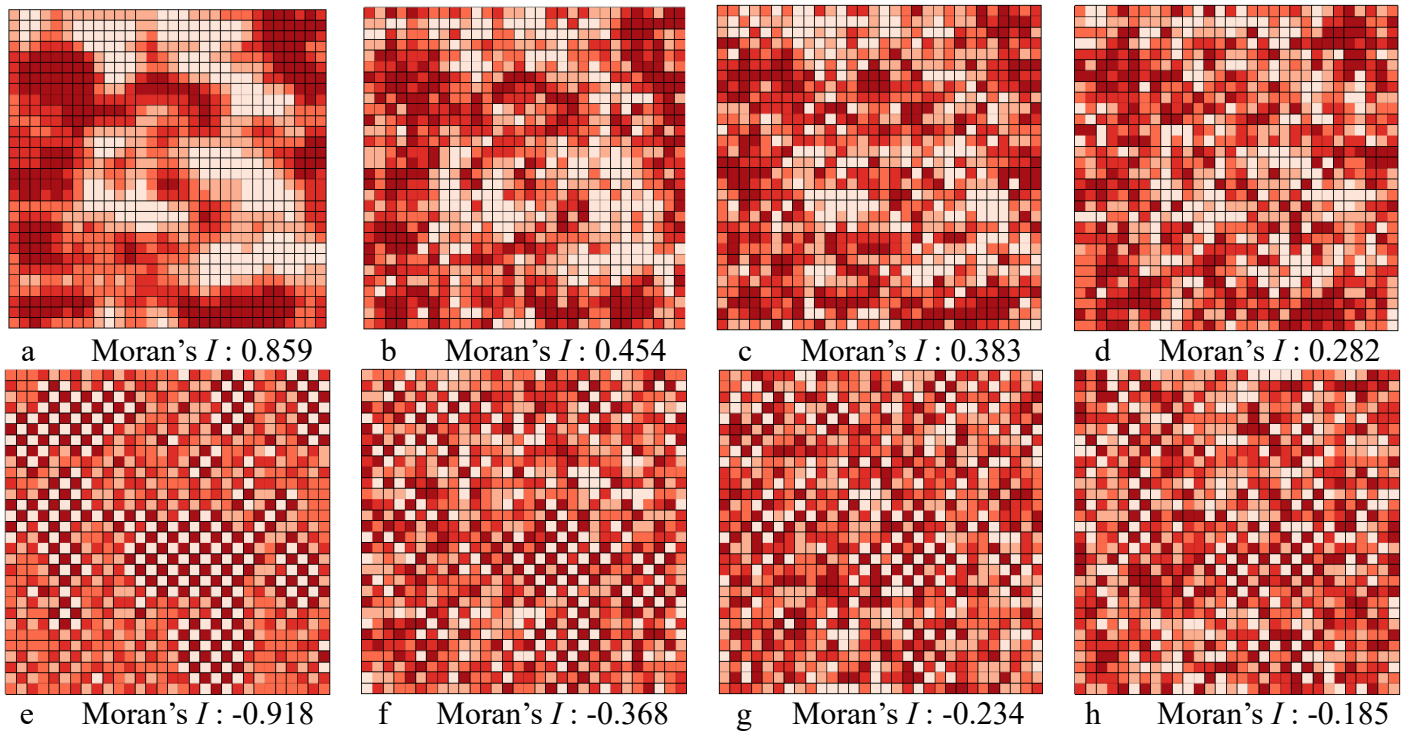
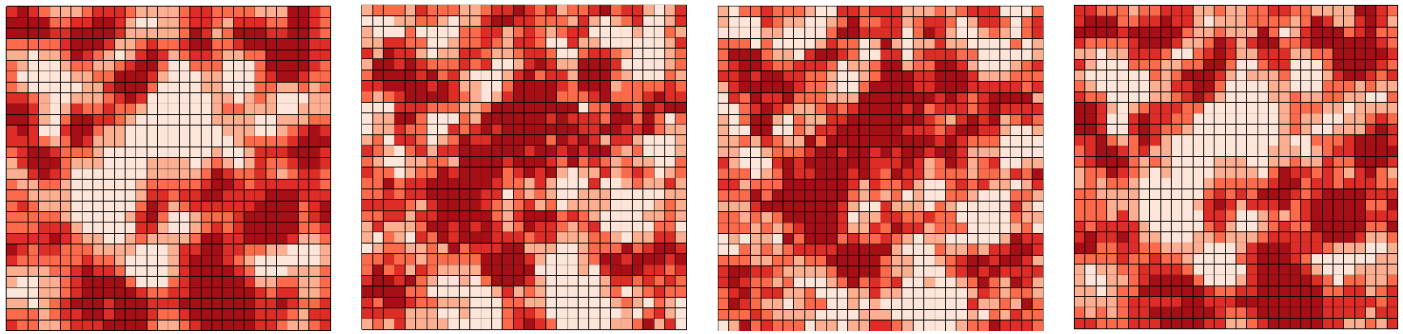
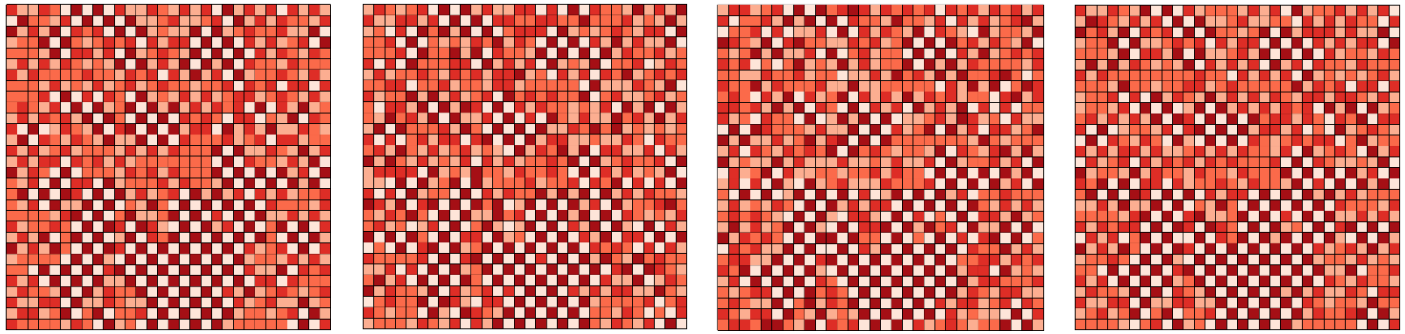


Figure A2.1.Map patterns of beta RVs. [a: PSA with ρ =0.95. b-d: beta RVs with PSA (b: positively skewed, c: symmetric, d: negatively skewed). e: NSA with ρ =−0.95. f-h: beta RVs with NSA (f: positively skewed, g: symmetric, h: negatively skewed)]

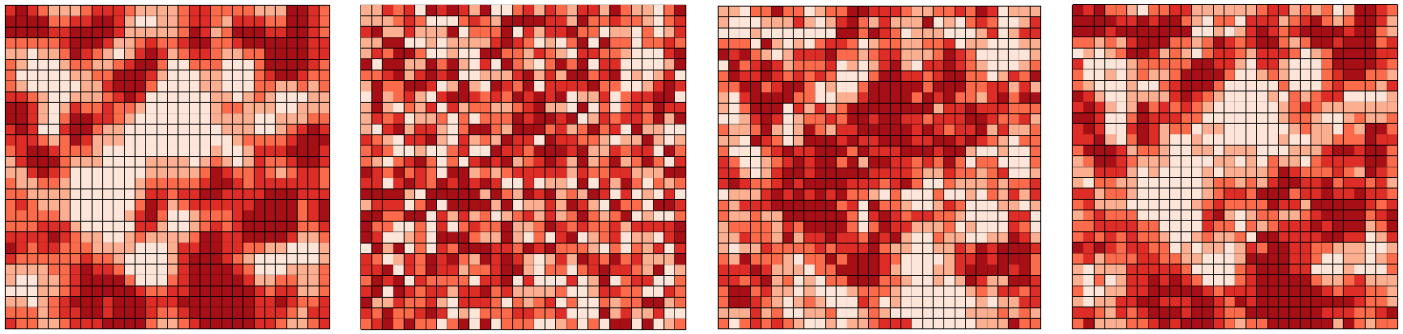


a Moran's I : 0.861 b Moran's I : 0.727 c Moran's I : 0.728 d Moran's I : 0.822

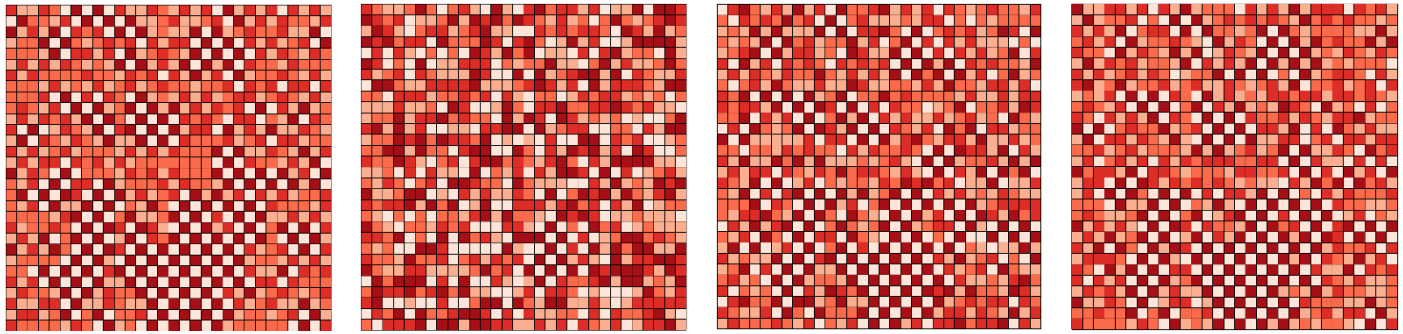


e Moran's I : -0.862 f Moran's I : -0.748 g Moran's I : -0.747 h Moran's I : -0.825

Figure A2.2. Map patterns of symmetric multinomial RVs. [a: PSA with ρ =0.95. b-d: multinomial RVs with PSA (b: π_1, c: π_2, d: π_3). e: NSA with ρ =−0.95. f-h: multinomial RVs with NSA (f: π_1, g: π_2, h: π_3)]



a Moran's I : 0.867 b Moran's I : 0.050 c Moran's I : 0.692 d Moran's I : 0.770



e Moran's I : -0.888 f Moran's I : -0.061 g Moran's I : -0.652 h Moran's I : -0.747

Figure A2.3. Map patterns of asymmetric multinomial RVs. [a: PSA with ρ =0.95. b-d: multinomial RVs with PSA (b: π_1, c: π_2, d: π_3). e: NSA with ρ =−0.95. f-h: multinomial RVs with NSA (f: π_1, g: π_2, h: π_3)]

CHAPTER 3

UNCOVERING A POSITIVE AND NEGATIVE SPATIAL AUTOCORRELATION MIXTURE PATTERN: A SPATIAL ANALYSIS OF BREAST CANCER INCIDENCES

Authors – Lan Hu, Yongwan Chun, Daniel A. Griffith

Geospatial Information Sciences Program, GR 31

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

Reproduced with permission from SpringerNature

Hu, L., Chun, Y., & Griffith, D. A. (2020). Uncovering a positive and negative spatial autocorrelation mixture pattern: a spatial analysis of breast cancer incidences in Broward County, Florida, 2000–2010. *Journal of Geographical Systems*, 1-18.

ABSTRACT

Spatial cancer data analyses frequently utilize regression techniques to investigate associations between cancer incidences and potential covariates. Model specification, a process of formulating an appropriate model, is a well-recognized task in the literature. It involves a distributional assumption for a dependent variable, a proper set of predictor variables (i.e., covariates), and a functional form of the model, among other things. For example, one of the assumptions of a conventional statistical model is independence of model residuals, an assumption that can be easily violated when spatial autocorrelation is present in observations. A failure to account for spatial structure can result in unreliable estimation results. Furthermore, the difficulty of describing georeferenced data may increase with the presence of a positive and negative spatial autocorrelation mixture, because most current model specifications cannot successfully explain a mixture of spatial processes with a single spatial autocorrelation parameter. Especially, properly accounting for a spatial autocorrelation mixture is challenging. This paper empirically investigates and uncovers a possible spatial autocorrelation mixture pattern in breast cancer incidences in Broward County, Florida during 2000 to 2010, employing different model specifications. The analysis results show that Moran eigenvector spatial filtering provides a flexible method to examine such a mixture.

3.1 Introduction

Spatial data analysis has been widely used to describe georeferenced cancer data. It often involves understanding the spatial patterns of cancer incidences, and further examining factors that are potentially associated with cancer occurrences. The development of geographic information systems (GISs) and spatial statistical analytics, such as exploratory spatial data analysis tools, have fostered investigations of spatial patterns of cancer data. For example, a collection of studies in the spatial cancer data analysis literature focuses on detecting spatial clusters of cancer incidences in space (e.g., Timander and McLafferty 1998, Meliker et al. 2009). Furthermore, regression models are utilized to examine potential covariates that can have significant associations with cancer incidences, including socio-economic status and demographic factors (e.g., MacKinnon et al. 2007, Dai 2010, Wang et al. 2012).

However, regression analysis results can be unreliable when spatial autocorrelation (SA) is present and it is not appropriately accounted for. For example, Dai (2010) uses linear regression and spatial lag models to evaluate the role of black residential segregation and spatial access to health care, with his results indicating that the spatial model outperforms the linear regression model by successfully addressing SA in the cancer data. In a spatial analysis, Moran's *I* or Geary's *c* are commonly used to test for SA in observed values and/or model residuals. However, a global measure of the underlying SA component may not reveal a complex spatial structure. For example, Hu et al. (2018) argue that the weak positive and (near-) zero SA that exist in a lung cancer dataset essentially are mixtures of positive and negative SA. The presence of a simultaneous SA mixture in a variable may lead to a failure of rejecting the null hypothesis for the SA parameter when positive SA (PSA) and negative SA (NSA) components cancel each other (Griffith 2006). When a SA mixture exists in data, a model specification should be able to accommodate both PSA and NSA properly to produce unbiased estimates. However, popular spatial regression model specifications (e.g., spatial autoregressive models) presently are not formulated to successfully account for a latent SA mixture in the geographic distribution of cancer incidences.

The purpose of this paper is to empirically investigate breast cancer incidences for females in Broward County, Florida (FL) with a focus on model specification issues. First, spatial patterns of age-adjusted breast cancer incidence rates are examined to determine whether or not a spatial pattern, including a mixture of PSA and NSA, is present. Next, potential risk factors for age-adjusted breast cancer incidence rates are examined using Poisson and negative binomial (NB) regression. These models are specified with expected counts of breast incidences as an offset variable, which is calculated with age-specific breast cancer rates in the United States (US). Furthermore, Poisson and NB model specifications are extended with the Moran eigenvector spatial filtering (MESF) methodology to account for SA, the estimated results are compared with those estimated with Besag-York- Mollié (BYM) model specifications. Importantly, this paper addresses a mixture of PSA and NSA in georeferenced data, and employs MESF model specifications to account for the PSA-NSA mixture feature in Poisson and NB regression.

3.2 Literature review

Breast cancer is one of the most commonly diagnosed cancers among women, and also is the leading cause of cancer deaths worldwide (Parkin et al. 2001). Breast cancer has drawn tremendous research attention in epidemiology and the social sciences, geography being among them. Many studies investigate the spatial patterns of breast cancer incidences (e.g., Vieira et al. 2008, Meliker et al. 2009), and examine associated risk factors (e.g., McPherson et al. 2000, Wang et al. 2012). Cancer rates are popularly utilized in these studies, and age-adjusted cancer rates generally are preferred over crude rates because they reflect different cancer risk levels among different population age cohorts (Anderson and Rosenber 1998, Ahmad et al. 2001). Commonly, age-adjusted rates are obtained by adjusting observed age-specific incidence rates based on the age structure of a reference population (Bray 2002). Although these cancer rates often are modeled with linear regression, a comparable model specification can be formulated with Poisson or NB regression for cancer counts using an offset variable (e.g., Sheehan 2004). This specification has the advantage that heterogeneous population sizes can be incorporated in the posited model (McCullagh and Nelder 1989).

Studies show that breast cancer incidences vary substantially over space; however, PSA in cancer incidences frequently is detected. For example, Zhou et al. (2015) identify significant clusters of breast cancer incidences in south-central Shenzhen in China, and Fukuda et al. (2005) report a spatially clustered pattern of breast cancer in Japan. Dai (2010) discusses PSA when modeling late-stage breast cancer in the Detroit metropolitan area, and Tian et al. (2011) show the presence of PSA in female breast cancer mortality rates in Texas. Some studies do not report significant SA. For example, Timander and McLafferty (1998) observe that breast cancer rates are fairly evenly distributed in West Islip, New York, and Muir et al. (2004) do not find spatial clusters among breast cancer incidences in Lincolnshire and Leicestershire in England.

Although PSA in breast cancer incidences frequently is identified in the literature, NSA also is observed in some geographically distributed phenomena. One of the few situations exhibiting NSA is that of spatial competition (Haining 1984). Griffith (2006) argues that a mixture of PSA and NSA may mask significant SA in a global SA test because they can cancel each other. In addition, the possible presence of NSA suggested by local area statistics (e.g., Le Gallo and Ertur 2003, Baumont et al. 2004, Odoi et al. 2003) appears in a mixture of both PSA and NSA latent in geographic distributions with a dominating global PSA.

Studies reveal that the spatial variability of breast cancer incidences can be explained by established risk factors (e.g., Gumpertz et al. 2006, Meliker et al. 2009). For example, Gumpertz et al. (2006) discuss that although the geographic pattern of breast cancer incidence rates differs within Los Angeles County, its geographic variation can be well explained with a set of biological and sociodemographic covariates in a generalized linear mixed model. Much research utilizes Gaussian linear regression (e.g., Dai 2010, Wang et al. 2012) or generalized linear regression models (e.g., Robert et al. 2004, Yang et al. 2011) to examine risk factors for breast cancer. To date, several significant contributing influences have been uncovered, including genetic factors, [e.g., family history] (e.g., McPherson et al. 2000, Yang et al. 2011), reproductive factors, [e.g., nulliparity] (e.g., Kelsey et al. 1993, Yang et al. 2011), demographic factors, [e.g., race and age] (e.g., Dai 2010, McPherson et al. 2000), socio-economic factors, [e.g., poverty and education] (e.g., MacKinnon et al. 2007, Hussain et al. 2008), and geographic

factors [e.g., primary care access and urban/rural disparities] (e.g., Wang et al. 2012, MacKinnon et al. 2007).

