Neural, Parallel, and Scientific Computations 25 (2017) 325-334

SPATIAL MODELING TECHNIQUES FOR LATTICE DATA

MITRA LAL DEVKOTA[1] AND GARY HATFIELD[2]

[1]Department of Management and Marketing, University of North Georgia,
Dahlonega, GA 30597, USA
[2]Department of Mathematics and Statistics, South Dakota State University,
Brookings, SD 57007, USA

ABSTRACT. Spatial Modeling Techniques for Lattice Data were discussed. In addition to Ordinary least squares, a conventional method of modeling spatial data; various types of spatial regression techniques, such as Simultaneous Autoregressive (SAR), Conditional Autoregressive (CAR), Generalized Least Squares (GLS), and Geographically Weighted Regression (GWR) were discussed. Comparative studies of these modeling techniques were carried out using a real world dataset and an artificially generated spatial dataset.The results showed that GWR was more suitable for the purpose of incorporating spatial autocorrelation of the data and assessing the local parameter estimates of the model.

AMS (MOS) Subject Classification. 39A10.

1. INTRODUCTION

Traditional modeling techniques used in spatial data analysis are based on standard regression techniques which assume that the observations in the data are independent. These techniques are based on the implicit assumption that the observations are independent to one another, a condition highly unlikely to occur in spatiotemporal phenomena. This technique violates the fundamental principle of geography, well known as Tobler's First Law of Geography (Tobler 1979) which states that everything is related to everything else, but near things are more closely related than the distant things. In spatially dependent data, the value of a variable at a location is dependent on the value of that variable in a neighboring location. This causes the errors of the regression model to be spatially autocorrelated (Wall, 2004). When the assumption of independence is invalid, the effects of autocorrelated predictors tend to be exaggerated (Gumpertz et al. 1997). Due to this, the degree of correlation will be higher than it should be, p-values will be significant when they are actually not, and coefficient of determination (R^2) will be higher than it should be.

Received February 1, 2017

The response in case of spatiotemporal / spatially autocorrelated data are usually modeled by autoregressive models (Simultaneous Autoregressive (SAR) and Conditional Autoregressive (CAR)). Since their first formulation, which is usually credited to Whittle (1954) and Besag (1974), these models have been used extensively in many fields of scientific research. The SAR model is preferred in likelihood inference, while the CAR model is more common in Bayesian inference as a prior distribution for spatially structured random effects. In the study of covariance structure implied by these models, Wall (2004) concluded that the implied correlation between a pair of neighboring areas is negatively associated with the number of neighbors of each region. She also showed that the association is not simple and much variability still remains unexplained.

Though these models were first developed for the analysis of regular lattice data, occurring for instance in agricultural field experiments or when decomposing an image in pixels, they are being used these days for irregular lattice data as well. For the analysis of lattice data, CAR and the SAR models are analogous to the stationary autoregressive time series model defined on the integers. i.e., SAR and CAR models are analogous in functional form and Markov property, respectively (Cressie, 1993). Despite the popularity of these models, some of their properties are not completely understood (see Wall, 2004).

Some research studies have followed a completely different way to consider the spatial pattern. These studies have used to model the response as a function of geographic coordinates (Miller et al. 2007). Pereira and Itami (1991) modeled a trend surface to the geographic coordinates and combined this information with a regression model using environmental predictor variables (Le Duc et al. 1992; Lichstein et al., 2002). While inclusion of geographic coordinates as predictors greatly improves the model accuracy, this effect should be referred to as geographic dependence, rather than spatial dependence (Miller et al. 2007). Besag (1972) suggested an autologistic model (a logistic model which incorporates spatial dependence in binary spatial data) to model a binary response. Dennis et al. (2002) used an autologistic model to model the presence of butterfly species and found that neighborhood models were more efficient than the models that used geographic coordinates. Tognelli and Kelt (2004) compared ordinary least squares regression model with CAR and SAR model, and found that the CAR and SAR models were more efficient in achieving a better fit of the model and that relative importance of the predictor variables shifted in such models.

