Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v40i1.36494

BIOMETRY, MODELING AND STATISTICS

Regression models for prediction of corn yield in the state of Paraná (Brazil) from 2012 to 2014

Rodolfo Seffrin*, Everton Coimbra de Araújo and Claudio Leones Bazzi

*Universidade Tecnológica Federal do Paraná, Campus Medianeira, Avenida Brasil, 3242, Bairro Parque Independência, 85884-000, Medianeira, Paraná, Brazil. *Author for correspondence. E-mail: seffrinrodolfo@gmail.com*

ABSTRACT. This study aimed to identify areas that showed spatial autocorrelation for corn yield and its predictive variables (i.e., average air temperature, rainfall, solar radiation, soil agricultural potential and altitude) and to determine the most appropriate spatial regression model to explain this culture. The study was conducted using data from the municipalities of the state of Paraná relating to the summer harvests in 2011/2012, 2012/2013, and 2013/2014. The statistical diagnostic of the OLS (Ordinary Least Square regression model) was employed to determine the most suitable regression model to predict corn yield. The SAR (Spatial Lag Model) was recommended for all crop years; however, the Spatial Error Model (CAR) was recommended only for the 2013/2014 crop year. The SAR and CAR spatial regressions chosen to predict corn yield in the various years had better results when compared to a regression model that does not incorporate data spatial autocorrelation (OLS). The coefficient of determination (R^2), the Bayesian information criteria (BIC) and the maximum value of the logarithm of likelihood function proved to be better for the estimation of corn yield when SAR and CAR were used.

Keywords: autoregressive spatial model; moran's index; spatial autocorrelation; spatial error model; spatial regression.

Modelos de regressão para a predição da produtividade de milho no Estado do Paraná (Brasil) entre os anos de 2012 a 2014

RESUMO. O presente estudo visou identificar áreas com correlação e autocorrelação espacial para a produtividade de milho e suas variáveis preditoras (temperatura média, precipitação pluvial, radiação solar, potencialidade agrícola do solo e altitude), e também, verificar o modelo de regressão espacial mais adequado para a explicação da cultura. O estudo foi realizado utilizando dados de municípios do estado do Paraná referente a safras de verão dos anos agrícolas de 2011/2012, 2012/2013 e 2013/2014. Para determinar o modelo de regressão mais apropriado para a estimativa da produtividade de milho, foi adotado o diagnóstico estatístico do modelo de regressão OLS - Ordinary Least Square. Para todos anos agrícolas foi recomendado a utilização do modelo de regressão espacial SAR – Spatial Lag Model, sendo que apenas para o ano agrícola 2013/2014 pode ser recomendado o modelo Spatial Error Model (CAR). A regressão espacial (SAR e CAR) adotada para a estimativa da produtividade de milho em diferentes anos, obteve melhores resultados quando comparada com os resultados da regressão que não incorpora a autocorrelação espacial dos dados (OLS). O coeficiente de determinação R^2, os critérios de informação bayesiano (BIC) e o máximo valor do logaritmo da função verossimilhança (Log-likelihood), apresentou melhora significativa na estimação da produtividade do milho quando utilizado SAR e CAR.

Palavras-chave: autocorrelação espacial; índice de moran; regressão espacial; modelo do erro espacial; modelo espacial auto regressivo.

Introduction

Two approaches are usually employed in a spatial correlation analysis that differ in the spatial environment used (global or local). In the global approach, data are considered as a whole, such as when the global Moran's spatial index is used. Alternatively, the LISA (Local Index of Spatial Association) method seeks to evaluate the spatial correlation for an area based on the values of attributes in relation to the values of neighbouring attributes (Anselin, 1995; Zheng, Myint, & Fan, 2014).

Once spatial dependence has been proven to exist (Spatial Correlation), a regression model can be used to identify the variables that best explain the spatial dependence that was found. These models can be used to determine the values of variables based on other variables in the model (Bailey & Gatrell, 1995).