Although well-designed aspatial model specifications with associated explanatory variables can account for spatial variability in breast cancer incidences, most empirical studies suffer from ignoring SA in their regression model specifications. SA invalidates the independence assumption in conventional statistics; hence, it needs to be appropriately addressed in a spatial analysis (Griffith 1987). Spatial specifications that are commonly utilized in cancer research include spatial autoregressive models (e.g., Antunes et al. 2001, Keitt et al. 2002), Bayesian spatial models (e.g., Lawson 2013), and MESF (e.g., Tiefelsdorf 2007, Jacob et al. 2011). Bayesian spatial models have been popularly used in disease mapping (Lawson 2013). For example, Torabi and Rosychuk (2012) use the well-known intrinsic conditionally autoregressive (ICAR) approach to accommodate spatial random effects in cancer incidence ratios. An ICAR prior also is adopted by Kazembe and Namangale (2007) to capture the spatial structure in childhood co-morbidity. Lee (2011) compares four different CAR models (e.g., the ICAR) and utilizes them for mapping cancer incidence rates in Greater Glasgow, Scotland.

The BYM model, an extension of the ICAR model specification that includes an additional random effect component for non-spatial heterogeneity, furnishes a useful approach to model areal count data of rare diseases (Gerber and Furrer, 2015). For example, López-Abente (2014) utilizes a BYM model to describe stomach cancer mortality rates in Spain, and reports that the geographical pattern is maintained across the study period. Although the random effect term in the BYM model can capture SA that is unexplained due to omitted covariates, it is considered not to be sufficiently flexible to account for the complex localized structure that possibly exists in residual SA because its random effects term exhibits a single global level of spatial smoothness determined by geographical adjacency, similar to the ICAR model (Lee et al., 2014; Hodges and Reich 2010). That is, the BYM and ICAR models are limited in their respective ability to accommodate a SA mixture pattern in georeferenced data. MESF, however, increasingly has been utilized with linear and generalized linear regression models to account for both PSA and NSA components simultaneously. For example, Jacob et al. (2011) posit an MESF

model specification to detect and adjust for hidden PSA and NSA components in their georeferenced data. Hu et al. (2018) use a MESF model to uncover a SA mixture pattern in lung cancer data.

3.3 Methodology

This research utilizes Poisson and NB models to describe cancer counts, which then are further extended to MESF and BYM specifications to uncover the underlying spatial pattern. This section briefly describes the MESF and BYM method specifications.

3.3.1. MESF model specification

This paper utilizes MESF methodology to account for SA in Poisson and NB regression. MESF utilizes a set of eigenvectors that are extracted from a transformed *n*-by-*n* spatial weights matrix, **C**, appearing in the numerator of the Moran Coefficient (e.g., Moran's *I*), which can be expressed as:

$$\mathbf{MCM} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n),$$

where **I** is an *n*-by-*n* identity matrix, **1** is a *n*-by-*1* vector of ones, *n* is the number of areal units, and superscript *T* is the matrix transpose operator. These *n* eigenvectors are mutually uncorrelated and orthogonal, and represent underlying components of SA. MESF introduces a subset of these eigenvectors as independent variables in a regression model to capture unexplained SA with its model specification (Griffith 2003, Griffith et al. 2019). This subset can be identified from a candidate eigenvector set, which is considerably smaller than *n*, with a stepwise regression procedure (Chun et al. 2016). In this paper, the stepwise procedure is conducted with the Akaike information criterion (AIC).

MESF can be flexibly specified to account for positive, negative, or a mixture of both types of SA. The *n* eigenvectors represent distinct underlying spatial patterns, and their corresponding eigenvalues represent their levels of SA when they are visualized with the spatial units used to generate the *n*-by-*n* spatial weight matrix **C**. Tiefelsdorf and Boots (1995) show that these

eigenvalues are equivalent to Moran's *I* values for their respective map patterns. Hence, eigenvectors with positive eigenvalues portray PSA, and eigenvectors with negative eigenvalues portray NSA (hereafter they are called PSA eigenvectors and NSA eigenvectors, respectively). Whereas MESF often uses a subset of only PSA eigenvectors to account for PSA, a mixture of PSA and NSA can be explained with both positive and NSA eigenvectors. Selected PSA and NSA eigenvectors capture, respectively, observed PSA and NSA described by a model specification. These three different models are specified with different candidate eigenvector sets in the stepwise procedure: that is, only PSA eigenvectors, only NSA eigenvectors, and a combined set of PSA and NSA eigenvectors.

A MESF model specification for a Poisson random variable (Griffith 2002) can be specified as:

$$E[\mathbf{Y}] = g^{-1}(\mathbf{X\beta} + \mathbf{E\gamma}) + \mathbf{O},$$

where $g(\cdot)$ is a link function that is the natural logarithm in most Poisson cases, $E[\cdot]$ denotes the expectation operator, **Y** is a *n*-by-*1* vector of the response variable assumed to follow a Poisson distribution, **O** is a *n*-by-*1* vector of offset values**. X** is a *n*-by-*k* matrix that contains covariates and *k* is number of covariate, **E** is a *n*-by-*p* matrix eigenvectors and *p* represents the number of selected eigenvectors , and **β** and **γ** are *1*-by-*k* and *1*-by-*p* vectors to be estimated respectively. Empirical studies show that this Poisson MESF specification successfully accounts for SA (e.g., Chun 2008, Griffith 2009). Similarly, a MESF specification can be employed with NB regression, which often is preferred when overdispersion occurs in Poisson regression (e.g., Patuelli et al. 2011, Chun 2014). However, because SA is a well-known source of overdispersion in a Poisson random variable (Haining et al. 2009, Griffith 2011), overdispersion becomes negligible in a Poisson MESF model when SA is successfully accounted for. In such a case, the Poisson MESF model should be equivalent to a NB MESF model and, hence, a NB specification is unnecessary.

3.3.2. The BYM model specification

The BYM model (Besag et al., 1991) includes an ICAR component to address SA and a non-spatial random effect component to capture uncorrelated heterogeneity. A BYM model for a Poisson variable (Riebler et al. 2016) can be specified as:

$$Y_i|O_i, \lambda_i \sim \text{Poisson}(O_i\lambda_i)$$
$$\log(\lambda_i) = \boldsymbol{X}_i\boldsymbol{\beta} + U_i + V_i$$

where O_i denotes the offset value for spatial unit i, λ_i denotes the mean value of a count for the i^{th} unit. $\boldsymbol{X}_i$ is a row vector of covariates for the i^{th} unit, $\boldsymbol{\beta}$ denotes a vector of regression coefficients, and U_i denotes a spatially correlated component for spatial unit i. This spatially correlated random effects term is normally distributed conditioned on its adjacent areas (Neyens et al., 2012). V_i denotes a non-spatial random effect component for the i^{th} unit. This non-spatial component is normally distributed and independent between areas (Neyens et al., 2012; Lee 2011); it is commonly used to account for overdispersion in count data modeling. This research utilizes an integrated nested Laplace (INLA) method to estimate the BYM model. INLA can produce comparable estimates in a short amount time compared to the Markov chain Monte Carlo (MCMC) approach (e.g., Gibbs sampler) (Rue et al., 2009; Gomez-Rubio et al., 2014). In this paper, the BYM model is estimated with the INLA package in R.

3.4 Results

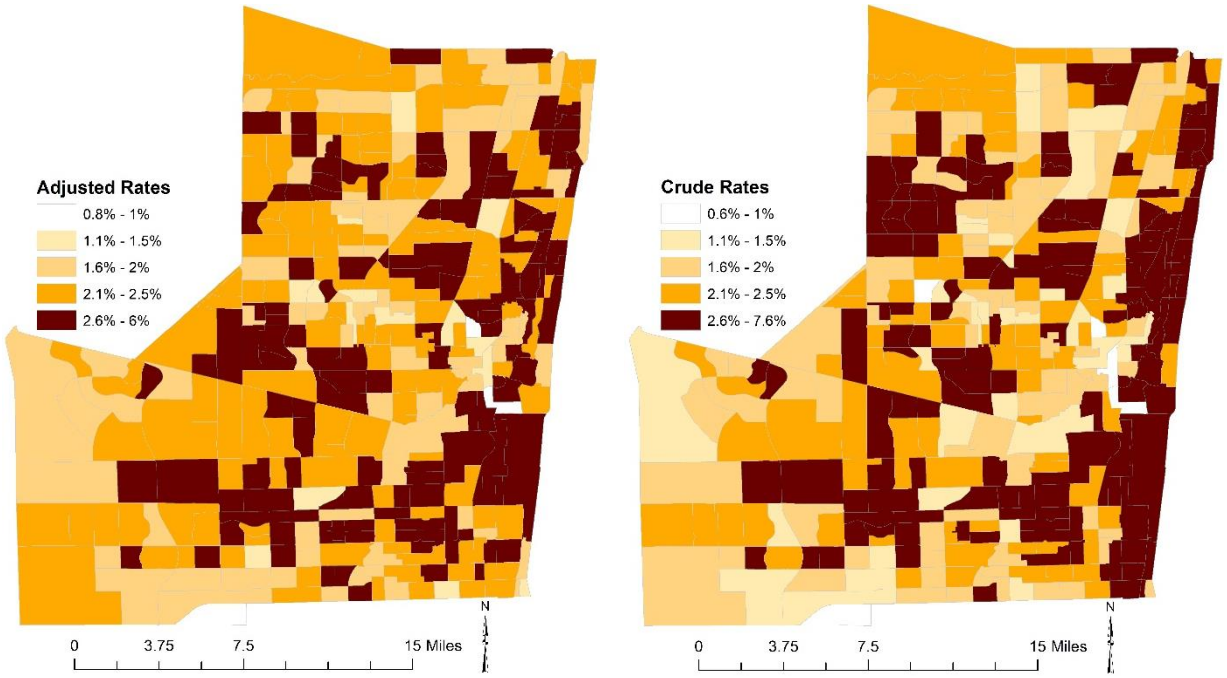
This section reports spatial patterns of age-adjusted breast cancer rates in Broward County, Florida, and then presents Poisson and NB regression results for the census tract resolutions. In addition, MESF model results including a PSA component only, and a simultaneous mixture of PSA and NSA components, are compared with those from standard Poisson and NB regression.

3.4.1. Breast cancer data and spatial patterns

The breast cancer data were retrieved from the Florida Cancer Registry. This dataset contains 18,905 cases of breast cancer in Broward County, FL for the 11-year period from 2000 to 2010.

These data underwent a data cleaning process by removing cases that are either unsuccessfully geocoded incidences (936), or duplicate registries (946), or male and unknown gender cases (249). Also, cases that are geocoded in census blocks with zero population in both the 2000 and 2010 US decennial censuses (234) were removed. A total of 16,541 cases were used in the data analysis after this data cleaning process. Next, age-adjusted breast cancer incidence rates were calculated. The nationwide 2010 US Census population counts by age cohort were collected from the US Census Bureau (https://factfinder.census.gov) for adjustment purposes. Then these rates were transformed to improve their normality using a Box-Cox type power transformation (Yeo and Johnson 2000).

Figure 3.1a depicts the spatial distribution of the age-adjusted breast cancer incidence rates. This map exhibits a pattern of PSA, confirmed by Moran's *I* test (z-score = 6.481). For comparison purposes, crude rates are portrayed in Figure 3.1b. This map pattern shows stronger PSA visually as well as with the Moran's *I* test (z-score = 7.62). This example empirically suggests that different rate calculations can lead to different spatial patterns, possibly due to the loss of some local PSA patterns: clusters of high rates in the upper middle area of the map and the coastal area are observed in Figure 3.1b, but not observed or weakened in Figure 3.1a Nevertheless, a visual inspection of Figure 3.1a and 3.1b suggests similar map patterns (e.g., a cluster of high rates at the southeastern corner, and a cluster of low rates in the middle and the western areas in Broward County).



(a) (b)

Figure 3.1. The spatial pattern of breast cancer rates in Broward County, FL, 2000-2010. (a) age-adjusted cancer rates for the census tract resolution [Moran's *I* (z-score, *p*-value): 0.19 (6.48, <0.0001)], and (b) crude cancer rates for the census tract resolution [Moran's *I* (z-score, *p*-value): 0.22 (7.62, < 0.0001)]. (Note that the *p*-values are calculated for a two-tail test.)

3.4.2. Regression results

This section summarizes regression results for breast cancer incidences in Broward County, FL. Standard Poisson and NB regression results are compared with their MESF counterparts; these models are specified with the logarithmic values of expected cancer cases as offset variables to be comparable with the age-adjusted incidence rates. The expected breast cancer cases were computed by applying the Broward County population-by-age cohorts to the 2013 US age-specific breast cancer incidence rates. The covariates included in the model specifications were identified from the literature, and summary statistics of ten covariates were reported in Table 3.1. These predictors mainly are socio-economic factors collected from US Census publications. *The primary care access variable* for the census tract resolution was created with 2013 primary care data retrieved from the US Department of Health and Human Service Administration

(https://www.hhs.gov/). Primary care availability has been identified as a significant risk factor for breast cancer; that is, people with poor primary care access are more likely to be diagnosed with breast cancer (e.g., Dai 2010, Wang et al. 2012).

Table 3.1. Independent variables included in regression analyses

Variables	Minimum	Mean	S.D.	Maximum
Female population	464	2,504	1,401,099	12,022
Female population density	0.0003	0.0055	0.0026	0.1738
The percentages of black population	0.000	0.2445	0.2626	0.982
The percentages of Hispanic population	0.002	0.219	0.132	0.638
The percentages of female population younger than 45	0.009	0.3323	0.086	0.639
The percentages of female-headed households	0.000	0.0541	0.030	0.207
The percentages of females with college degrees or higher	0.547	0.869	0.0924	0.989
Median household income	18,208	56,553	23,772	163,478
Urban/rural disparities (dummy variable)	---	---	---	---
The percentages of foreign born population	0.028	0.296	0.112	0.629
The number of physicians per 1,000 people	0.000	0.289	0.951	10.07

Note: S.D. denotes standard deviation

Table 3.2 reports Poisson and NB model specifications results. The standard Poisson model contains only covariates, and the two Poisson MESF models are specified with eigenvectors to account for SA. The first MESF model is specified with only PSA eigenvectors (MESF-Pos), with 41 PSA eigenvectors selected by a stepwise procedure. The second MESF model is specified with both PSA and NSA eigenvectors (MESF-Mix), with 53 eigenvectors selected by a stepwise procedure. These results show that the two MESF models perform better results, achieving smaller AIC and larger pseudo-R^2 values, than the standard Poisson model. However, the MESF-Mix describes the data better, having the lowest AIC and the highest pseudo-R^2 values. In addition, overdispersion decreases from 1.88 (standard Poisson model specification) to 1.2 (MESF-Pos), and then to 0.94 (MESF-Mix), becoming very close to its theoretical value of one.

Table 3.2. Estimation results for Poisson and negative binomial model specifications with further extensions to the Moran eigenvector spatial filtering technique

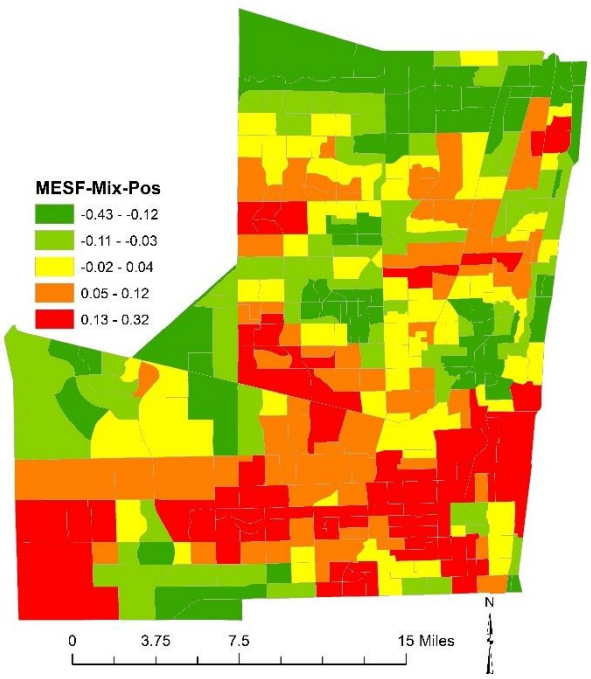
	Poisson						Negative Binomial					
	Standard Poisson			MESF-Pos		MESF-Mix			Standard NB			MESF-Pos		MESF-Mix	
Variables	coeff.	std. error	vif	coeff.	std. error	coeff.	std. error		coeff.	std. error	vif	coeff.	std. error	coeff.	std. error
Intercept	0.119	0.366	---	0.332	0.367	-0.091	0.315		0.049	0.373	---	0.332	0.335	-0.091	0.324
Female population density	-0.131	5.402	1.64	-0.984	4.870	-1.767	4.292		-2.904	5.365	1.63	-0.980	4.442	-1.768	4.418
% of black population	-0.045	0.079	3.14	-0.120.	0.071	-0.055	0.063		-0.045	0.078	3.31	-0.120	0.065	-0.055	0.064
% of Hispanic population	0.039	0.127	2.48	-0.493***	0.138	-0.223*	0.112		0.010	0.129	2.50	-0.490***	0.126	-0.223	0.115
% of females younger than 45	-0.499***	0.137	1.51	-0.408**	0.135	-0.214	0.112		-0.463**	0.144	1.40	-0.410***	0.123	-0.214	0.115
% of female householders	-0.712	0.798	1.74	-0.645	0.702	0.434	0.488		-0.701	0.779	1.76	-0.643	0.697	0.432	0.487
% of well-educated females	0.410	0.211	2.71	0.361	0.184	0.564***	0.168		0.406*	0.205	2.71	0.361*	0.168	0.564**	0.173
Median household income	-0.002	0.039	2.44	-0.005	0.037	-0.002	0.033		0.006	0.040	2.26	-0.010	0.034	-0.002	0.034
Urban/rural disparities	0.085	0.052	1.12	-0.035	0.048	0.064	0.041		0.080	0.054	1.13	-0.040	0.044	0.064	0.043
% of foreign born population	-0.529***	0.141	2.03	-0.323*	0.131	-0.387***	0.115		-0.498***	0.141	2.05	-0.320**	0.119	-0.387**	0.119
Physicians per 1,000 people	0.012	0.014	1.02	-0.009	0.013	0.001	0.011		0.004	0.014	1.03	-0.010	0.011	0.001	0.012
Psuedo-R^2	0.176			0.511		0.625			0.185			0.511		0.625	
AIC	2671.9			2470.0		2402.4			2595.9			2472.0		2404.4	
Overdispersion	1.88			1.20		0.94			0.02			<0.001		<0.001	
Moran's *I p-value* (one-tail test)	<0.001			0.999		0.335			<0.001			0.999		0.531	
# selected eigenvectors	---			41/137		53/358			---			41/137		53/358	

Significance codes: ***0.001, **0.01, *0.05

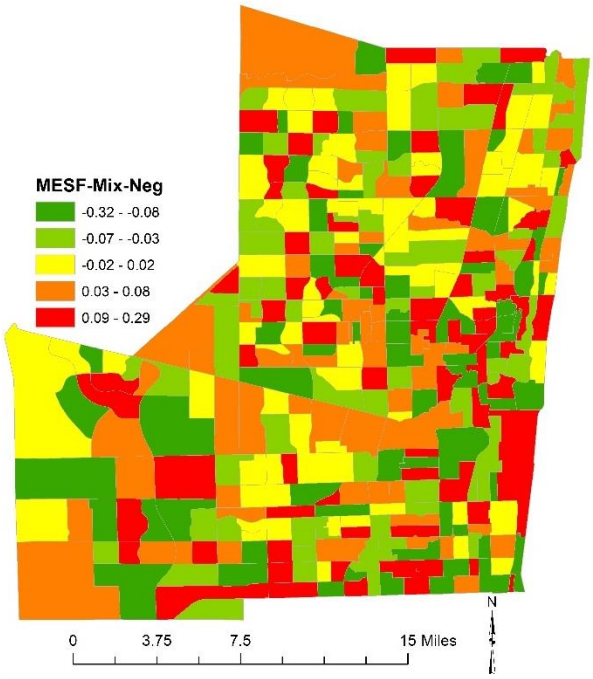
Note: NB denotes negative binomial, MESF denotes Moran eigenvector spatial filtering, MESF-Pos denotes spatial eigenvector spatial filtering model with positive spatial autocorrelation eigenvector only, MESF-Mix denotes spatial eigenvector spatial filtering model with both positive and negative spatial autocorrelation eigenvector, and vif denotes variance inflation factor

Moran's *I* test based on the method by Lin and Zhang (2007) indicates the presence of SA in the standard Poisson model residuals, NSA in the MESF-Pos model residuals, and insignificant SA in the MESF-Mix model residuals. These Moran's *I* statistics suggest that the standard Poisson model fails to account for SA in the cancer rates, and that the MESF-Pos model accounts for PSA components but still is unable to address a NSA component, leaving significant NSA in the residuals. In contrast, the MESF-Mix model successfully accounts for both PSA and NSA components. Hence, the MESF-Mix model is preferred to the other two models, especially from a SA perspective. This outcome suggests a possible mixture of PSA and NSA components in the data, although the global Moran's *I* exhibits only PSA in the residuals of the standard Poisson. Figure 3.2 portrays the two spatial filter components of the MESF-Mix model. Figure 3.2a portrays a PSA-only spatial filter component that is a linear combination of selected PSA eigenvectors, and Figure 3.2b depicts a NSA-only spatial filter component that is a linear combination of selected NSA eigenvectors.