The main purpose of this research work is to carry out a brief comparative study of different modeling techniques for spatial data. The results of the CAR and SAR models will be compared with the model fitted using Ordinary Least Squares (OLS),

Generalized Least Squares (GLS), and finally with Geographically Weighted Regression (GWR). In section 2, the different models such as CAR, SAR, OLS, GLS, and GWR are defined. Section 3 presents a comparative study of these models using a real-world data set with crop residue yield potential, temperature, and precipitation of two states of the North-Central region (Illinois and Indiana) of the USA followed by an artificially generated spatial data set. Results of these modeling approaches and discussion are in section 4 and summary of the results are in section 5.

2. DESCRIPTION OF MODELS

2.1. CONDITIONAL AUTOREGRESSIVE (CAR) MODEL.

As defined by Wall (2004), the Conditional Autoregressive (CAR) model is specified through a set of conditional distributions

$$y_i|y_j : j \neq i \sim N(\mu_i + \sum_{j=1}^{n} c_{ij}(y_j - \mu_j), \tau_i^2) \tag{2.1}$$

where $E(y_i) = \mu_i, \tau_i^2$ is the conditional variance, c_{ij} are known or unknown constants with $c_{ii} = 0$ for all $i = 1, 2, ..., n$. Then, for finite n, we form $\boldsymbol{C} = (c_{ij})$ and by the factorization theorem (Besag, 1974), $\boldsymbol{Y}$ has a multivariate normal distribution given by

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}) \tag{2.2}$$

where

$$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)^T \tag{2.3}$$

and

$$\boldsymbol{M} = diag(\tau_i^2) \tag{2.4}$$

for all $i = 1, 2, ..., n$. Also, $\boldsymbol{I}$ is an $n \times n$ identity matrix. For a CAR model to be well defined, we require $\boldsymbol{I} - \boldsymbol{C}$ to be nonsingular.

2.2. SIMULTANEOUS AUTOREGRESSIVE (SAR) MODEL.

Wall (2004) defines a SAR model by simultaneous equations

$$y_i = \mu_i + \sum_{j=1}^{n} s_{ij}(y_j - \mu_j) + \epsilon_i \tag{2.5}$$

where $\epsilon = (\epsilon_1, \epsilon_2, .., \epsilon_n)^T \sim N(0, \boldsymbol{D})$ with $\boldsymbol{D}$ diagonal, $E(y_i) = \mu_i$, and s_{ij} are known or unknown constants with $s_{ii} = 0$ for all $i = 1, 2, ..., n$. This model is called simultaneous because the random variables are simultaneously determined by the n equations in (2.5). If n is finite, we form $\boldsymbol{S} = (s_{ij})$. The joint distribution of $\boldsymbol{Y} = (y_1, y_2, .., y_n)^T$ is given by,

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2(\boldsymbol{I} - \boldsymbol{S})^{-1}(\boldsymbol{I} - \boldsymbol{S'})^{-1}) \tag{2.6}$$

It is important to note that for the SAR model to be defined, we require $(\boldsymbol{I} - \boldsymbol{S})$ to be nonsingular (Bivand et al. 2008 and Schabenberger and Gotway 2004).

2.3. **ORDINARY LEAST SQUARES (OLS) MODEL.** In the general linear model

$$Y = X\beta + \epsilon, \tag{2.7}$$

Y is an $n \times 1$ vector of dependent or response variables, X is the design matrix of independent (explanatory) variables, which includes a column of 1s for the intercept, β is the vector of regression coefficients and ϵ is a random vector of residuals whose distribution is $N(0, \sigma^2)$ (Kutner, 2004). The residuals are assumed to be independently and identically distributed with mean 0 and the constant variance, σ^2 .

The maximum likelihood estimate of β is given by

$$\hat{\beta} = (X^T X)^{-1}(X^T Y) \tag{2.8}$$

2.4. **GENERALIZED LEAST SQUARES (GLS) MODEL.** In generalized least squares (GLS) model (Pinhiero and Bates, 2000, pp. 204), the residuals are considered themselves a random variable which has a covariance structure given by,

$$Y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2\Sigma) \tag{2.9}$$

where Σ is a variance covariance matrix of the model residuals which is also a positive definite matrix. Here, the error variances are considered random, and hence the random effect σ represents both the spatial structure of the residuals from the fixed effects model, and the unexplainable noise. The solution to (2.9) is found by least squares, which is given by,

$$\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}Y) \tag{2.10}$$

The gls (Generalized Least Squares) function of the nlme library in R was used to estimate model parameters.