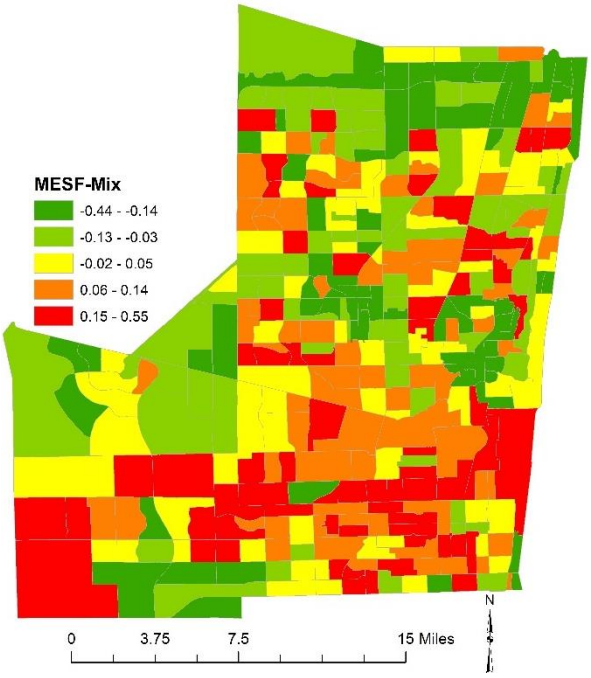
Accounting for SA leads to significance level changes for estimated covariate coefficients. First, one independent variable is not significant in the standard Poisson specification, but becomes significant at the 5% level in both MESF model specifications; it is *the percentage of Hispanic population*. Second, two independent variables have different significance results between the two MESF model specifications. *The percentage of younger female population* is significant in the standard Poisson specification and the MESF-Pos model, but is not significant at the 5% level in the MESF-Mix model. *The percentage of well-educated population* is not significant at the 5% level in the standard Poisson and MESF-Pos models, but is significant in the MESF-Mix models at the 1% level.


MESF-Mix-Pos
-0.43 - -0.12
-0.11 - -0.03
-0.02 - 0.04
0.05 - 0.12
0.13 - 0.32
N
0
3.75
7.5
15 Miles

(a)


MESF-Mix-Neg
-0.32 - -0.08
-0.07 - -0.03
-0.02 - 0.02
0.03 - 0.08
0.09 - 0.29
N
0
3.75
7.5
15 Miles

(b)


MESF-Mix
-0.44 - -0.14
-0.13 - -0.03
-0.02 - 0.05
0.06 - 0.14
0.15 - 0.55
N
0
3.75
7.5
15 Miles

(c)

Figure 3.2. Spatial filter components in the Poisson Moran eigenvector spatial filtering specifications for the census tract resolution: (a) the positive spatial autocorrelation spatial filter component [Moran's *I* (z-score, *p*-value): 0.59 (20.02, < 0.0001)], (b) the negative spatial autocorrelation spatial filter component [Moran's *I* (z-score, *p*-value): −0.25 (−8.14, <0.0001)], and (c) the spatial autocorrelation mixture spatial filter component [Moran's *I* (z-score, *p*-value): 0.37 (12.36, <0.0001)]. (Note that the *p*-values are calculated for a two-tail test.)

Results for the three NB regression models align with those for the Poisson specifications. The NB MESF-Mix specification outperforms the others by having the smallest AIC and the largest pseudo-R^2 values. Also, as for the Poisson specifications, the Moran's *I* tests suggest the presence of PSA in the standard NB residuals, NSA in the first MESF model residuals, and no SA in the second MESF model residuals. This result indicates a potential misspecification of the standard NB model as well as the first MESF model, similar to the Poisson case. The three NB models have coefficients very similar to their Poisson counterparts, and the significance levels for these estimated covariate coefficients are almost the same. Especially, the Poisson and NB MESF-Mix models have almost identical coefficients and significance levels. This outcome confirms that a Poisson specification produces the same results as a NB model specification when overdispersion is successfully explained, and a NB specification with additional parameters that capture extra variance does not produce a better model description of SA for these particular breast cancer data. In other words, successfully accounting for SA renders a more parsimonious model specification.

Table 3.2 also indicates that the following three significant covariates at the 5% level appear in the Poisson MESF-Mix model: *the percentage of Hispanic population*, *the percentage of well-educated population*, and *the percentage of foreign born population*. The regression coefficient estimates suggest that, on average, the percentage of Hispanic population (−0.223) and foreign born population (−0.387) with less education (0. 564) is less likely to develop breast cancer. This observation is consistent with findings stated in the literature. For example, Hussain et al. (2008) state that increased risk for *in situ* and invasive breast cancer is significantly associated with highly educated females; however, better educated females are more likely to survive breast cancer. DeSantis et al. (2014) find that the Hispanic population has a relatively lower breast cancer incidence rate. Carrière et al. (2013) also use an area-based methodology identify an

inverse relationship between breast cancer incidence rates and the concentration of foreign-born population in Canada. However, they argue that this association is difficult to interpret due to limited information of the study area, such as socio-economic status and health behaviors.

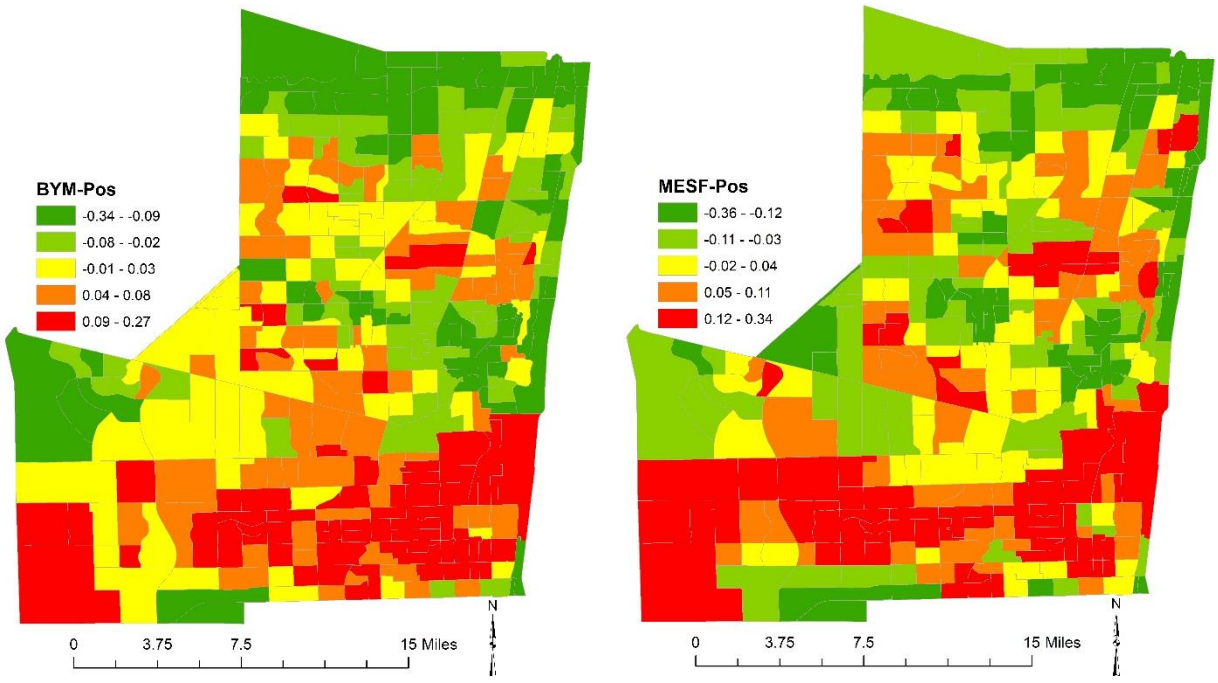
Table 3.3 summarizes the INLA estimation results for Poisson and NB model specifications. The INLA Poisson model results are comparable with those for the GLM standard Poisson model appearing in Table 3.2, identifying the same significant variables: *the percentage of females younger than 45*, and *the percentage of foreign born population*. In addition, the parameter estimates of the BYM Poisson (BYM-Pos) model are comparable with those for the Poisson MESF-Pos model in Table 3.2. Three variables appear significant in both of these specifications: the *percentage of Hispanic population, the percentage of female population younger than 45*, and *the percentage of foreign born population*. Similarly, a majority of the variables have almost identical coefficient estimates; one exception is *the female population density*. Figure 3.3 illustrates the spatial pattern captured by the random effects terms with the BYM-Pos model (Figure 3.3a), and by eigenvectors with the Poisson MESF-Pos model (Figure 3.3b). These map patterns are very similar, with clusters of high values observed in the south, and clusters of low values in the north and the east coastal areas. A comparison of these model results indicate that the BYM-Pos and Poisson MESF-Pos model specifications produce very similar results. This outcome may indicate that the Bayesian model successfully accounts for PSA but not the SA mixture component in the data.

Table 3.3. Estimation results for Poisson and negative binomial model specifications with the Besag-York- Mollié algorithm

	INLA Poisson			**Poisson-BYM**			**INLA NB**			**NB-BYM**		
	coeff.		std. error	coeff.		std. error	coeff.		std. error	coeff.		std. error
Intercept	0.054		0.377	0.206		0.425	0.075		0.381	0.214		0.431
Female population density	-2.182		5.327	-4.384		5.325	-2.410		5.371	-4.541		5.421
% of black population	-0.047		0.079	-0.111		0.093	-0.050		0.079	-0.112		0.094
% of Hispanic population	0.008		0.131	-0.342	*	0.166	0.002		0.132	-0.314		0.168
% of females younger than 45	-0.456	**	0.145	-0.408	**	0.154	-0.455	**	0.147	-0.400	**	0.157
% of female householders	-0.665		0.672	-0.477		0.312	-0.683		0.689	-0.469		0.301
% of well-educated females	0.402		0.206	0.388		0.217	0.397	*	0.208	0.362	*	0.220
Median household income	0.005		0.040	0.000		0.041	0.004		0.040	0.001		0.042
Urban/rural disparities	0.079		0.054	0.034		0.059	0.076		0.055	0.040		0.059
% of foreign born population	-0.513	***	0.142	-0.387	**	0.146	-0.505	***	0.143	-0.413	**	0.148
Physicians per 1,000 people	<0.001		<0.001	<0.001		<0.001	<0.001		<0.001	<0.001		<0.001
Watanable AIC	2576.2			2529.3			2596.1			2541.4		
DIC	2573.7			2511.5			2593.8			2536.7		

Significance codes: ***0.001, **0.01, *0.05

Note: NB denotes negative binomial, MESF denotes Moran eigenvector spatial filtering, BYM denotes Besag-York- Mollié, INLA denotes integrated nested Laplace, and vif denotes variance inflation facto



(a) (b)

Figure 3.3. Estimated spatial effects: (a) the random effects term estimated with the BYM-Pos model [Moran's *I* (z-score, *p*-value): 0.66 (25.23, < 0.0001)], and (b) the positive spatial autocorrelation spatial filter component estimated with the Poisson MESF-Pos model [Moran's *I* (z-score, *p*-value): 0.61 (20.60, < 0.0001)].

A comparison of the standard NB model estimates appearing in Tables 3.2 and 3.3 indicates that the estimated regression coefficients are almost identical for all variables, and significant variables in both models are the same. The estimated coefficients for the NB MESF-Pos and BYM NB (BYM-NB) models are very similar, except for two variables: *female population density*, and *urban/rural disparities*. Two variables are significant in both models; *the percentage of females younger than 45* variable is significant in the NB MESF-Pos model, wheras it is nonsignificant in the BYM-NB model. These results indicate that the NB MESF-Pos model is comparable with the BYM-NB model because they yield similar results, which may suggest that the BYM-NB model, unlike the NB MESF-mix model, is unable to account for positive and negative SA simultaneously.

3.5 Conclusions

This research utilizes the MESF methodology to uncover a SA mixture pattern in breast cancer data for Broward County, FL. Several conclusions can be drawn from an investigation of model specification issues with a breast cancer dataset for Broward County, FL. First, MESF model specifications successfully improve model performance. Specifically, the MESF-Pos models outperform the standard Poisson and NB models with a successful correction for PSA; however, the MESF-Pos model specifications leave significant NSA unexplained in the model residuals. The MESF-Mix models further improve model performance with the smallest AIC and the largest pseudo-R^2 values by accounting for both latent PSA and NSA components, with the Moran's *I* statistics suggest no SA in the MESF-Mix model residuals. These regression results and the spatial filter maps confirm a possible presence of a mixture of PSA and NSA in the age-adjusted rates, although the Moran's *I* suggests PSA only in the standard Poisson and NB model residuals, possibly because the PSA component outweighs the NSA component. This empirical analysis implies that the MESF-Mix model specification furnishes an efficient method to account for SA mixture components in georeferenced data.

Second, the BYM-Pos model generates results very similar to those for the Poisson MESF-Pos model, although they have very different model structures, which indicates that both the MESF and BYM models can furnish useful approaches to accommodate PSA in georeferenced data. However, a BYM model is limited in its ability to explain the presence of a PSA-NSA mixture because it fails to account for the hidden NSA that may be partially or fully masked by a globally dominating PSA component. This paper shows that MESF can address PSA and NSA components simultaneously with a set of PSA and NSA eigenvectors. Essentially, the PSA eigenvectors capture the PSA component, whereas NSA eigenvectors capture the NSA component. The difference in structure between a MESF and BYM model is worth noting here because the latter is based on a conditional autoregressive model with a first order variance structure, whereas the former can represent either a first- or a second-order variance structure (e.g., a simultaneous autoregressive model).

Third, excess Poisson variation is successfully accounted for by the MESF model specifications. Overdispersion is observed in the standard Poisson specification, and decreases in the MESF-Pos model specifications. Furthermore, the estimate of the overdispersion parameter becomes very close to one when both positive and NSA eigenvectors are included in an MESF model specification. This outcome confirms the impact of SA on overdispersion, as reported in the literature (e.g., Griffith 2007, Haining et al. 2009, Chun and Griffith 2011). The NB model specifications are not necessary when overdispersion is properly explained, and the need for a NB specification may suggest that overdispersion in the standard Poisson and MESF-Pos models is an outcome of model misspecification in terms of mixtures of PSA and NSA.

Fourth, the two different cancer rate measurements, crude rates and age-adjusted rates, reveal somewhat different spatial patterns of breast cancer incidence rates. Both the crude and age-adjusted rates visually exhibit PSA patterns that also are confirmed by Moran's *I* tests. However, the degree of PSA is stronger for the crude rates than for the age-adjusted rates; these discrepancies between the two rate measures may result from the generation of clusters of dissimilar rates while adjusting for cancer rates. For example, outliers (e.g., high adjusted cancer rates) can be triggered by small population counts in age cohorts, which potentially can distort the spatial structure of cancer rates. Although the Moran's *I* statistics in Figure 3.1 indicate the age-adjusted rates display PSA, clusters of similar and dissimilar values are simultaneously observed in the maps. The results of MESF model estimations also imply the presence of SA mixtures in these data.

Fifth, the MESF-Mix models are preferred to others because they yield the smallest AIC and the largest pseudo-R^2 values. These model results suggest that, on average, a higher breast cancer risk is associated with better educated non-Hispanic and non-foreign born females, which corroborates findings from previous studies. Age is considered as a critical risk factor for breast cancer, but the younger female population variable is insignificant for both resolutions. This outcome may be because the impact of age is already addressed in age-adjusted cancer rates.

Findings summarized in this paper can be further leveraged with future research. Mixtures of PSA and NSA need additional theoretical investigations to better understand their impact on modeling georeferenced data. Because findings reported in this paper are based on a single empirical data analysis, subsequent data analyses for other cancer types and study areas would better illuminate general aspects of the geographic distribution of cancer cases, such as the data containing a mixture of PSA and NSA, and contribute to the replicability of findings reported here.

CHAPTER 4

SPACE-TIME STATISTICAL INSIGHTS ABOUT GEOGRAPHIC VARIATION IN LUNG CANCER INCIDENCE RATES: FLORIDA, 2000-2011

Authors – Lan Hu, Daniel A. Griffith, Yongwan Chun

Geospatial Information Sciences, GR31

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

Hu, L., Griffith, D. A., & Chun, Y. (2018). Space-Time Statistical Insights about Geographic Variation in Lung Cancer Incidence Rates: Florida, USA, 2000–2011. *International journal of environmental research and public health*, *15*(11), 2406.