2.5. **GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) MODEL.** As described above, standard regression techniques model the data assuming that the errors are independent. On the other hand, the spatial modeling techniques such as CAR, SAR, and GLS address the local nature of the data by explicitly modeling the covariance structure of the error terms. These later techniques assume spatial stationarity and are location independent; and that the results are eventually equations with global parameter estimates, in that the relationships they describe between the response and the predictor variables are constant throughout the region of interest

(Miller et al. 2007). Another possibility could be to fit a Geographically Weighted Regression (GWR) model (Brunsdon et al. 1999) which addresses the issue of spatial stationarity directly by allowing the relationships to vary over space so that parameter estimates of the regression model do not need to be the same everywhere. As defined by Brunsdon et al. 1999, GWR is an extension of OLS regression model given by

$$y_i = \sum_k X_{ik}\beta_k + \epsilon_i \tag{2.11}$$

by allowing the regression parameter estimates to vary over space as given by

$$y_i = \sum_k X_{ik}\beta_k(u_i, v_i) + \epsilon_i \tag{2.12}$$

where (u_i, v_i) are the geographic coordinates of the i th observation in space. The vectors of estimated coefficients for GWR models are given by

$$\hat{\beta}_i = (X^T W_i X)^{-1}(X^T W_i Y) \tag{2.13}$$

where W_i represents the square matrix of weights relative to the position i. Detailed explanations about the shape, bandwidth, and the functional form of the spatial kernel are given in Fotheringham et al. 2002. The spgwr package (Bivand and Yu 2013) for GWR in the statistical software package R, version 3.0.1 (R Development core team 2013) was used to estimate model parameters. We have chosen adaptive kernel bandwidth. The kernel bandwidth is estimated using cross validation technique, and used a Gaussian distance decay function. GWR is often referred to as a local regression model as it provides parameter estimates to local statistics and hence it is more appropriate when the relationships vary spatially (Fotheringham et al. 2002). Tulbure et al. 2011, Brown et al. 2012 have shown that GWR performs better than OLS regression model in modeling spatially varying data. Pez et al. 2011 mention that GWR is not recommended in situations with small sample sizes ($n \approx 160$) in their experiments), so it is important to note that caution needs to be exercised using GWR for the purpose of assessing spatial heterogeneity of individual parameters when analyzing data sets with small samples. Jetz et al. 2005 suggest that GWR should not be used instead of OLS regression, but rather as a supplement to OLS regression. Bivand et al. 2008 emphasize that GWR is an exploratory technique mainly intended to specify the location where nonstationarity is taking place on the map.

3. APPLICATION

3.1. **DATA.** The real world data set was acquired from the USDA National Agricultural Statistics Service (USDA-NASS, 2009). Two states, Illinois and Indiana, were selected for this comparative study. The data set consists of the variables: County

names, FIPS ID of the county, crop residue yield potential, temperature, and precipitation for the years 1970-2008. For this analysis, a subset of the data for the year 2008 was considered. The dry crop residue yield potential (Mg ha-1) was calculated using crop yield data collected from the USDA-NASS, 2009; temperature was the mean temperature ° C during crop growing seasons (April-October), and precipitation was the annual precipitation (mm) derived from the monthly gridded PRISM weather data. The crop residue yield potential was considered as a dependent variable and the climate variables temperature and precipitation were considered as independent variables. The original version of the data set with crop residue yield potential, temperature and precipitation for the years 1970 to 2008, has already been analyzed by averaging over time and then Conditional Autoregressive and Simultaneous Autoregressive models (Schabenberger and Gotway 2004) were fit (Chintala et al., 2014) for modeling crop residue yield potential. The second data set was an artificially generated spatial data set for 400 locations.

3.2. MODELING CROP RESIDUE YIELD POTENTIAL OF TWO STATES (ILLINOIS AND INDIANA) OF THE US.

CAR, SAR, OLS, GLS, and GWR regression models were fit for the crop residue yield potential as a function of two climate variables (temperature and precipitation) for 194 counties of the two states Illinois and Indiana. We compared the performance of the models using Akaike Information Criterion (AIC) and Moran's I. The conventional way of interpreting AIC is that the smaller the AIC, the better the performance of the model in explaining the response. Moran's I measures the degree of spatial autocorrelation of the regression residuals. Global Moran's I statistic was calculated to test the presence of spatial autocorrelation of the residuals under all the models. Its value ranges from -1 to 1 where the values of 1, 0, and -1 respectively indicate perfect positive spatial autocorrelation, no spatial autocorrelation, and perfect negative spatial autocorrelation.

4. RESULTS AND DISCUSSION

The results from the following table (corresponding to the analysis of real world data set) show that the GWR model has the smallest value of AIC. The table also shows that residuals under CAR, SAR, and GWR models are negatively spatially autocorrelated while those under the OLS and GLS models are positively spatially autocorrelated. On comparison, the residuals under SAR and GWR models exhibit poor spatial autocorrelation while those under the rest of the models exhibit moderate spatial autocorrelation. Similarly, from the output of the analysis of artificially generated spatial data, we see that AIC values for CAR, SAR, and GWR are very close to each other. The Moran's I for residuals of CAR, SAR, and GWR models are smaller compared to the other two models. It is very important to note that GWR

incorporates the spatial autocorrelation and also addresses the local nature of the data, and, hence, we can conclude from these data analyses that the GWR model is the most suitable among these five regression models.

Table 1. Comparison of models using AIC and Moran's I for residuals

Models	Real World Data Set		Generated Data Set	
	AIC	Moran's I	AIC	Moran's I
OLS	629.20	0.35	−124.4118	0.196614762
CAR	577.71	−0.23	−151.4060	−0.138195393
SAR	583.80	−0.04	−152.3936	0.004871265
GLS	461.23	0.50	−137.1612	0.196724377
GWR	448.72	−0.0034	−152.3911	0.159412444

It should be noted that CAR requires a symmetric matrix of spatial weights (Schabenberger and Gotway, 2004) and models autocorrelation within the local neighborhood. CAR model is more suitable for large raster datasets. It is also used as a prior in hierarchical Bayesian modeling of spatial data. On the other hand, SAR does not require a symmetric matrix of spatial weights and models a gradual decay in autocorrelation across multiple neighborhoods. It is widely used method of analyzing irregular lattices. Maps showing spatial autocorrelation of residuals under different models (Figure 1) for an artificial data indicate that there is not much difference in the spatial pattern of the CAR, SAR, and GWR models. The map of the GLS residuals shows that there is somewhat stronger positive autocorrelation of the residuals.

Traditional modeling techniques used in spatially varying data assume that the observations in the data are independent to one another, a condition unlikely to occur in spatial data. When untreated, the spatial dependency of the data can create underestimated standard errors, resulting in Type I errors (Legendre 1993; Legendre and Legendre 1998; Legendre et al. 2002). The consequences of such spatial dependency also could be that correlation coefficients and coefficients of determination of the regression model will appear higher than they really are. In addition, the standard errors might appear smaller than they should be which gives us the false impression that the spatial predictions we make are better than they really are. The another problem caused by such dependency of spatial data is that each observation provides less information about the data and the degrees of freedom used in analyses are exaggerated (Miller et al. 2007).

5. SUMMARY

In this paper, we briefly reviewed different modeling techniques for spatial data and provided a brief review of the literature. It presented a comparative study of OLS,