ABSTRACT

The geographic distribution of lung cancer rates tends to vary across a geographic landscape, and covariates (e.g., smoking rates, demographic factors, socio-economic indicators) commonly are employed in spatial analysis to explain the spatial heterogeneity of these cancer rates. However, such cancer risk factors often are not available, and conventional statistical models are unable to fully capture hidden spatial effects in cancer rates. Introducing random effects in the model specifications can furnish an efficient approach to account for variations that are unexplained due to omitted variables. Especially, a random effects model can be effective for a phenomenon that is static over time. The goal of this paper is to investigate geographic variation in Florida lung cancer incidence data for the time period 2000–2011 using random effects models. In doing so, a Moran eigenvector spatial filtering technique is utilized, which can allow a decomposition of random effects into spatially structured (SSRE) and spatially unstructured (SURE) components. Analysis results confirm that random effects models capture a substantial amount of variation in the cancer data. Furthermore, the results suggest that spatial pattern in the cancer data displays a mixture of positive and negative spatial autocorrelation, although the global map pattern of the random effects term may appear random.

4.1 Introduction

Spatial scientists, practitioners, and policy makers are interested in understanding the spatial variation in cancer rates at various geographic scales and resolutions (e.g., d'Onofrio et al., 2016). One commonly employed geographic resolution is the county, because aggregating especially counts of rare cancer cases by county for ecological analyses almost always preserves patient confidentiality (Wieland et al., 2008). This ethical and legal goal is at the expense of data analysis accuracy and precision, as well as data analytic complications such as the ecological fallacy. Meanwhile, depicting finer geographic resolution rates with choropleth or kernel density smoothed maps can maintain patient confidentiality while improving data analysis accuracy and precision (e.g., Lee et al., 2018), and help avoid or minimize such complications as the ecological fallacy.

Often, one objective of such research is to investigate associations between cancer rates and socio-economic/demographic characteristics. The availability of such covariates, generally retrieved from government census publications, tends to be very limited for fine geographic resolutions (e.g., census blocks). Furthermore, as a rare event, many cancer rates are often zero in a sizeable number of areal units at a very fine geographic resolution. Because a spatial analysis of cancer rates defies conducting scientific human subject experiments on ethical grounds, and, hence, is observational/correlational in nature, researchers seldom have a priori information about covariates that would make significant contributions to the geographic variation in cancer rates. Quasi-experimental designs have uncovered selected surrogate covariates supportable by scientific rationales, such as income/poverty and access to cancer screening/diagnosis/treatment (e.g., Smith et al., 2017), age and increased risk of developing cancer (e.g., Roquette et al., 2018), and education and lifestyle cancer-prevention behaviors (e.g., Wang et al., 2018).

A random effects (RE) model seeks to control for time-invariant unobserved heterogeneity in data. RE may be viewed as areal unit specific effects attributable to unknown latent variables. Although omitted variables influence a regression analysis of georeferenced cancer rates,

introducing a RE term can account for some of their effects. With regard to RE, Besag et al. (1991) specify that their geographic distributions comprise the following two components: spatially structured random effects (SSRE), and spatially unstructured random effects (SURE). A Bayesian hierarchical model, in which a prior distribution can substitute for repeated measures in a space-time series, can be specified to estimate both SSRE and SURE terms. In this Bayesian context, a SSRE component often is modelled with a conditional autoregressive (CAR) specification that captures spatial autocorrelation (SA) (SA refers to Tobler's first law of geography: everything is related to everything else, but nearby phenomena are more related than distant phenomena. In other words, either similar (positive SA) or dissimilar (negative SA) attribute values tend to cluster together on a map. Georeferenced data rarely are void of map pattern (i.e., have no SA)) latent in georeferenced data, and a SURE component is specified with an independent normal distribution. In this specification, SA is accounted for directly through parameters. In contrast, for space-time data, a second-order model is specified in which SA is accounted for through correlations among observations. That is, repeated measures are furnished for each spatial unit over time in a space-time data series. In this paper, an estimated RE term represents omitted variables, indicating that geographic distributions of cancer rates contain a considerable amount of unexplained variation, particularly at the census tract resolution. This cancer data expectation is attributable to the lengthy exposure lag that characterizes many cancers (i.e., a cancer being triggered long before its actual diagnosis), combined with the movement of people, and the departure of carcinogenic sources over time.

The specification of Besag et al. (1991) addresses only positive SA (PSA) situations. More recently accumulated empirical evidence indicates that a number of georeferenced phenomena seemingly exhibiting (near-) zero SA actually contain a mixture of PSA and negative SA (NSA) as two compensating components (Griffth and Arbia 2010). PSA arises from cooperative processes that involve intensifying spatial externalities, whereas NSA arises from competitive processes that involve abating spatial externalities. In this paper, a SSRE term consistently decomposes into a PSA-NSA mixture, which is barely investigated in the literature. Possible reasons for this mixed spatial pattern in cancer data include: (1) that geographic distributions of cancer cases display some degrees of global, regional, and local map patterns, which potentially

arise from a collocation of similar socio-economic/demographic characteristics in space (e.g., the Schelling model) (e.g., Mao et al., 2001); and (2) that these geographic distributions may display a degree of alternating map pattern trend because of local social networks that can induce an increasing cancer screening rate when someone in a neighborhood has a positive cancer diagnosis.

4.2 Literature review

A number of risk factors that are associated with lung cancer incidence have been examined and characterized in the literature (e.g., MacLennan et al., 1977; Mao et al., 2001; Molina et al., 2008); cigarette smoking is the most well-known factor that can trigger lung cancer. Studies also show that nonsmokers exposed to secondhand tobacco have higher risks of developing lung cancer (e.g., Alberg and Samet 2003). Other life style related risk factors, such as an unhealthy diet and alcohol consumption, increase the risk of developing lung cancer (e.g., Feskanich et al., 2000). Another suspicious contributor to human lung cancer burden is outdoor air pollution (e.g., fine particulate matter and a concentration of ozone); a number of studies examine, and the findings support, an association between air pollution and lung cancer risk (e.g., Pope et al., 2002; Vineis et al., 2004).

Indicators of socio-economic status also tend to be highly correlated with lung cancer risk. Due to their availability, the use of these variables has been popular amongst researchers to describe lung cancer incidence rates in the literature. For example, Mao et al. (2001) report a significant inverse relationship between high socio-economic status, and lung cancer risk. Socio-economic status reflects one's lifestyle, including diet, working and living conditions, enabling them to be treated as surrogates, and assumed to be associated with lung cancer (Osler 1993; Pomerleau et al., 1997). Specifically, the part of the population with lower socio-economic status (e.g., less educated, below a poverty level, unemployed) tends to have a higher risk of developing lung cancer than their counterparts that are classified with higher socio-economic status (Alberg et al., 2005).

In addition, the risk of developing lung cancer tends to vary across racial/ethnic, age, and sex groups. Alberg et al. (2005) argue that lung cancer incidence rates are similar for African and white Americans; however, a higher risk is observed for black American men than white American men. Haiman et al. (2006) comment that this difference is attributable to varying smoking behaviors among these ethnic/racial groups. Some case-control studies suggest higher risks of smoking-related lung cancer in women than men (e.g., Risch et al., 1993; Zang and Wynder 1996); however, this sex-difference in susceptibility to lung cancer still lacks supporting evidence. Age has been an important risk factor for most of cancers; the risk of lung cancer increases as age increases, seemingly as a part of the natural maturation process. Studies also report that immigration status plays a role in lung cancer risk; for example, United States (U.S.). Asian immigrants have higher lung cancer mortality rates than their U.S.-born counterparts; whereas the rates are lower among U.S. black immigrants than U.S.-born blacks (e.g., Singh and Miller 2004), This variation may be attributable to differences in smoking prevalence between the U.S. and the countries of origin, and differences across socio-economic classes (e.g., Blue and Fenelon 2011; Bosdriesz et al., 2013).

Lung cancer incidence rates generally are observed to vary substantially across geographic space. The literature suggest that air pollution is one of the major contributors to this geographic variation (Alberg and Samet 2003). For example, Jacquez and Greiling (2003) observe clusters of significantly high lung cancer incidence rates in central Long Island coinciding with a concentration of air toxics. The spatial variation of risk for lung cancer also is attributable to the geographic distribution of population. For example, Kelsall and Diggle (1998) report that the prevalence of lung cancer incidence is higher in areas with high social deprivation, which may directly link to smoking behavior and eating habits. A range of spatial models, including Bayesian space-time joint models (e.g., Richardson et al., 2006), spatial multilevel regression models (e.g., Jerrett et al., 2005), and conditional autoregressive models (e.g., Jin et al., 2005), have been applied to account for the geographic variation present in geospatial cancer data analyses.

A RE model frequently is utilized for a longitudinal data analysis exploiting repeated measures over time (Verbeke et al., 2010). For example, it has popularly been applied to model economic/social phenomena. Frondel and Vance (Frondel and Vance 2010) specify a RE model to estimate fuel price elasticities with household data. Clarke et al. (2010) use a model with both fixed and random effects to analyze the determinants of pupil achievement in primary school, finding that a RE model outperforms a fixed effects only model, based on statistical efficiency. Chen and Tarko (2014) employ a RE model to investigate traffic safety in highway work zones, with their results indicating that a RE model furnishes a good option for that type of research.

4.3 Data and methodologies

Lung cancer cases were obtained from the Florida cancer registry. After a data cleaning process that led to removal of duplicates (e.g., patients were diagnosed with lung cancer as a secondary cancer), records containing missing information (e.g., age and sex), and unsuccessfully geocoded records (i.e., failed-to-be-geocoded cases were deleted for the entire state, and then subsets were extracted from the clean dataset for specific study areas), 172,495 cancer incidences were used in data analyses. Cancer points are distributed unevenly across the 67 counties of the state, sample size ranging from 13,918 (Broward County) to 31 (Liberty County), with a median of 1,277 (Santa Rosa County). These lung cancer incidences occurred in a 12-year span, from 2000 to 2011. At the block group resolution, a relatively fine geographic resolution, many block groups have zero cancer incidences. In contrast, only 1.98% of the census tracts, a coarser geographic resolution, have zero cancer counts. To avoid the issue of excessive zeros, this research focuses on two geographic resolutions, namely county and census tract, for comparison purpose. In addition, this paper limits its study area to six different metropolitan statistical areas (MSAs) focusing on relatively highly densely populated areas in the state: Pensacola, Tallahassee, Jacksonville, Orlando, Miami, and Tampa.

4.3.1. Lung Cancer Incidence Rates

The crude cancer incidence rate, the ratio of cancer counts and population size at risk, generally is considered as a limited measure because cancer generally occurs at different rates based on

age, gender, and even racial group composition of a population. A comparison of crude cancer rates over time or across different geographic areas is likely to be plagued by bias because of different local population compositions (Anderson and Rosenberg 1998). Standardization of disease rates has been proposed to control for changes in population structure. Adjustments of cancer rates for age is a frequently applied standardization (Ahmad et al., 2001). The Centers for Disease Control and Prevention (CDC) also adopts this approach for statistical report purposes. With the availability of age and sex information for lung cancer patients, this research adjusts lung cancer incidence rates for both age and sex.

Figure 4.1 portrays the geographic distribution of adjusted lung cancer incidence rates across the State of Florida and its six MSAs. The Moran coefficient (MC) and Geary Ratio (GR) statistics suggest adjusted cancer rates exhibit a very weak PSA map pattern at the county resolution (Figure 4.1a), and random spatial patterns at the census tract resolution (Figures 4.1b–4.1g). Compared with the crude lung cancer incidence rates summarized in the Appendix 4A (Figure A4), the standardization process tends to reduce spatial clusters of similar cancer rates (i.e., clusters of high values or low values), and generate alternating patterns (i.e., a low lung cancer rate is surrounded by high rates for its neighbors, or a high lung cancer rate is surrounded by low rates for its neighbors) at both the county and census tract resolutions. In addition, due to relatively small populations at the census tract resolution, rate adjustment triggers outliers (e.g., high cancer rates). For example, the highest adjusted cancer rate in the Miami MSA reaches 2.73%, whereas the highest crude rate is 0.36%. Also, more census tracts stand out with high adjusted cancer rates compared with their corresponding crude ones. To mitigate negative impacts of extreme outliers, census tracts with small populations but some cancer counts are aggregated with their neighboring tracts for the analyses summarized in this paper. Specifically, the Miami MSA has 19 such census tracts that were merged into their adjacent tracts; the Tampa and Orlando MSAs have, respectively, five and one such census tracts. Most of these merged census tracts involve commercial, industrial, or coastal land use.

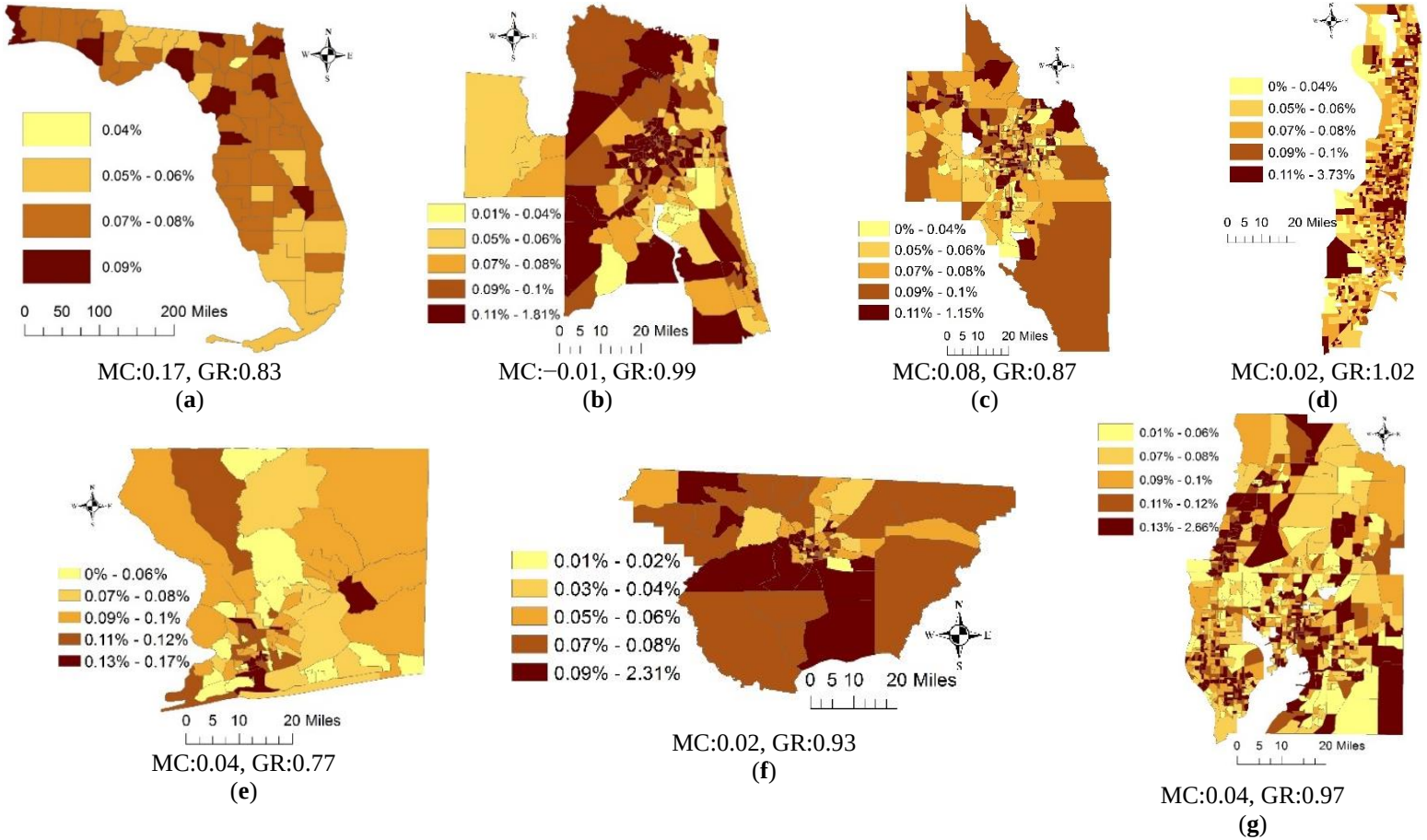


Figure 4.1. The spatial patterns of adjusted lung cancer incidence rates. (**a**) the State of Florida counties. Census tracts for: (**b**) the Jacksonville MSA; (**c**) the Orlando MSA; (**d**) the Miami MSA; (**e**) the Pensacola MSA; (**f**) the Tallahassee MSA; (**g**) the Tampa MSA

4.3.2. Moran Eigenvector Spatial Filtering

Moran Eigenvector spatial filtering (MESF) is a spatial statistical methodology that introduces a set of eigenvectors into a regression model specification to capture SA. Eigenvectors can be extracted from a transformed spatial weights matrix $\mathbf{C}$, which can be expressed as:

$$\mathbf{MCM} = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n), \quad (1)$$

where I is an n-by-n identity matrix, 1 is a n-by-1 vector of ones, n is the number of areal units, and T is the matrix transpose operator. This transformed spatial weights matrix generates n eigenvectors; however, only a subset of them serves as independent variables to be included in a model specification (Griffith 2003). This subset can be identified from a candidate eigenvector set with a stepwise regression procedure (Chun et al., 2016).

A RE model can be specified as:

$$Y = X\beta X + Z + \varepsilon, \quad (2)$$

where Y denotes a response variable, X denotes a matrix of covariates, βX denotes regression coefficients for covariates, Z denotes a RE term, and ε denotes a regression error term. The RE term, Z, is commonly assumed to be normally distributed and uncorrelated with both covariates and residuals, and to have a mean of zero. In order to estimate the RE term and separate it from the residual error ε, additional information (e.g., repeated measures furnished in a space-time series, or priors in a Bayesian analysis) are necessary (e.g., Griffith 2006). A RE model can be further extended with MESF, in order to accommodate both SSRE and SURE terms simultaneously, as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_\mathbf{X} + \mathbf{E}_k\boldsymbol{\beta}_\mathbf{E} + \mathbf{Z}_{\text{SURE}} + \boldsymbol{\varepsilon}, \quad (3)$$

where $\mathbf{E}_k$ denotes a subset of eigenvectors, and $\boldsymbol{\beta}_\mathbf{E}$ are unknown coefficients for these eigenvectors. $\mathbf{E}_k\boldsymbol{\beta}_\mathbf{E}$ furnishes a SSRE term, and $\mathbf{Z}_{\text{SURE}}$ denotes a SURE term. That is, the RE term, $\mathbf{Z}$, is decomposed into the linear combination of $\mathbf{E}_k\boldsymbol{\beta}_\mathbf{E}$ and $\mathbf{Z}_{\text{SURE}}$. Furthermore, a separation of

the selected eigenvectors, $\mathbf{E}_k$, into PSA and NSA eigenvectors, can furnish a way to investigate PSA and NSA components in a SSRE.

In this paper, space-time lung cancer counts (e.g., n-by-T = 67-by-12 for the county resolution) furnish the repeated measures for the response variable. The count variable is described with a Poisson probability model by including the logarithmic values of expected lung cancer counts as an offset variable. After a RE term successfully is estimated using the Poisson RE model, a MESF model is specified to estimate the SSRE and SURE components, with the estimated RE term as the independent variable. Essentially, a linear combination of the selected eigenvectors constructs a SSRE term, which is further decomposed into a PSA-NSA mixture (Griffith 2006), and the MESF model residual constitutes the SURE term. Poisson RE and MESF models were implemented in R 3.4.2.; the glmer procedure (package lme4) was utilized to estimate the RE components.

4.4 Results and discussion

This section summarizes analysis results for both county and census tract resolutions. Regression results for quasi-Poisson and Poisson RE models are compared, and the estimated RE components are portrayed with maps.

4.4.1. The State Scale and County Resolution

Seven variables were retrieved to describe lung cancer incidence rates at the county resolution, including smoking rates from the Florida Department of Health, and socio-economic variables, which are median household income, the percentage of population with a college or higher degree, the percentage of population below a poverty threshold, the percentage of Hispanic population, the percentage of black population, and immigrants, from the U.S. Census Bureau. Table 4.1 summarizes the estimation results for a Poisson RE model, as well as the results of a quasi-Poisson model for comparison purpose. It shows that the lung cancer data has considerable overdispersion (i.e., excess Poisson variation). However, the extra-Poisson variation successfully is accounted for in the RE model, with the overdispersion parameter decreasing from 13.36 to

2.15. Moreover, an inclusion of the RE term leads to an increase in the pseudo-R^2, increasing it from 0.30 to 0.74. The VIF values are all less than 10 (e.g., O'brien 2007; Craney and Surles 2002), indicating no excessive multi-collinearity among the covariates.

Table 4.1 also reports standard errors increase in the Poisson RE model specification, which results in significance level changes for some covariates, compared with the results of the covariates-only quasi-Poisson specification. For example, the ratio of population with a college or higher degree, the ratio of population under poverty, and the ratio of black population become insignificant in the RE model. The immigrant variable is included mainly because the State of Florida has gained a large number of immigrants, and papers in the literature argue that lung cancer risk may vary among U.S. residents and immigrants, as discussed in the preceding background. However, the immigrant variable does not have a significant association with lung cancer risk in both models. The only significant variable in the Poisson RE model is the smoking rate, which exhibits a positive relationship with lung cancer risk. The estimated RE term has a mean of zero, and is not correlated with the covariates, as expected.

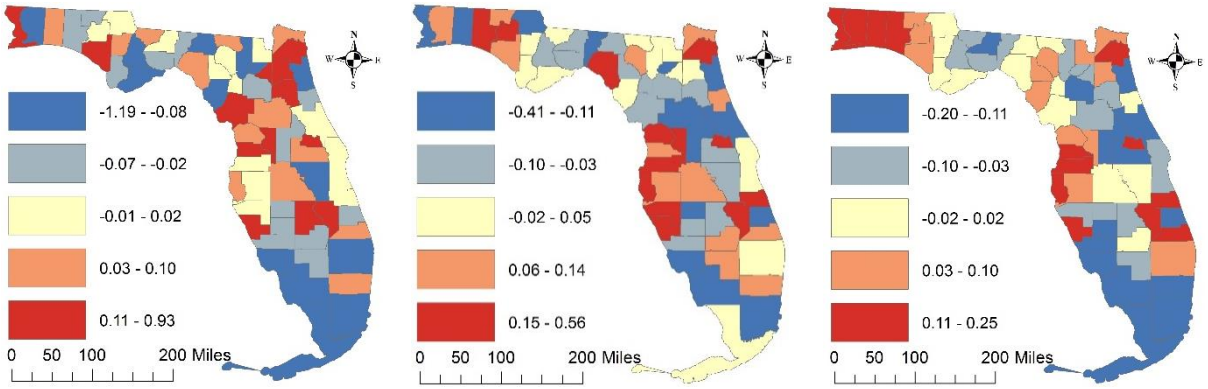
Table 4.1. Estimation results for Poisson models at the county resolution.

Variables	**Quasi-Poisson Model**			**Poisson Random Effects Model**		
	Coeff.	Std. Error	VIF	Coeff.	Std. Error	Cor. †
Smoking	4.060 ***	0.317	2.158	1.355 *	0.994	<0.001
Income	−0.262	0.262	2.763	0.191	0.617	−0.034
Education	−0.983 *	0.443	4.150	1.116	0.928	<0.001
Poverty	−4.368 ***	1.027	7.584	1.608	2.191	<0.001
Hispanic pop	−0.027	0.161	4.738	0.051	0.074	0.074
Black pop	1.587 ***	0.284	6.005	−0.627	0.427	0.067
Immigrants	−0.015	0.013	2.449	0.033	0.050	0.021
Overdispersion	13.02			2.12		
Pseudo-R^2	0.30			0.75		

Significance codes: ***0.001, **0.01, *0.05, · 0.1.
† This represents correlation coefficients between the RE term and the covariates.

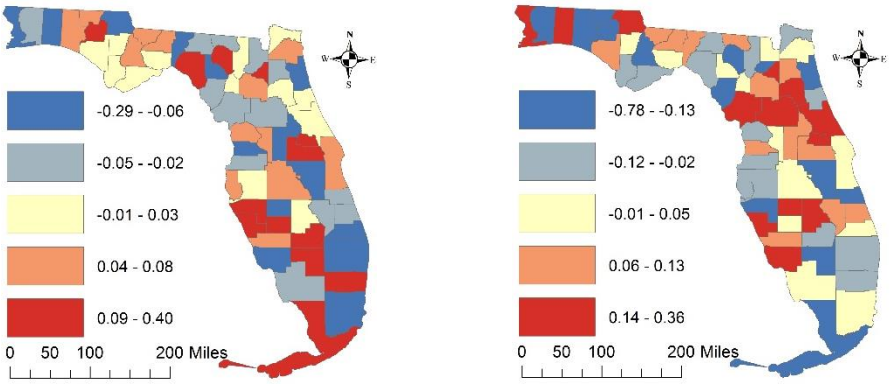
Figure 4.2 portrays the geographic distributions of RE components at the county resolution. The counties with high/low adjusted lung cancer rates in Figure 4.1a also are conspicuous in Figure 4.2a, which captures the major spatial pattern of lung cancer rates. However, the MC values suggest that both the RE and SSRE terms contain trace amounts of SA, which means inclusion of

the covariates in the Poisson mixed model explains some degrees of the PSA component observed on Figure 4.1a. The *p*-values of the Shapiro-Wilk (S-W) normal diagnostic statistic indicate that neither closely conforms to a normal distribution. The decomposition of the SSRE term yields a mixture of moderate-to-strong PSA (Figure 4.2c) and moderate NSA (Figure 4.2d). The *p*-values of the S-W statistic indicate that SSRE-PSA and SSRE-NSA are normally distributed. The MC suggests no significant SA in the SURE component, and that it deviates from a bell-shape curve.



MC: 0.080, *P*(S-W): <0.001 MC: 0.027, *P*(S-W): 0.035 MC: 0.616, *P*(S-W): 0.758

(**a**) (**b**) (**c**)



MC: −0.506, *P*(S-W): 0.159 MC: 0.104, *P*(S-W): 0.001

(**d**) (**e**)

Figure 4.2. Spatial patterns of RE components for the county resolution. (**a**) the RE term; (**b**) the SSRE term; (**c**) the SSRE-PSA term; (**d**) the SSRE-NSA term; (**e**) the SURE term.

4.4.2. The Metropolitan Statistical Area Scale and Census Tract Resolution

Smoking prevalence data are not available at the census tract resolution. So only socio-economic and demographic variables were included to describe lung cancer incidence rates. Results for Poisson RE models for each MSA are compared with covariate-only quasi-Poisson regression results. The overdispersion values larger than one in Table 4.2 indicate the lung cancer counts are slightly overdispersed for all MSA cases. However, all of them get closer to one for the mixed models. The pseudo-R^2 increases suggest improvements of model performance for all MSAs. A comparison of Tables 4.2 and 4.3 shows that standard errors get larger for the Poisson RE model specifications, which may have an impact on the significance level of independent variables. Including the RE terms also enhances model performance; all RE specifications have larger pseudo-R^2 values.

Tables 4.2 and 4.3 show that median household income is significant in all specifications, and has a negative association with lung cancer risk. Although the well-educated population variable is significant in some cases, exhibiting an inverse relationship, the population below poverty variable tends to be positively associated with lung cancer rates. The relationships between these socio-economic indicators and lung cancer risk corroborates the findings in the literature (e.g., Ward et al., 2004; Clegg et al., 2009). For demographic factors, the estimated results suggest lower lung cancer risks for Hispanics, blacks, and immigrants. Stellman et al. (2003) comment that the white and black populations have similar lung cancer risks if their smoking habits are similar. However, studies (e.g., Muscat et al., 2002) find that Caucasians are more likely to be heavier smokers than African-American, which makes them more susceptible to lung cancer. Singh and Miller (2004) observe that although lung cancer risk varies among different racial/ethical groups, it tends to be lower among U.S. immigrants due to a relatively lower smoking prevalence.

Intercept-only RE models are specified for each study area to examine the spatial variation in lung cancer incidence rates. Table 4.4 summarizes the amount of variation explained by the RE terms. It indicates that the RE terms explain a substantially smaller amount of variations at the

census tract resolution than at the county resolution. In addition, this percentage varies across the six MSAs, with the Tallahassee MSA having the lowest statistical explanation (11.64%), and the Pensacola MSA having the highest statistical explanation (27.13%). The average percentage of variation accounted for by the RE terms is roughly 21%, indicating a tremendous amount of unexplained geographic variation in lung cancer rates, particularly at the census tract resolution. Figure 4.3 depicts the amount of variation accounted for by each RE component beyond that by the covariates. The SSRE and SURE components constituting a RE term explain almost the same amount of variation across all MSAs. Meanwhile, for the two sub-terms of the SSRE, the SSRE-NSA term outperforms the SSRE-PSA term for the Orlando, Pensacola, Tallahassee, and Tampa MSAs.

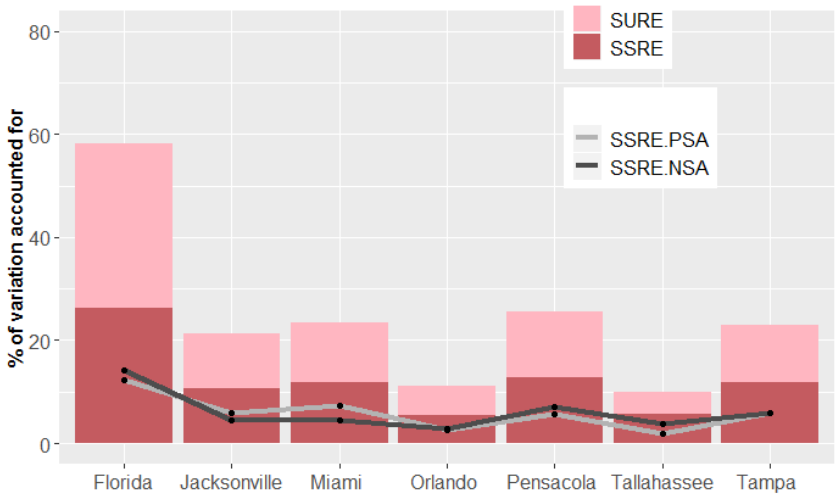
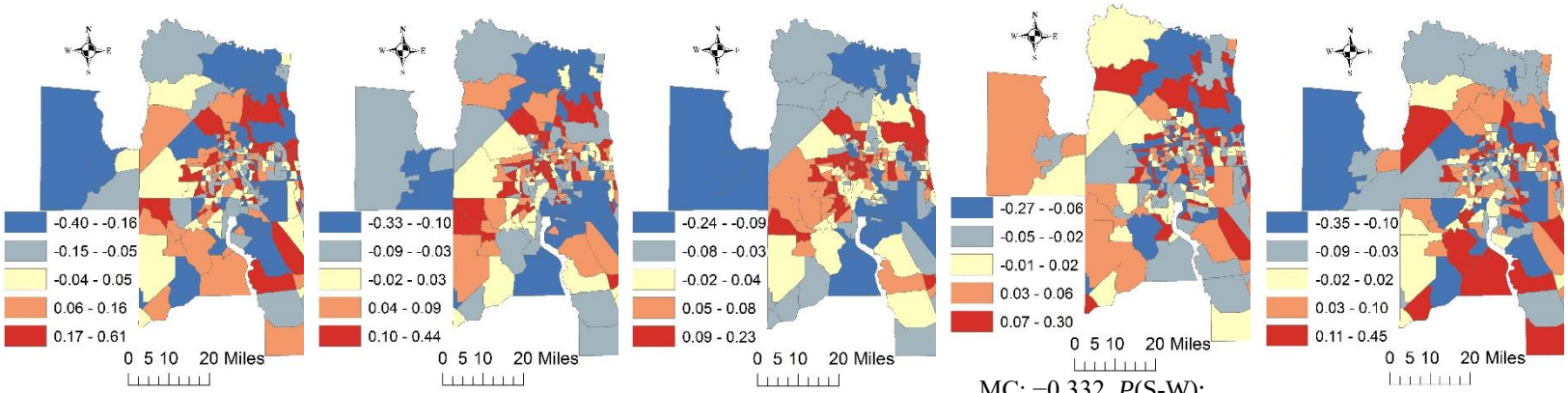


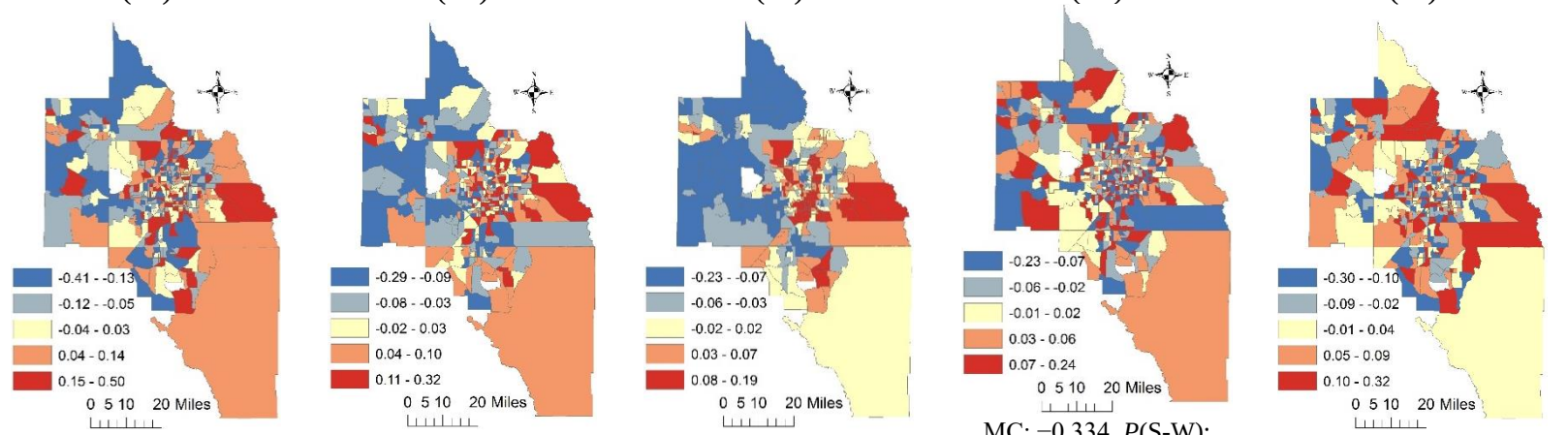
Figure 4.3. The amount of geographic variation in lung cancer incidence rates accounted for by the RE terms. The first bar is for Florida at the county resolution, and the other six are for the MSAs at the census tract resolution.

Figure 4.4 portrays the spatial patterns of RE components for the six MSAs. Because the RE components account for relatively low percentages of the geographic variation at the census tract resolution, Figures 4.4a1–4.4a6 do not reflect the map patterns of adjusted lung cancer rates well; however, they capture high cancer rates in urban areas, and low rates in rural areas for most of the MSAs, which also are highlighted on their corresponding cancer rates maps. For example, Figure 4.4a3 highlights census tracts within Fort Lauderdale and Pompano Beach that have relatively high cancer rates, which also stand out in Figure 4.1d. The MCs imply a presence of

weak PSA in the RE components, except for the Pensacola and Tallahassee MSAs, and the *p*-values of the S-W statistic indicate that they all barely conform to normal distributions. After a removal of the SURE components from the RE terms, stronger PSA is detected in the SSRE components, with increasing MC values for most MSAs (Figures 4.4b1–4.4b6). However, the SSRE components in the Pensacola and Tallahassee MSAs still exhibit (near-) zero SA. Similarly, a decomposition of these SSRE terms yields mixtures of moderate-to-strong PSA components (Figures 4.4c1–4.4c6) and weak-to-moderate NSA components (Figures 4.4d1–4.4d6) for all MSAs. The *p*-values of the S-W statistic suggest that all of the SSRE-PSA and SSRE-NSA terms closely conform to normal distributions, except for the Jacksonville MSA. Map patterns displayed in Figures 4.4e1–4.4e6 appear random, an outcome confirmed by their insignificant MCs. All of the SURE components are normally distributed.