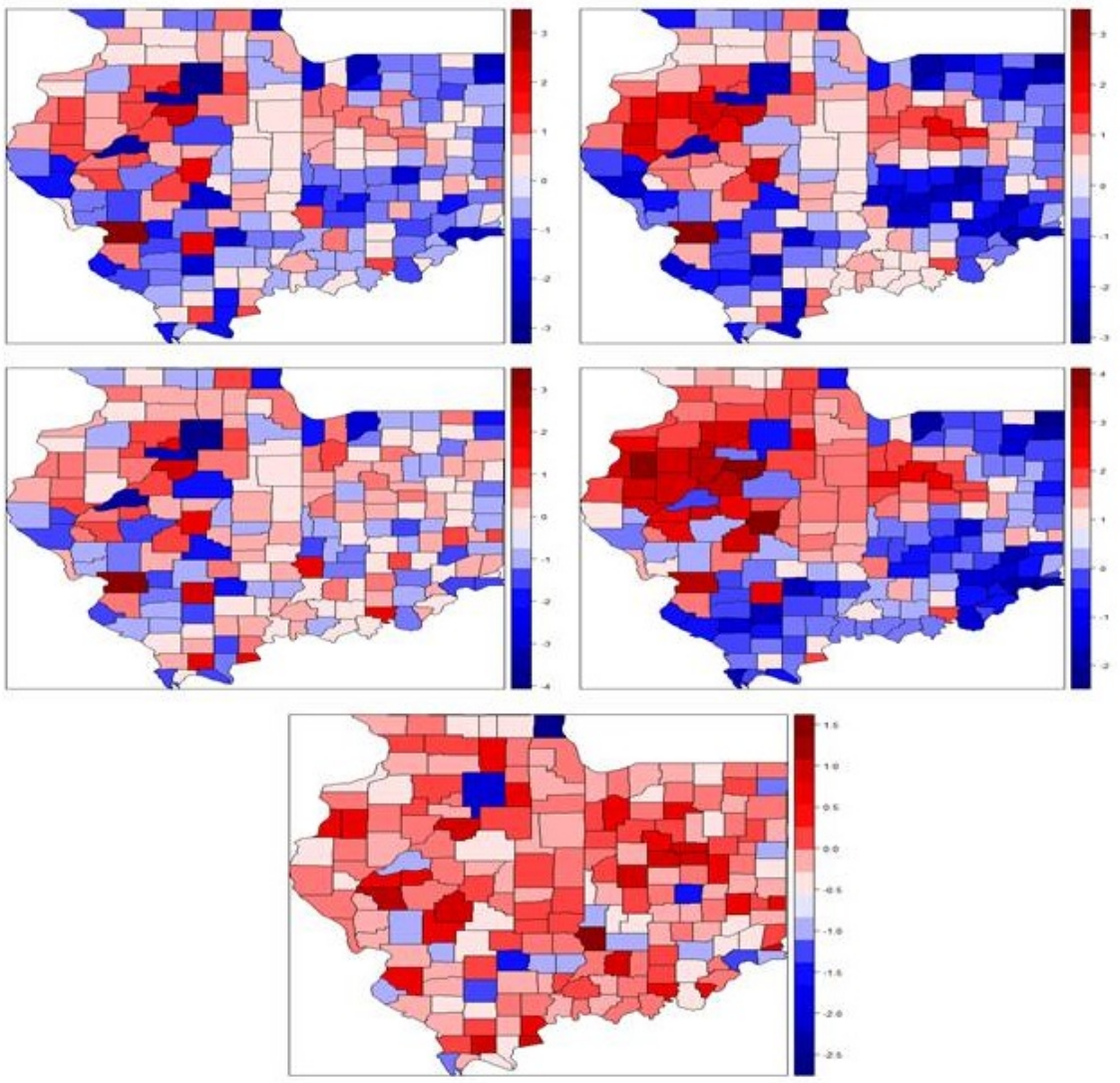

FIGURE 1. Plots showing spatial autocorrelation of residuals under different models. *From left to right: Residuals under OLS, CAR, SAR, GLS and GWR models*

CAR, SAR, GWR, and GLS models using a real-world data set with crop residue yield potential, temperature, and precipitation of two states of the North-Central region (Illinois and Indiana) of the USA followed by an artificially generated data set. The results from these data analyses showed that GWR was more suitable for the purpose of incorporating spatial autocorrelation of the data and assessing the local parameter estimates of the model.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that assisted greatly in improving the quality of the paper.

REFERENCES

[1] Assuno, R., & Krainski, E. (2009). *Neighborhood dependence in Bayesian spatial models.* Biometrical Journal, 51(5), 851-869.

[2] Besag, J. (1972). *Nearest-neighbour systems and the autologistic model for binary data.* J. Roy. Stat. Soc. B 34, 7583.

[3] Besag, J. (1974). *Spatial interaction and the statistical analysis of lattice systems.* Journal of the Royal Statistical Society. Series B (Methodological), 192-236.

[4] Bivand, R. S., Pebesma, E. J., & Rubio, V. G. (2008). *Applied spatial data analysis with R.* Springer.