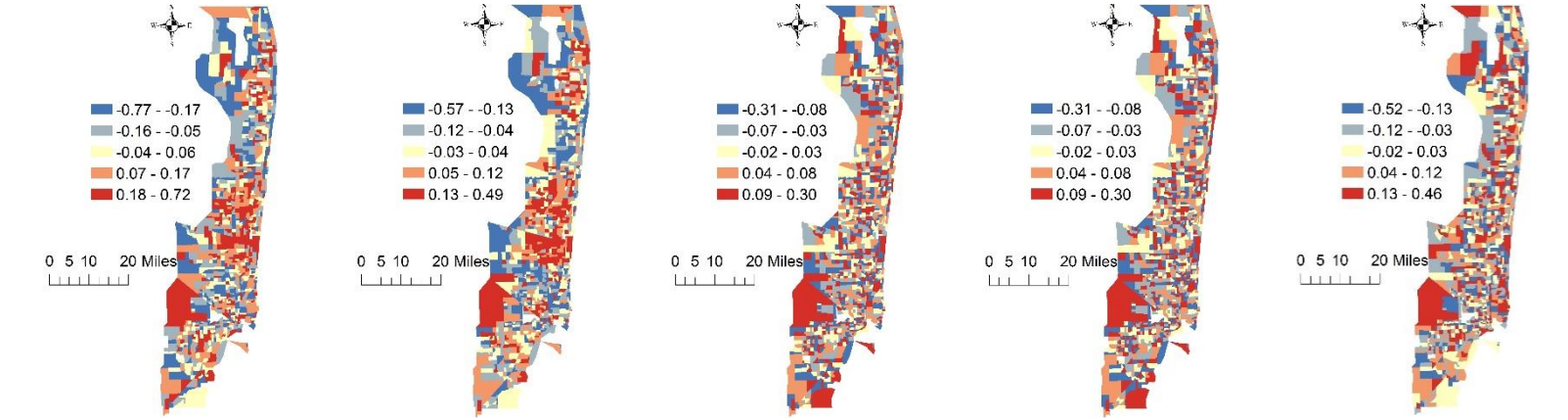


MC: 0.133, *P*(S-W): 0.062 (**a1**) MC: 0.259, *P*(S-W): 0.006 (**b1**) MC: 0.693, *P*(S-W): 0.022 (**c1**) MC: −0.332, *P*(S-W): 0.010 (**d1**) MC: 0.002, *P*(S-W): 0.125 (**e1**)

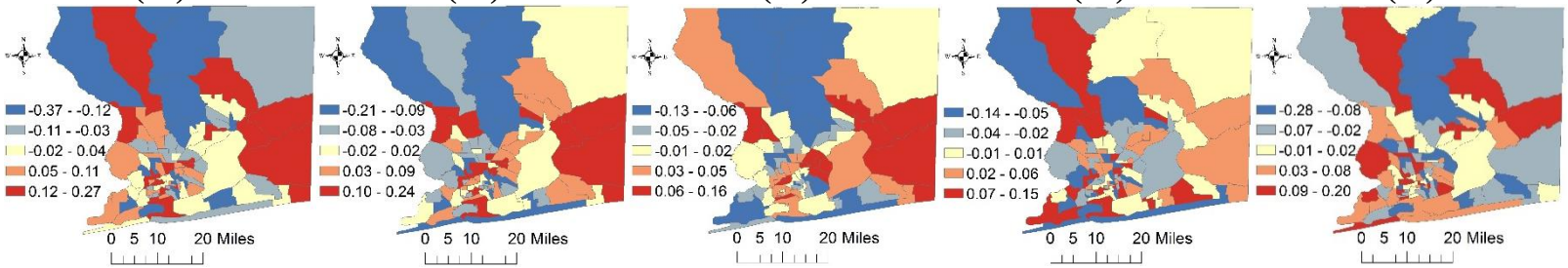


MC: 0.100, *P*(S-W): 0.039 (**a2**) MC: 0.173, *P*(S-W): 0.680 (**b2**) MC: 0.765, *P*(S-W): 0.148 (**c2**) MC: −0.334, *P*(S-W): 0.961 (**d2**) MC: 0.027, *P*(S-W): 0.680 (**e2**)

Figure 4.4. Spatial patterns of RE components at the census tract resolution. (**a1-a6**) the RE terms; (**b1**–**b6**) the SSRE terms; (**c1**–**c6**) the SSRE-PSA terms; (**d1**–**d6**) the SSRE-NSA terms; (**e1**–**e6**) the SURE terms. Rows from top to bottom: Jacksonville, Orlando, Miami, Pensacola, Tallahassee, and Tampa MSAs.

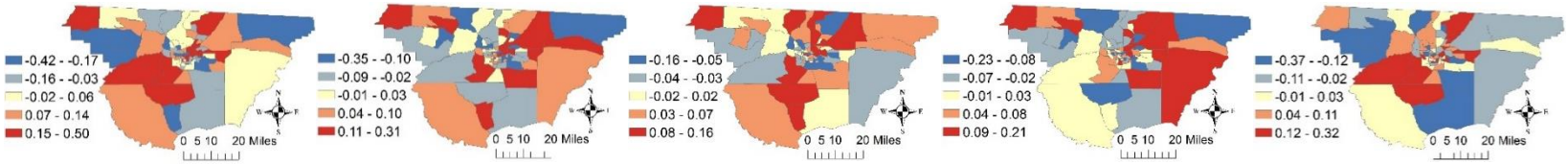


MC: 0.182, *P*(S-W): 0.019 (**a3**) | MC: 0.343, *P*(S-W): 0.505 (**b3**) | MC: 0.756, *P*(S-W): 0.259 (**c3**) | MC: −0.336, *P*(S-W): 0.502 (**d3**) | MC: 0.003, *P*(S-W): 0.852 (**e3**)

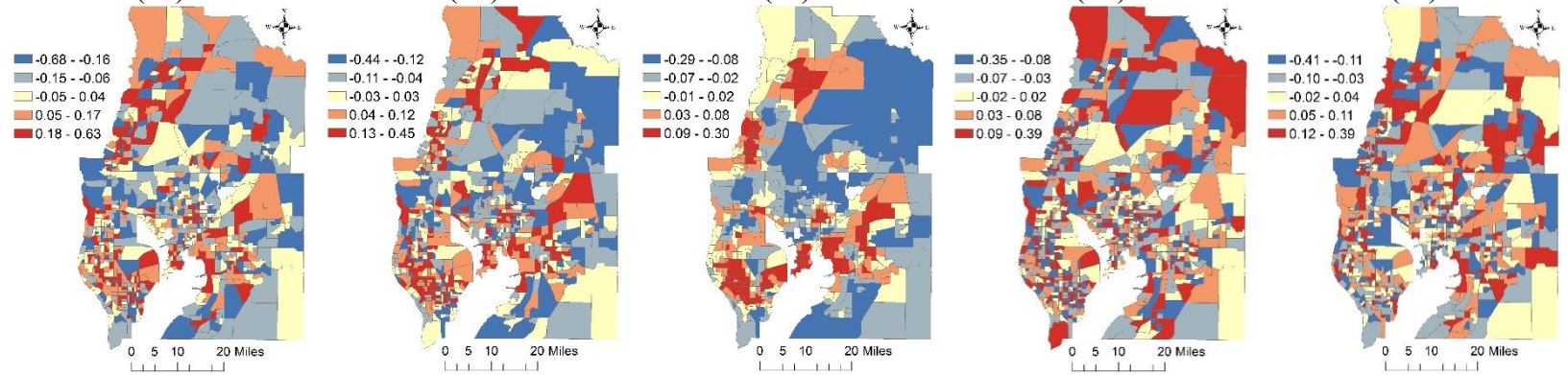


MC:0.050, *P*(S-W): 0.016 (**a4**) | MC:0.038, *P*(S-W): 0.500 (**b4**) | MC:0.551, *P*(S-W): 0.751 (**c4**) | MC:−0.300, *P*(S-W): 0.300 (**d4**) | MC:0.050, *P*(S-W): 0.514 (**e4**)

Figure 4.5. (Continued) Spatial patterns of RE components at the census tract resolution. (**a1-a6**) the RE terms; (**b1**–**b6**) the SSRE terms; (**c1**–**c6**) the SSRE-PSA terms; (**d1**–**d6**) the SSRE-NSA terms; (**e1**–**e6**) the SURE terms. Rows from top to bottom: Jacksonville, Orlando, Miami, Pensacola, Tallahassee, and Tampa MSAs.



MC:0.054, *P*(S-W): 0.808 (**a5**) MC:0.067, *P*(S-W): 0.759 (**b5**) MC:0.796, *P*(S-W): 0.278 (**c5**) MC:0.036, *P*(S-W): 0.870 (**d5**) MC:0.036, *P*(S-W): 0.870 (**e5**)



MC: 0.129, *P*(S-W): 0.160 (**a6**) MC: 0.218, *P*(S-W): 0.087 (**b6**) MC: 0.788, *P*(S-W): 0.455 (**c6**) MC: −0.336, *P*(S-W): 0.217 (**d6**) MC: 0.014, *P*(S-W): 0.100 (**e6**)

Figure 4.6. (Continued) Spatial patterns of RE components at the census tract resolution. (**a1-a6**) the RE terms; (**b1**–**b6**) the SSRE terms; (**c1**–**c6**) the SSRE-PSA terms; (**d1**–**d6**) the SSRE-NSA terms; (**e1**–**e6**) the SURE terms. Rows from top to bottom: Jacksonville, Orlando, Miami, Pensacola, Tallahassee, and Tampa MSAs.

Table 4.2. Estimation results for quasi-Poisson model specifications at the census tract resolution.

Variables	Pensacola MSA				Tallahassee MSA				Jacksonville MSA				Orlando MSA				Miami MSA				Tampa MSA			
	Coeff.		Std. Error	Vif	Coeff.		Std. Error	Vif	Coeff.		Std. Error	Vif	Coeff.		Std. Error	Vif	Coeff.		Std. Error	Vif	Coeff.		Std. Error	Vif
Income	−1.10	***	0.19	3.25	−0.91	**	0.17	2.68	−0.80	***	0.09	2.46	−0.73	***	0.07	1.75	−0.40	***	0.03	1.75	−0.58	***	0.05	1.77
Education	−0.62		0.33	2.80	0.43		0.40	2.32	−0.86	***	0.20	2.56	−0.44	*	0.20	2.47	−0.35	***	0.09	4.00	−0.56	***	0.12	2.29
Poverty	0.85	**	0.31	3.15	0.94	*	0.37	3.15	0.12		0.19	2.66	1.23	***	0.19	2.18	0.01		0.10	2.54	0.54	***	0.11	2.24
Hispanic pop	0.96		0.72	1.14	−0.68		0.54	1.14	0.81	***	0.24	1.11	−0.58	***	0.07	1.39	−0.51	***	0.03	2.22	−0.17	***	0.06	1.44
Black pop	−0.55	***	0.13	2.56	−0.72	***	0.16	2.15	−0.19	***	0.06	2.10	−0.30	***	0.07	1.95	−0.19	***	0.03	1.92	−0.08	.	0.05	1.50
Immigrants	−6.61	*	3.12	1.16	−2.69		4.29	1.28	−3.07	.	1.60	1.06	−9.15	***	1.32	1.29	10.21	***	1.29	1.11	−6.06	**	1.86	1.21
Overdispersion	1.08				1.16				1.26				1.22				1.27				1.30			
Pseudo-R^2	0.14				0.17				0.17				0.40				0.43				0.36			

Significance codes: ***0.001, **0.01, *0.05, · 0.1.

Table 4.3. Estimation results for Poisson RE model specifications at the census tract resolution.

Variables	Pensacola MSA				Tallahassee MSA				Jacksonville MSA				Orlando MSA				Miami MSA				Tampa MSA			
	Coeff.		Std. Error	Cor. †	Coeff.		Std. Error	Cor. †	Coeff.		Std. Error	Cor. †	Coeff.		Std. Error	Cor. †	Coeff.		Std. Error	Cor. †	Coeff.		Std. Error	Cor. †
Income	−1.06	***	0.27	0.02	−0.99	***	0.23	0.08	−0.78	***	0.12	<0.01	−0.79	***	0.10	−0.02	−0.41	***	0.05	<0.01	−0.65	***	−0.65	−0.04
Education	−0.79	*	0.48	0.01	0.53		0.56	−0.05	−0.86	***	0.31	<0.01	−0.48	*	0.28	0.01	−0.47	***	0.14	<0.01	−0.66	***	−0.66	<0.01
Poverty	0.73		0.46	−0.01	0.91		0.48	−0.07	0.17		0.28	−0.01	1.02	***	0.27	0.01	0.06		0.15	<0.01	0.30	*	0.30	0.05
Hispanic pop	0.44		1.02	0.01	−0.61		0.79	0.10	0.86	***	0.37	<0.01	−0.69	***	0.10	0.01	−0.58	***	0.04	<0.01	−0.17	***	−0.17	−0.02
Black pop	−0.50	***	0.20	−0.01	−0.69	***	0.22	−0.04	−0.17	***	0.09	<0.01	−0.31	***	0.10	<0.01	−0.27	***	0.05	<0.01	−0.03		−0.03	−0.02
Immigrants	−7.82	*	4.52	0.02	−4.91		5.80	−0.05	−2.86		2.58	<0.01	−9.59	***	1.95	<0.01	8.06	***	2.13	<0.01	−9.33	***	−9.33	−0.02
Overdispersion	1.03				1.12				1.10				1.06				1.09				1.07			
Pseudo-R^2	0.19				0.23				0.22				0.44				0.51				0.45			

Significance codes: ***0.001, **0.01, *0.05, · 0.1.

† This represents correlation coefficients between the RE term and covariates.

Table 4.4. The amount of variation accounted for by the RE terms.

Models	Florida	Pensacola MSA	Tallahassee MSA	Jacksonville MSA	Orlando MSA	Miami MSA	Tampa MSA
RE models intercept-only	58.39%	27.13%	11.64%	25.14%	13.68%	24.46%	23.88%
RE models with covariates	58.19%	25.53%	9.91%	21.20%	11.14%	23.47%	22.98%

4.5 Conclusions

This research examines the spatial patterns of lung cancer incidence rates at different geographic resolutions and scales in Florida, and also investigates factors that are associated with lung cancer risk. Major findings are as follows. First, lung cancer count data contain a substantial amount of overdispersion (13.36) at the county resolution, whereas they are much less overdispersed (less than 2) at the census tract resolution. A RE model specification successfully addresses this issue. Because the estimated overdispersion parameter is closer to 1 for the RE model specifications, substitution of a negative binomial model becomes unnecessary, which is a desirable outcome given reservations expressed by (Diggle and Milne 1983) concerning the suitability of this latter specification for SA situations. Second, a RE model furnishes an efficient method to correct for biased estimation (e.g., underestimated standard errors). Regression results indicate that an inclusion of a RE term, which can serve as a proxy for omitted variables, improves model performance (e.g., it increases pseudo-R^2 values). Third, estimated results suggest that a risk of lung cancer is positively associated with smoking behavior, and the percentage of population with low socio-economic status (e.g., low household income, poor education), and negatively associated with the percentage of black/Hispanic population, and the percentage of immigrants. These positive/negative relationships corroborate findings already appearing in the literature.

This research contributes to the literature in the following two ways. First, this research shows that the RE model specifications improve model performance by including a RE terms that successfully accounts for variation beyond that attributable to covariates. Here the RE terms account for 58.39% of the geographic variation in lung cancer incidence rates at the county resolution, and 21% of this variation, on average, at the census tract resolution. This outcome indicates that considerable unexplained variation exists in the lung cancer data at the census tract resolution. This poor statistical explanation probably is attributable to two major factors: one is aggregating cancer cases into a coarser resolution (e.g., county) averages out noises that present in a finer resolution (e.g., census tract) (Openshaw 1984). Second is due to the massive immigration to Florida. Generally speaking, population migration over time can contribute to a

change in cancer rates, and results in an introduction of a source of variation that is not well described with RE. A purposeful migration for health issues can have a large impact. For example, unhealthy immigrates would choose to move closer to health facilities, or move away from contaminated areas, whereas healthy people relocate to regions that are economically better off (e.g., Bentham 1988; Boyle et al. 2002). Immigration, thus, may muddle disease rates in a region with rates increasing in some areas while decreasing in others (Hughes 2016). In addition, the State of Florida is a well-known destination for retired people. Such movement of elderly people can distort the age pyramid of the state, resulting in an impact on adjusted cancer rates.

Second, the RE term comprises SSRE and SURE components; their MCs indicate the existence of weak-to-moderate PSA (e.g., the Miami MSA) or (near)-zero SA (e.g., the Tallahassee MSA) in the SSRE components. However, a decomposition of the SSRE terms explicitly shows that they essentially are mixtures of moderate-to-strong PSA and weak-to-moderate NSA. Griffith and Arbia (2010) utilize a two-SA-parameter spatial simultaneous autoregressive model to uncover a mixture of SA, where the PSA component counterbalances the NSA component. A discovery of SA mixtures has rarely been reported in literature, especially in epidemiology, and its detection can help researchers gain a better understanding of the geographic distribution of, geographic variation of, and risk factors for a disease. As discussed earlier, the moderate-to-strong PSA largely is associated with the geographic distribution of socio-economic phenomenon (e.g., employment status, population migration), whereas the weak-to-moderate NSA likely is linked to mechanisms such as a decrease of lung cancer rates because of increasing cancer screening when lung cancer cases are detected in neighboring places.

This study furnishes motivation for a number of future research efforts. First, a comparison of research outcomes at the county and census tract resolutions reveal a presence of substantial heterogeneity in lung cancer data, and more noise is expected if a spatial analysis is conducted at a finer resolution (e.g., block groups). Thus, extending current research to a finer resolution would be beneficial. Second, a comparison of crude and adjusted lung cancer incidence rates suggests the disappearance of some prominent spatial patterns (e.g., PSA) at both geographic resolutions. However, this observation has rarely been discussed in the literature, and hence a

further examination of rate standardization and/or more similar case studies is necessary. Third, to date, the literature about PSA-NSA mixtures is relatively scant. This study only explores the scenario that a weak-to-moderate PSA or (near)-zero SA can be partitioned into a mixture of moderate-to-strong PSA and weak-to-moderate NSA. Other scenarios (e.g., a global strong PSA; moderate NSA) remain to be investigated. Finally, SA mixtures are discovered in the lung cancer data in Florida. Similar research should be conducted to examine if consistent results would be obtained with different empirical data, or for different study areas.

4.6 Appendix A4. Crude Lung Cancer Incidence Rates Maps

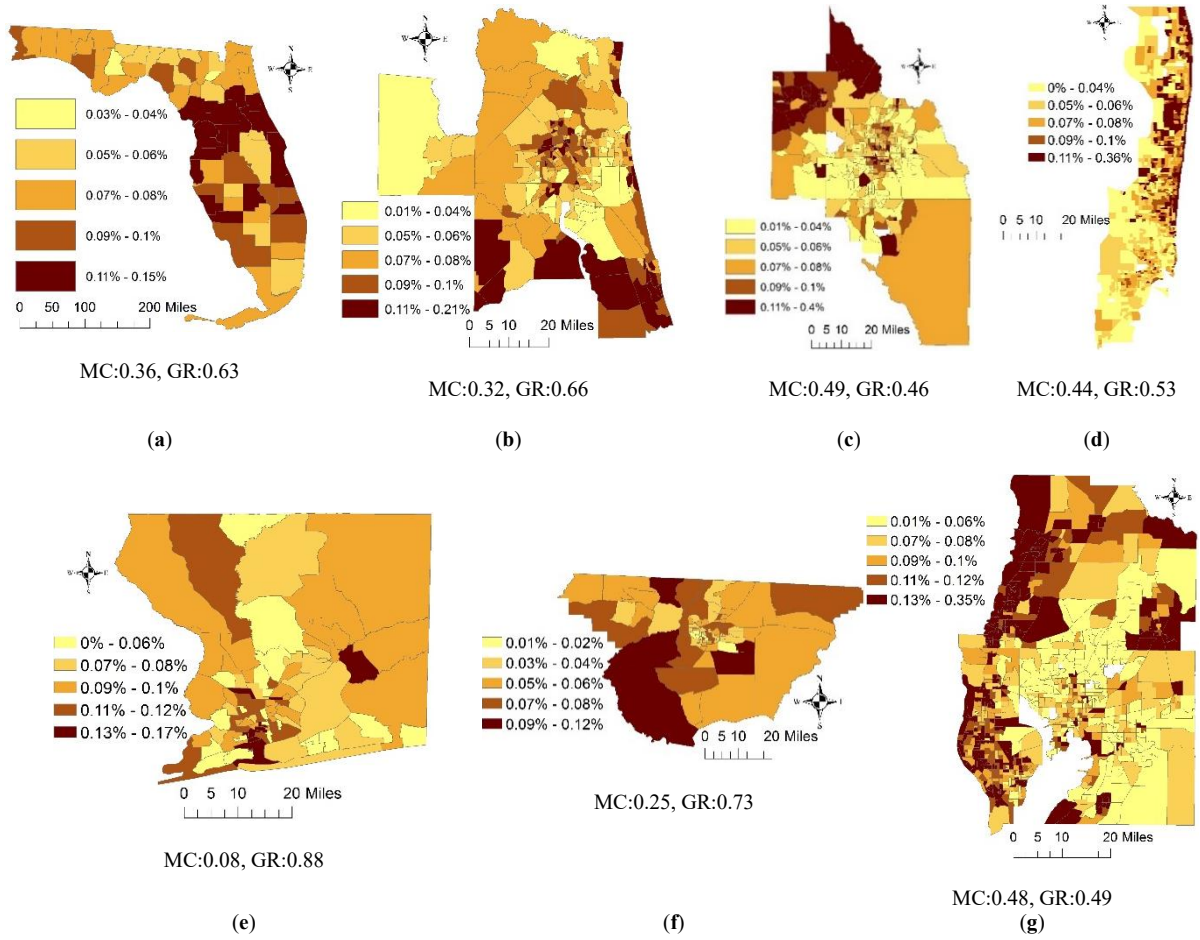


Figure A4.1. The spatial patterns of crude lung cancer incidence rates. (**a**) the State of Florida counties. Census tracts for: (**b**) the Jacksonville MSA; (**c**) the Orlando MSA; (**d**) the Miami MSA; (**e**) the Pensacola MSA; (**f**) the Tallahassee MSA; (**g**) the Tampa MSA.

CHAPTER 5

CONCLUSION

This dissertation contributes to the understanding of negative spatial autocorrelation (SA), which has been largely neglected in the spatial analysis literature. It explores negative SA in two broad ways. First, it investigates the impacts of SA on the distributional characteristics of random variables (RVs), with a consideration of both negative SA and positive SA. This dissertation particularly focuses on the beta and multinomial RVs, which become increasingly important in GISciences but have not drawn a much attention in the literature. Second, this dissertation examines the presence of positive and negative SA mixture in empirical data and evaluates the capability of statistical model specifications in accommodating a mixture pattern of positive SA and negative SA. Detailed summaries and implications are discussed as below for each individual chapter.

The second chapter evaluates the impact of SA upon the histograms of beta and multinomial RVs, considering both positive SA and negative SA. The most critical finding is that positive SA and negative SA behave differently when a RV is skewed Specifically, negative SA fails to converge on its maximum, and the gap between the maximum it achieves and the theoretical maximum becomes more conspicuous as a level of skewness increases. In contrast, positive SA always converges on its maximum and is not impacted by skewness. Second, the simulation results show that a RV mean is unaffected by SA. However, SA inflates variance, with this inflation becoming more pronounced as the SA level increases. Meanwhile, positive SA generally creates more inflation than negative SA. The third and fourth moments: skewness and excess kurtosis, can also be altered by SA. For example, while skewness remains unchanged when beta RV closely conforms on a normal distribution, it deviates from theoretical values when beta RV is positively or negatively skewed.

The third chapter utilizes the Moran eigenvector spatial filtering (MESF) methodology to uncover a SA mixture pattern in breast cancer data for Broward County, FL. One major conclusion that can be drawn from it is that the MESF model specifications improve model

performance. Specifically, the MESF model that accounts for both positive and negative SA in the data is preferred. In contrast, both MESF with positive eigenvectors and Besag-York-Mollié (BYM) models successfully accommodate positive SA, but fail to address negative SA that may be hidden by a globally dominating pattern of positive SA. The analysis results confirm a possible presence of a mixture of positive and negative SA in the age-adjusted cancer rates, although the Moran's *I* suggests positive SA only in the standard Poisson and negative binomial (NB) model residuals, possibly because the positive SA component outweighs the negative SA component. This empirical analysis implies that the MESF model specification with both positive and negative eigenvectors furnishes an efficient method to account for SA mixture components in georeferenced data.

The forth chapter examines the spatial patterns of lung cancer incidence rates at different geographic resolutions and scales in Florida. There are two major contributions of this research. First, the results show that a random effects (RE) model furnishes an efficient method to correct for biased estimation (e.g., underestimated standard errors). The analysis results indicate that an inclusion of a RE term, which can serve as a proxy for omitted variables, improves model performance (e.g., it increases pseudo-R^2 values). Second, the RE term comprises spatially structured RE (SSRE) and spatially unstructured RE (SURE) components; their Moran's *I* statistics indicate the existence of positive SA (e.g., the Miami metropolitan statistical area (MSA)) or (near)-zero SA (e.g., the Tallahassee MSA) in the SSRE components. However, a decomposition of the SSRE terms explicitly shows that they essentially are mixtures of positive and negative SA. The positive SA largely is associated with the geographic distribution of socio-economic phenomenon (e.g., employment status, population migration), whereas the negative SA likely is linked to mechanisms such as a decrease of lung cancer rates because of increasing cancer screening when lung cancer cases are detected in neighboring places.

There are several limitations of this dissertation. First, a beta RV can mimic different RVs, however, this research only investigates three different RVs. For the multinomial RV, scenarios with only three categories are investigated in this study. Second, simulation experiments designed for this dissertation are based on a 30-by-30 square tessellation surface. Different

surface partitioning (e.g., a hexagonal/irregular tessellation) with different sizes and configurations are not evaluated. Third, the spatial analysis of lung cancer rates focuses on six MSAs which are relatively highly densely populated, spatial pattern and heterogeneity of cancer rates may differ in areas less populated. Overall, further theoretical investigations about negative SA and mixtures of positive and negative SA are necessary in order to better understand their impacts on modeling georeferenced data. Because findings reported in this dissertation are based on a simulation experiment and two empirical data analyses, subsequent and further data analyses would be insightful.

REFERENCES

Alberg, A. J., & Samet, J. M. (2003). Epidemiology of lung cancer. *Chest*, 123(1), 21S-49S.

Alberg, A. J., Brock, M. V., & Samet, J. M. (2005). Epidemiology of lung cancer: looking to the future. *Journal of Clinical Oncology*, 23(14), 3175-3185.

Antunes, J. L. F., Biazevic, M. G. H., De Araujo, M. E., Tomita, N. E., Chinellato, L. E. M., & Narvai, P. C. (2001). Trends and spatial distribution of oral cancer mortality in São Paulo, Brazil, 1980–1998. *Oral Oncology*, 37(4), 345-35.

Anderson, R. N., & Rosenberg, H. M. (1998). Age standardization of death rates: implementation of the year 2000 standard. *National vital statistics reports*, 47(3), 1-17.

Ahmad, O. B., Boschi-Pinto, C., Lopez, A. D., Murray, C. J., Lozano, R., & Inoue, M. (2001). Age standardization of rates: a new WHO standard. Geneva: *World Health Organization*, 9.

Baumont, C., Ertur, C., & Gallo, J. (2004). Spatial analysis of employment and population density: the case of the agglomeration of Dijon 1999. *Geographical Analysis*, 36(2), 146-176.

Bentham, G. (1988). Migration and morbidity: implications for geographical studies of disease. *Social science & medicine*, 26(1), 49-54.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1), 1-20.

Blazer, D. G., & Wu, L. T. (2009). The epidemiology of at-risk and binge drinking among middle-aged and elderly community adults: National Survey on Drug Use and Health. *American Journal of Psychiatry*, 166(10), 1162-1169.

Blue, L., & Fenelon, A. (2011). Explaining low mortality among US immigrants relative to native-born Americans: the role of smoking. *International journal of epidemiology*, 40(3), 786-793.

Bolduc, D. (1999). A practical technique to estimate multinomial probit models in transportation. *Transportation Research Part B: Methodological*, 33(1), 63-79.

Boarnet, M. G., & Glazer, A. (2002). Federal grants and yardstick competition. *Journal of urban Economics*, 52(1), 53-64.

Bosdriesz, J. R., Lichthart, N., Witvliet, M. I., Busschers, W. B., Stronks, K., & Kunst, A. E. (2013). Smoking prevalence among migrants in the US compared to the US-born and the population in countries of origin. *PloS one*, 8(3).

Boyle, P., Norman, P., & Rees, P. (2002). Does migration exaggerate the relationship between deprivation and limiting long-term illness? A Scottish analysis. *Social science & medicine*, 55(1), 21-31.

Branscum, A. J., Johnson, W. O., & Thurmond, M. C. (2007). Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease. *Australian & New Zealand Journal of Statistics*, 49(3), 287-301.

Bray, F. (2002). Age-standardization. *Cancer incidence in five continents*, 87-89.

Can, A. (1990). The measurement of neighborhood dynamics in urban house prices. *Economic geography*, 66(3), 254-272.

Carrière, G. M., Sanmartin, C., Bryant, H., & Lockwood, G. (2013). Rates of cancer incidence across terciles of the foreign-born population in Canada from 2001–2006. *Canadian Journal of Public Health*, 104(7), 443-449.

Cepeda-Cuervo, E., & Núñez-Antón, V. (2013). Spatial double generalized beta regression models: extensions and application to study quality of education in Colombia. *Journal of Educational and Behavioral Statistics*, 38(6), 604-628.

Chen, J., Hori, Y., Yamamura, Y., Shiyomi, M., & Huang, D. (2008). Spatial heterogeneity and diversity analysis of macrovegetation in the Xilingol region, Inner Mongolia, China, using the beta distribution. *Journal of arid environments*, 72(6), 1110-1119.

Chen, E., & Tarko, A. P. (2014). Modeling safety of highway work zones with random parameters and random effects models. *Analytic methods in accident research*, 1, 86-95.

Chun, Y. (2008). Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems*, 10(4), 317-344.

Chun, Y., & Griffith, D. A. (2011). Modeling network autocorrelation in space–time migration flow data: an eigenvector spatial filtering approach. *Annals of the Association of American Geographers*, 101(3), 523-536.

Chun, Y. (2014). Analyzing space-time crime incidents using eigenvector spatial filtering: an application to vehicle burglary. *Geographical Analysis*, 46(2), 165-184.

Chun, Y., Griffith, D. A., Lee, M., & Sinha, P. (2016). Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. *Journal of Geographical Systems*, 18(1), 67-85.

Chun, Y and D. A. Griffith. (2018) Impacts of negative spatial autocorrelation on frequency distributions, *Chilean Journal of Statistics*, 9(1), 3-17.

Clarke, Paul; Crawford, Claire; Steele, Fiona; Vignoles, Anna (2010): The choice between fixed and random effects models: Some considerations for educational research, *IZA Discussion Papers*, No. 5287, Institute for the Study of Labor (IZA), Bonn

Clegg, L.; Reichman, M.; Miller, B.; Hankey, B.; Singh, G.; Lin, Y.; Goodman, M.T.; Lynch, C.F.; Schwartz, S.M.; Chen, V.W.; et al. (2009). Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. *Cancer causes & control*, 20(4), 417-435.

Cohen, J. P., & Paul, C. J. M. (2004). Public infrastructure investment, interstate spatial spillovers, and manufacturing costs. *The Review of Economics and Statistics*, 86(2), 551-560.

Conley, T. G., & Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17(4), 303-327.

Cordeiro, R., Donalisio, M. R., Andrade, V. R., Mafra, A. C., Nucci, L. B., Brown, J. C., & Stephan, C. (2011). Spatial distribution of the risk of dengue fever in southeast Brazil, 2006-2007. *BMC Public Health*, 11(1), 355.

Craney, T.; Surles, J.G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14, 391–403.

Cressie, N. (1991). Statistics for Spatial Data. *Wiley*, New York.

Crespo Cuaresma, J., and M. Feldkircher. (2013). "Spatial Filtering, Model Uncertainty and the Speed of Income Convergence in Europe." *Journal of Applied Econometrics* 28(4), 720–41.

Dai, D. (2010). Black residential segregation, disparities in spatial access to health care facilities, and late-stage breast cancer diagnosis in metropolitan Detroit. *Health & Place*, 16(5), 1038-1052.

Debella-Gilo, M., & Etzelmüller, B. (2009). Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. *Catena*, 77(1), 8-18.

Dendoncker, N., Rounsevell, M., & Bogaert, P. (2007). Spatial analysis and modelling of land use distributions in Belgium. Computers, *Environment and Urban Systems*, 31(2), 188-205.

DeSantis, C., Ma, J., Bryan, L., & Jemal, A. (2014). Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 64(1), 52-62.

Diggle, P. J., & Milne, R. K. (1983). Negative binomial quadrat counts and point processes. *Scandinavian journal of statistics*, 257-267.

Do, D. P., Wang, L., & Elliott, M. R. (2013). Investigating the relationship between neighborhood poverty and mortality risk: a marginal structural modeling approach. *Social Science & Medicine*, 91, 58-66.

d'Onofrio, A.; Mazzetta, C.; Robertson, C.; Smans, M.; Boyle, P.; Boniol, M. (2016). Maps and atlases of cancer mortality: A review of a useful tool to trigger new questions. *Ecancermedicalscience*, 10, 670.

Eckey, H. F., C. Dreger, and M. Türck. (2006). European Regional Convergence in a Human Capital Augmented Solow Model (No. 88). *Volkswirtschaftliche Diskussionsbeiträge*.

Elhorst, J. P., & Zigova, K. (2014). Competition in research activity among economic departments: Evidence by negative spatial autocorrelation. *Geographical Analysis*, 46(2), 104-125.

Ferrari, S. and Cribari, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, 31, 799-815.

Feskanich, D.; Ziegler, R.; Michaud, D.; Giovannucci, E.; Speizer, F.; Willett, W.; Colditz, G.A. (2000). Prospective study of fruit and vegetable consumption and risk of lung cancer among men and women. *Journal of the National Cancer Institute*, 92, 1812–1823.

Frondel, M.; Vance, C. (2010). Fixed, random, or something in between. A variant of Hausman's specification test for panel data estimators. *Economic Letters*. 107, 327–329.

Fukuda, Y., Umezaki, M., Nakamura, K., & Takano, T. (2005). Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. *International Journal of Health Geographics*, 4(1), 16.

Gerber, F., & Furrer, R. (2015). Pitfalls in the implementation of Bayesian hierarchical modeling of areal count data: An illustration using BYM and Leroux models. *Journal of Statistical Software*, Code Snippets, 63(1), 1-32.

Gorman, D. M., Speer, P. W., Gruenewald, P. J., & Labouvie, E. W. (2001). Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *Journal of studies on alcohol*, 62(5), 628-636.

Greene, W. H. (2003). Econometric Analysis, (5th ed.), *Upper Saddle River*, New Jersey: Prentice Hall.

Griffith, D. A. (1987). Spatial autocorrelation. *A Primer (Washington, DC, Association of American Geographers)*.

Griffith, D. A. (2002). A spatial filtering specification for the auto-Poisson model. *Statistics and Probability Letters* 58(3), 245–51.

Griffith, D. A. (2003). Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization. *Berlin: Springer*.

Griffith, D. A. (2006). Hidden negative spatial autocorrelation. *Journal of Geographical Systems*, 8(4), 335-355.

Griffith, D. A. (2007). Spatial structure and spatial interaction: 25 years later. *The Review of Regional Studies*, 37(1), 28-38.

Griffith, D. A. (2009). Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 Germany journey-to-work flows. *Journal of Geographical Systems*, 11(2), 117-140.

Griffith, D. A., & Arbia, G. (2010). Detecting negative spatial autocorrelation in georeferenced random variables. *International Journal of Geographical Information Science*, 24(3), 417-437.

Griffith, D. A. (2011). Positive spatial autocorrelation impacts on attribute variable frequency distributions. *Chilean Journal of Statistics*, 2(2), 3-28.

Griffith, D. (2013). Estimating missing data values for georeferenced Poisson counts. *Geographical Analysis*. 45, 259–284.

Griffith, D. A., & Chun, Y. (2016). Spatial autocorrelation and uncertainty associated with remotely-sensed data. *Remote Sensing*, 8(7), 535.

Griffith, D. A. (2017). Some robustness assessments of Moran eigenvector spatial filtering. *Spatial Statistics*, 22, 155-179.

Griffith, D. A. (2019). Negative Spatial Autocorrelation: One of the Most Neglected Concepts in Spatial Statistics. *Stats*, 2(3), 388-415.

Gomez-Rubio V, Bivand RS, Rue H (2014) Spatial models using Laplace approximation methods. *In: Fischer MM, Nijkamp P (eds) Handbook of regional science*. Springer, New York, pp 1401–1417

Guolo, A., and C. Varin. (2014). "Beta Regression for Time Series Analysis of Bounded Data, with Application to Canada Google® Flu Trends." *The Annals of Applied Statistics* 8(1), 74–88.

Gumpertz, M. L., Pickle, L. W., Miller, B. A., & Bell, B. S. (2006). Geographic patterns of advanced breast cancer in Los Angeles: associations with biological and sociodemographic factors (United States). *Cancer Causes & Control*, 17(3), 325-339.

Haiman, C.; Stram, D.; Wilkens, L.; Pike, M.; Kolonel, L.; Henderson, B.; Le Marchand, L. (2006). Ethnic and racial differences in the smoking-related risk of lung cancer. *New England Journal of Medicine*, 354, 333–342.

Haining R (1984) Testing a spatial interacting-markets hypothesis. *Review of Economics and Statistics*, 66(4):576–583

Haining, R., Law, J., & Griffith, D. (2009). Modelling small area counts in the presence of overdispersion and spatial autocorrelation. *Computational Statistics & Data Analysis*, *53*(8), 2923-2937.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in medicine*, 22(9), 1433-1446.

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4), 325-334.

Hu, L., Griffith, D., & Chun, Y. (2019). Impacts of Spatial Autocorrelation in Georeferenced Beta and Multinomial Random Variables. *Geographical Analysis*.

Hu, L., Chun, Y., & Griffith, D. A. (2020). Uncovering a positive and negative spatial autocorrelation mixture pattern: a spatial analysis of breast cancer incidences in Broward County, Florida, 2000–2010. *Journal of Geographical Systems*, 1-18.

Hu, L., Griffith, D., & Chun, Y. (2018). Space-Time Statistical Insights about Geographic Variation in Lung Cancer Incidence Rates: Florida, USA, 2000–2011. *International journal of environmental research and public health*, *15*(11), 2406.

Hughes, A.E. (2016) Residential Mobility and CRC Screening: A Spatial Analysis of CRC Screening in an Urban Safety-Net Clinic; *The University of Texas at Dallas: Dallas, TX, USA.*

Hussain, S. K., Altieri, A., Sundquist, J., & Hemminki, K. (2008). Influence of education level on breast cancer risk and survival in Sweden between 1990 and 2004. *International Journal of Cancer*, 122(1), 165-169.

Jacob, B. G., Griffith, D. A., Mwangangi, J., Gathings, D. A., Mbogo, C. C., & Novak, R. J. (2011). A cartographic analysis using spatial filter logistic model specifications for implementing mosquito control in Kenya. *Urban Geography*, 32(2), 263-300.