[5] Bivand, R. (2013). *spdep: Spatial dependence: weighting schemes, statistics and models.* R package version 0.5-60. http://CRAN.R-project.org/package=spdep

[6] Bivand, R. and Yu, D. (2013). *spgwr: Geographically weighted regression.* R package version 0.6-22. http://CRAN.R-project.org/package=spgwr

[7] Brown, S., Versace, V. L., Laurenson, L., Ierodiaconou, D., Fawcett, J., & Salzman, S. (2012). *Assessment of spatiotemporal varying relationships between rainfall, land cover and surface water area using geographically weighted regression.* Environmental Modeling & Assessment, 17(3), 241-254.

[8] Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1999). *Some notes on parametric significance tests for geographically weighted regression.* Journal of Regional Science, 39(3), 497-524.

[9] Chintala, R., Djira, G. D., Devkota, M. L., Prasad, R., & Kumar, S. (2014). *Modeling the effect of temperature and precipitation on crop residue potential for the North Central Region of the United States.* Agricultural Research, 3(2), 148-154.

[10] Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

[11] Dennis, R.L.H., Shreeve, T.G., Sparks, T.H., Lhonore, J.E. (2002). *A comparison of geographical and neighbourhood models for improving atlas databases.* The case of the French butterfly atlas. Biol. Conserv., 108.

[12] Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression.* New York: Wiley.

[13] Gumpertz, M. L., Graham, J. M., & Ristaino, J. B. (1997). *Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: effects of soil variables on disease presence.* Journal of Agricultural, Biological, and Environmental Statistics, 131-156.

[14] Haining, R. (1993).*Spatial data analysis in the social and environmental sciences.* Cambridge University Press.

[15] Jetz, W., Rahbek, C., & Lichstein, J. W. (2005). *Local and global approaches to spatial data analysis in ecology.* Global Ecology and Biogeography, 14(1), 97-98.

[16] Kutner, Michael H., Chris Nachtsheim, and John Neter (2004). *Applied linear regression models.* McGraw-Hill/Irwin

[17] Le Duc, M., Hill, M., Sparks, T., 1992. *A method for predicting the probability of species occurrence using data from systematic surveys.* Watsonia 19, 97105.

[18] Legendre, P. (1993). *Spatial autocorrelation: problem or new paradigm?* Ecology 74, 16591673.

[19] Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M., Myers, D.(2002). *The consequences of spatial structure for the design and analysis of ecological field surveys.* Ecography 25, 601615.

[20] Legendre, P., Legendre, L. (1998). *Numerical Ecology, 2nd English ed.* Elsevier, Amsterdam.

[21] Lichstein, J. W., Simons, T. R., Shriner, S. A., & Franzreb, K. E. (2002). *Spatial autocorrelation and autoregressive models in ecology.* Ecological monographs, 72(3), 445-463.

[22] Miller, J., Franklin, J., & Aspinall, R. (2007). *Incorporating spatial dependence in predictive vegetation models.* Ecological Modeling, 202(3), 225-242.

[23] Pez, A., Farber, S., & Wheeler, D. (2011). *A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships.* Environment and Planning-Part A, 43(12), 2992.

[24] Pereira, J., Itami, R. (1991). *GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham Red Squirrel.* Photogr. Eng. Remote Sensing 57, 14751486.

[25] Pinheiro, Jos C., and Douglas M. Bates (2000). *Linear mixed-effects models: basic concepts and examples.* Springer, New York.

[26] Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R.* CRC Press.

[27] R Development Core Team (2013) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria

[28] Schabenberger, O., & Gotway, C. A. (2004). *Statistical methods for spatial data analysis.* CRC Press.

[29] Tobler, W. R. (1979). Cellular geography. In Philosophy in geography (pp. 379-386). Springer Netherlands.

[30] Tognelli, M.F., Kelt, D.A., 2004. *Analysis of determinants of mammalian species richness in South America using spatial autoregressive models.* Ecography 27, 427436.

[31] Tulbure, M. G., Wimberly, M. C., Roy, D. P., & Henebry, G. M. (2011). *Spatial and temporal heterogeneity of agricultural fires in the central United States in relation to land cover and land use.* Landscape Ecology, 26(2), 211-224.

[32] Wall, M. M. (2004). *A close look at the spatial structure implied by the CAR and SAR models.* Journal of Statistical Planning and Inference, 121(2), 311-324.

[33] Whittle, P. (1954). *On stationary processes in the plane.* Biometrika, 434-449.