Jacquez, G.; Greiling, D.A. (2003). Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York. *International Journal of Health Geographics*, 2, 4.

Jerrett, M.; Burnett, R.; Ma, R.; Pope, C.A., III.; Krewski, D.; Newbold, K.; Thurston, G.; Shi, Y.; Finkelstein, N.; Calle, E.E.; et al. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, 16, 727–736.

Jin, X.; Carlin, B.; Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61, 950–961.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181), 990-993.

Kaiser, M. S., Cressie, N., & Lee, J. (2002). Spatial mixture models based on exponential family conditional distributions. *Statistica Sinica*, 449-474.

Kao, Y.-H. 2016. Three Essays on Spatial Econometrics with an Emphasis on Testing. Ph.D. Thesis, *University of Illinois at Urbana-Champaign, Urbana, IL, USA*. Unpublished doctoral dissertation.

Kao, S. Y. H., & Bera, A. K. (2016). Spatial regression: The curious case of negative spatial dependence. *Urbana-Champaign, Mimeo: University of Illinois*.

Kazembe, L. N., & Namangale, J. J. (2007). A Bayesian multinomial model to analyses spatial patterns of childhood co-morbidity in Malawi. *European journal of epidemiology*, 22(8), 545-556.

Kalhori, L., & Mohhamadzadeh, M. (2017). Spatial Beta Regression Model with Random Effect. *Journal of Statistical Research of Iran JSRI*, 13(2), 215-230.

Kavousi, A., Meshkani, M. R., & Mohammadzadeh, M. (2011). Spatial analysis of auto-multivariate lattice data. *Statistical Papers*, 52(4), 937-952.

Kelsall, J. E., & Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4), 559-573.

Kelsey, J. L., Gammon, M. D., & John, E. M. (1993). Reproductive factors and breast cancer. *Epidemiologic Reviews*, 15(1), 36.

Keitt, T. H., Bjørnstad, O. N., Dixon, P. M., & Citron-Pousty, S. (2002). Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, 25(5), 616-625.

Lawson, A. B. (2013). Statistical methods in spatial epidemiology. *John Wiley & Sons*.

Lawson, A. B. (2013). Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. *Boca Raton, FL: CRC Press*.

Le Gallo, J., & Ertur, C. (2003). Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980−1995. *Papers in Regional Science*, 82(2), 175-201.

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

Lee, D., Rushworth, A., & Sahu, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70(2), 419-429.

Lee, M.; Chun, Y.; Griffith, D. (2018). An evaluation of kernel smoothing to protect confidentiality of patient locations. *International Journal of Urban Sciences*. in press.

Lee, G., O'Leary, J. T., Lee, S. H., & Morrison, A. (2002). Comparison and contrast of push and pull motivational effects on trip behavior: An application of a multinomial logistic regression model. *Tourism Analysis*, 7(2), 89-104.

LeSage, J. P., and R. K. Pace. (2008). "Spatial Econometric Modeling of Origin-Destination Flows." Journal of Regional Science 48(5), 941–67.

Lin, G., & Zhang, T. (2007). Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky breast cancer data. *Geographical Analysis*, 39(3), 293-310.

López-Abente, G., Aragonés, N., García-Pérez, J., & Fernández-Navarro, P. (2014). Disease mapping and spatio-temporal analysis: importance of expected-case computation criteria. *Geospatial health*, 27-35.

MacKinnon, J. A., Duncan, R. C., Huang, Y., Lee, D. J., Fleming, L. E., Voti, L., Rudolph, M., & Wilkinson, J. D. (2007). Detecting an association between socioeconomic status and late stage breast cancer using spatial analysis and area-based measures. *Cancer Epidemiology Biomarkers & Prevention*, 16(4), 756-762.

MacLennan, R.; Da Costa, J.; Day, N.; Law, C.; Ng, Y.; Shanmugaratnam, K. (1977). Risk factors for lung cancer in Singapore Chinese, a population with high female incidence rates. *International journal of cancer*, 20, 854–860.

Mao, Y.; Hu, J.; Ugnat, A.; Semenciw, R.; Fincham, S. (2001). Socioeconomic status and lung cancer risk in Canada. *International journal of epidemiology*. 30, 809–817.

McPherson, K., Steel, C., & Dixon, J. M. (2000). Breast cancer—epidemiology, risk factors, and genetics. *BMJ: British Medical Journal*, 321(7261), 624-628.

McCullagh, P., and J. Nelder. (1989). Generalized Linear Models. *London: Chapman & Hall*.

Meliker, J. R., Jacquez, G. M., Goovaerts, P., Copeland, G., & Yassine, M. (2009). Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data. *Cancer Causes & Control*, 20(7), 1061-1069.

Millington, J. D., Perry, G. L., & Romero-Calcerrada, R. (2007). Regression techniques for examining land use/cover change: a case study of a Mediterranean landscape. *Ecosystems*, 10(4), 562-578.

Molina, J.; Yang, P.; Cassivi, S.; Schild, S.; Adjei, A.A. (2008). Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. In *Mayo Clinic Proceedings*; Elsevier: Amsterdam, The Netherlands; Volume 83, pp. 584–594.

Montgomery, R. A., & Chazdon, R. L. (2001). Forest structure, canopy architecture, and light transmittance in tropical wet forests. *Ecology*, 82(10), 2707-2718.

Muir, K., Rattanamongkolgul, S., Smallman-Raynor, M., Thomas, M., Downer, S., & Jenkinson, C. (2004). Breast cancer incidence and its possible spatial association with pesticide application in two counties of England. *Public Health*, *118*(7), 513-520.

Murray, A. T., I. McGuffog, J. S. Western, and P. Mullins. (2001). "Exploratory Spatial Data Analysis Techniques for Examining Urban Crime: Implications for Evaluating Treatment." *British Journal of criminology* 41(2), 309–29.

Muscat, J.; Richie, J.; Stellman, S.D. (2002). Mentholated cigarettes and smoking habits in whites and blacks. *Tobacco Control*, 11, 368–371.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403(6772), 853-858.

Neyens, T., Faes, C., & Molenberghs, G. (2012). A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and spatio-temporal epidemiology*, 3(3), 185-194.

O'brien, R.M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41, 673–690.

Odoi, A., Martin, S. W., Michel, P., Holt, J., Middleton, D., & Wilson, J. (2003). Geographical and temporal distribution of human giardiasis in Ontario, Canada. *International Journal of Health Geographics*, 2(1), 5.

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques in Modern Geography* No.38. GeoBooks, Norwich, England.

Osler, M. (1993). Social class and health behavior in Danish adults: A longitudinal study. *Public Health*, 107, 251–260.

Pace, R. K., J. P. LeSage, and S. Zhu. (2013). "Interpretation and Computation of Estimates from Regression Models Using Spatial Filtering." *Spatial Economic Analysis* 8(3), 352–69.

Paez, A. (2018). "Using Spatial Filters and Exploratory Data Analysis to Enhance Regression Models of Spatial Data." *Geographical Analysis*, 1–25.

Parkin D.M., Bray F.I., & Devesa S.S. (2001). Cancer burden in the year 2000. The global picture. *European Journal of Cancer*, 37, S4–S66.

Patuelli, R., Griffith, D. A., Tiefelsdorf, M., & Nijkamp, P. (2011). Spatial filtering and eigenvector stability: space-time models for German unemployment data. *International Regional Science Review*, 34(2), 253-280.

Pomerleau, J.; Pederson, L.; Østbye, T.; Speechley, M.; Speechley, K.N. (1997). Health behaviours and socio-economic status in Ontario, Canada. *European journal of epidemiology*, 13, 613–622.

Pope, C.A., III, Burnett, R.; Thun, M.; Calle, E.; Krewski, D.; Ito, K.; Thurston, G.D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, 287, 1132–1141.

Richardson, S.; Abellan, J.; Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical methods in medical research*, 15, 385–407.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4), 1145-1165.

Risch, H.A.; Howe, G.R.; Jain, M.; Burch, J.D.; Holowaty, E.J.; Miller, A.B. (1993). Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histologic type. *American journal of epidemiology*, 138, 281–293.

Robert, S. A., Trentham-Dietz, A., Hampton, J. M., McElroy, J. A., Newcomb, P. A., & Remington, P. L. (2004). Socioeconomic risk factors for breast cancer: distinguishing individual-and community-level effects. *Epidemiology*, 15(4), 442-450.

Roquette, R.; Nunes, B.; Painho, M. (2018). The relevance of spatial aggregation level and of applied methods in the analysis of geographical distribution of cancer mortality in mainland Portugal (2009–2013). *Population health metrics*, 16, 6, doi:10.1186/s12963-018-0164-6.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society*: Series b (statistical methodology), 71(2), 319-392.

Sheehan, T. J., DeChello, L. M., Kulldorff, M., Gregorio, D. I., Gershman, S., & Mroszczyk, M. (2004). The geographic distribution of breast cancer incidence in Massachusetts 1988 to 1997, adjusted for covariates. *International Journal of Health Geographics*, 3(1), 17.

Shiyomi, M., Takahashi, S., & Yoshimura, J. (2000). A measure for spatial heterogeneity of a grassland vegetation based on the beta-binomial distribution. *Journal of Vegetation Science*, 11(5), 627-632.

Singh, G.; Miller, B.A. (2004). Health, life expectancy, and mortality patterns among immigrant populations in the United States. *Canadian journal of public health*, 95, 14–21.

Sinha, P. (2017). Modeling Land use change using an eigenvector spatial filtering model specification for discrete responses. *In Advances in Geocomputation* (pp. 335-344). Springer, Cham.

Smith, H.; Seal, S.; Sullivan, D. (2017). Impact of race, poverty, insurance coverage and resource availability on breast cancer across geographic regions of Mississippi. *Journal of the Mississippi Academy of Sciences*., 62, 353–369.

Stellman, S.; Chen, Y.; Muscat, J.; Djordjevic, I.; Richie, J.R.; Lazarus, P.; Thompson, S.; Altorki, N.; Berwick, M.; Citron, M.L.; et al. (2003). Lung cancer risk in white and black Americans. *Annals of epidemiology*., 13, 294–302.

Tian, N., Wilson, J., & Zhan, F. (2011). Spatial association of racial/ethnic disparities between late-stage diagnosis and mortality for female breast cancer: where to intervene? *International Journal of Health Geographics*, 10(1), 24.

Tiefelsdorf, M., and B. Boots. (1995). The exact distribution of Moran's I. *Environment and Planning A*, 27, 985-999.

Tiefelsdorf, M., & Griffith, D. A. (2007). Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A*, 39(5), 1193-1221.

Timander, L. M., & McLafferty, S. (1998). Breast cancer in West Islip, NY: a spatial clustering analysis with covariates. *Social Science & Medicine*, 46(12), 1623-1635.

Torabi, M., & Rosychuk, R. J. (2012). Hierarchical Bayesian spatiotemporal analysis of childhood cancer trends. *Geographical Analysis*, 44(2), 109-120.

Verbeke, G.; Molenberghs, G.; Rizopoulos, D. (2010). Random effects models for longitudinal data. In *Longitudinal Research with Latent Variables*, Springer: Berlin/Heidelberg, Germany,; pp. 37–96.

Verburg, P. H., van Eck, J. R. R., de Nijs, T. C., Dijst, M. J., & Schot, P. (2004). Determinants of land-use change patterns in the Netherlands. *Environment and Planning B: Planning and Design*, 31(1), 125-150.

Vieira, V. M., Webster, T. F., Weinberg, J. M., & Aschengrau, A. (2008). Spatial-temporal analysis of breast cancer in upper Cape Cod, Massachusetts. *International Journal of Health Geographics*, 7(1), 46.

Vineis, P.; Forastiere, F.; Hoek, G.; Lipsett, M. (2004). Outdoor air pollution and lung cancer: Recent epidemiologic evidence. *International Journal of Cancer*, 111, 647–652.

Wang, F., Guo, D., & McLafferty, S. (2012). Constructing geographic areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois. *Applied Geography*, 35(1), 1-11.

Wang, N.; Mengersen, K.; Kimlin, M.; Zhou, M.; Tong, S.; Fang, L.; Wang, B.; Hu, W. (2018). Lung cancer and particulate pollution: A critical review of spatial and temporal analysis evidence. *Environmental research*, 164, 585–596.

Ward, E.; Jemal, A.; Cokkinides, V.; Singh, G.; Cardinez, C.; Ghafoor, A.; Thun, M. (2004). Cancer disparities by race/ethnicity and socioeconomic status. *CA: a cancer journal for clinicians*, 54, 78–93.

Wieland, S. C., Cassa, C. A., Mandl, K. D., & Berger, B. (2008). Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences*, 105(46), 17608-17613.

Yang, X. R., Chang-Claude, J., Goode, E. L., Couch, F. J., Nevanlinna, H., Milne, R. L, Gaudet, M., Schmidt, M. K., Broeks, A., Cox, A., & Fasching, P.A. (2011). Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. *Journal of the National Cancer Institute*, 103(3), 250-263.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.

Zang, E.A.; Wynder, E.L. (1996). Differences in lung cancer risk between men and women: Examination of the evidence. *Journal of the National Cancer Institute*, 88, 183–192.

Zhou, H. B., Liu, S. Y., Lei, L., Chen, Z. W., Peng, J., Yang, Y. Z., & Liu, X. L. (2015). Spatio-temporal analysis of female breast cancer incidence in Shenzhen, 2007–2012. *Chinese Journal of Cancer*, 34(3),13.

BIOGRAPHICAL SKETCH

Lan Hu earned her bachelor's degree in urban planning at the China University of Mining and Technology (Beijing) in 2010. She came to the USA to continue her study and earned a master's degree in Applied Geography from University of North Texas in 2013. She worked as a GIS analyst for an engineering company after graduation, and then she went back to China in 2015 and worked as a research assistant in China Academy of Science in Beijing, where she published her first academic paper. She came back to the USA in 2016 for her PhD at The University of Texas at Dallas, where she worked as a teaching assistant while studying. During the four years' PhD period, she published four papers with her advisors and has one more under review.

CURRICULUM VITAE

EMPLOYMENT

GIS intern, 2M Research Services	12/2019-Current
Lecturer (Statistics), University of Texas at Dallas	08/2019-12/2019
Teaching Assistant, University of Texas at Dallas	08/2016-08/2019
Research Assistant, China Academy of Science	05/2015-12/2015
GIS Analyst, New York Air Brake	06/2014-02/2015
Data/Operation Analyst, Diversegy, LLC	06/2013-06/2014
Lectuer/Teaching Assistant, University of North Texas	09/2010-05/2013

EDUCATION

PhD, Geospatial Information Sciences
The University of Texas at Dallas, TX, USA 08/2016-05/2020

Master of Science, Applied Geography
University of North Texas, Denton, TX, USA 09/2010-06/2013

Bachelor of Science, Urban Planning
China University of Mining & Technology, Beijing, China 09/2006-06/2010

PUBLICATIONS

Journal Articles

- Xu, X., Cai, H., Sun, D., Hu, L., & Banson, K. E. (2016). Impacts of Mining and Urbanization on the Qin-Ba Mountainous Environment, China. *Sustainability*, 8(5), 488.

- Hu, L., Griffith, D. A, & Chun, Y. (2018). Space-Time Statistical Insights about Geographic Variation in Lung Cancer Incidence Rates: Florida, USA, 2000–2011. *International journal of environmental research and public health*, 15(11), 2406.

- Hu, L., Griffith, D. A., & Chun, Y. (2019). Impacts of Spatial Autocorrelation in Georeferenced Beta and Multinomial Random Variables. *Geographical Analysis*.

- Hu, L., Chun, Y., & Griffith, D. A. (2019). A Multilevel Eigenvector Spatial Filtering Model of House Prices: A Case Study of House Sales in Fairfax County, Virginia. *ISPRS International Journal of Geo-Information*, 8(11), 508.

- Hu, L., Chun, Y., & Griffith, D. A. (2020). Uncovering a positive and negative spatial autocorrelation mixture pattern: a spatial analysis of breast cancer incidences in Broward County, Florida, 2000–2010. *Journal of Geographical Systems*, 1-18.

Book Chapters

- Lyons, D., Rice, M., & Hu, L. (2015). Industrial waste management improvement: a case study of Pennsylvania. *International Perspectives on Industrial Ecology*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar, 108-125.

Paper Under Review

- Hu, L., Chun, Y., & Griffith, D. A. Incorporating spatial autocorrelation into house sale price prediction using a random forest model

CONFERENCE PRESENTATIONS

- Hu. L & Lyons. D, 2011. Material Cycling: geographic destination and treatment strategies of Pennsylvania's industrial wastes from 1992 to 2008. Presented at the *Southwest Division of the Association of American Geographers*, Austin, TX, November 11.

- Hu, L., Chun, Y., & Griffith, D. A, 2017. A spatial analysis of 2000-2010 breast cancer cases in Duval County, FL: model specification issues. Presented at the *American Association of Geographers*, Boston, MA, April 4.

- Hu, L., Griffith, D. A., & Chun, Y, 2018. Impacts of Negative Spatial Autocorrelation on Exponential and Beta Random Variables. Presented at the *American Association of Geographers*, New Orleans, LA, April 11.

- Hu, L., Chun, Y., & Griffith, D. A, 2018. Small Area Estimation of Cancer Rates: A Case Study of Lung Cancer in Florida, 2000-2010. Presented at the *Spatial Accuracy*, Beijing, China, May 22.

- Hu, L., Griffith, D. A., & Chun, Y, 2019. Space-Time Statistical Insights about Geographic Variation in Lung Cancer Incidence Rates: Florida, USA, 2000–2011. Presented at the *American Association of Geographers*, Washington DC, April 4.

- Hu, L., Griffith, D. A., & Chun, Y, 2019. A multilevel eigenvector spatial filtering model of house prices: an application to the house sales in Fairfax County, Virginia. Presented at the *Southwest of American Association of Geographers*, Fort Worth, TX, October 11.

TEACHING EXPERIENCE

Lecturer

Earth Science (GEOG 1710), 08/2010-05/2013
Department of Geography and Environment,
University of North Texas

Methods of Quantitative Analysis (EPPS 2302), 08/2019-current
School of Economics, Public & Policy Science (EPPS),
The University of Texas at Dallas

Teaching assistant

Department of Geography and Environment, University of North Texas, 08/2010-05/2013

- GEOG 2170 Culture, Environment and Society
- GEOG 1200 World Regional Geography

School of EPPS, The University of Texas at Dallas, 08/2016-08/2019

- GISC 6381 Geographic Information Science Fundamental
- GISC 4326 Cartography and Geovisualization
- GISC 6388 Advanced GIS Programming
- GISC 6301 GIS Data Analysis Fundamental
- GISC 7310 Advanced GIS Data Analysis
- GISC 7360 GIS Pattern Analysis

AWARDS

Pioneer Student Research Grants – 2016, 2017, 2018
PhD Research Small Award -2019

PROFESSIONAL MEMBERSHIP

Association of American Geographers (AAG) – 2016 - present