Some studies have used spatial regression models for data estimation, such as studies on estimated

milk production (Ponciano & Scalon, 2010), the analysis of mixed forests in northeastern China (Lou, Zhang, Lei, Li, & Zang, 2016), the estimated impact of urbanization on air quality (Fang, Liu, Li, Sun, & Miao, 2015), the application of a spatial model to predict the number of electric vehicles in the Philadelphia area (Chen, Wang, & Kockelman, 2015), the estimated unemployment rate in Romania (Simionescu, 2015), the estimated malaria incidence in Northern Namibia in 2009 (Alegana et al., 2013), the delimitation of disease risk zones (Charras-Garrido et al., 2013), the relationship between tax evaluation bands and domestic energy consumption in London (Tian, Song, & Li, 2014), the application of a spatial regression model to identify land-cover types in China (Song, Du, Feng, & Guo, 2014), the estimated impact of agriculture on the sale of houses in Pennsylvania (Yoo & Ready, 2016), the estimated frequency of floods in the Northern United States (Ahn & Palmer, 2016), understanding the causes of the reforestation in Vietnam in the 1990s (Meyfroidt & Lambin, 2008), the estimated soil carbon stocks in the city of Chahe, China (Guo et al., 2017), the use of regression models to determine whether residual analysis through electromagnetic induction can be used to survey soil properties (Lu, Zhou, Zhu, Lai, & Liao, 2017), and the use of spatial regression on Cellular Automata to explain and simulate soil alterations (Ku, 2016).

In the present study, regression models were generated to predict corn yield. The OLS (Ordinary Least Square) diagnostic developed by Anselin (2005) was chosen as a criterion for selection, and the following variables were considered: soil classification, altitude, average temperature, rainfall and average solar radiation. The corn crop was chosen as the variable to be analysed spatially because of its national and international importance. This study had the following objectives: 1) to use area spatial statistical techniques to investigate the autocorrelation and spatial correlation between the average corn yield in the summer harvest, soil classification, altitude, rainfall, average air temperature and average global solar radiation in the municipalities of the Paraná State; and 2) to determine the most appropriate regression model to be applied to predict corn yield in the municipalities of the Paraná State using OLS diagnostics.

Material and methods

The area of study comprises the state of Paraná, and this study uses data related to the average corn yield of the main harvest (summer harvest) in the state municipalities, considering variables related to the altitude (m), soil agricultural potential, precipitation (mm), average temperature (°C) and solar radiation (KJ/M^2). The crop years considered in the study were 2011/2012, 2012/2013, and 2013/2014, and the period considered for the agrometeorological data was from early September to late May, following the planting time and harvesting time of the main corn harvest in the state of Paraná according to CONAB (2016).

Data relating to the average corn yields and soil agricultural potential were obtained from the Brazilian Institute of Geography and Statistics (IBGE), and data on the average altitude of the municipalities were obtained from Paranaense Institute for Economic and Social Development (IPARDES). Climate variables relating to precipitation (mm) and average temperature (°C) were obtained from Meteorological System of Paraná (SIMEPAR), National Institute of Meteorology (INMET) and Agronomic Institute of Paraná (IAPAR), and data relating to solar radiation (KJ M^{-2}) were obtained from INMET and IAPAR. Data related to the soil agricultural potential and the altitude were obtained for the 399 municipalities in the state of Paraná despite the fact that no record of corn production existed for 35 municipalities during the study period, so these were excluded from the study. Climate variables (i.e., rainfall, average temperature and solar radiation) were obtained for the municipalities if a meteorological station was available.

The PostgreSQL database version 9.5 (POSTGREE, 2016) with the PostGIS spatial extension version 2.3.0 (POSTGIS, 2016) was used to organize the spatial data (i.e., the geometry of the municipalities) and the non-spatial data (i.e., the variables used in the study). The spatial analysis of the area was carried out using Geoda software, version 1.6.7 (OPENGEODA, 2016) and ArcMap 9.3 (ESRI, 2011).

Weather data related to rainfall and average temperature were obtained from 94 weather stations located in 75 municipalities in the states of Paraná, Santa Catarina, Mato Grosso do Sul and São Paulo. Because data on solar radiation were not available, this information was obtained from 43 weather stations located in 41 municipalities in the states of Paraná, Santa Catarina, Mato Grosso do Sul, and São Paulo. These variables were used following Hoogenboom (2000) recommendation, which indicates that the agrometeorological variables that most affect crop development, growth and

yield are the precipitation rate, air temperature and solar radiation.

The values of weather variables for the municipalities for which a weather station was not available were simulated according to the Thiessen Polygon method, also known as Voronoi polygons, which consists of calculating the area of effect of the weather stations at each polygon (municipality). According to Unwin and Unwin (1998), the area of effect for each weather station can be estimated using this method. The values of the daily weather data for the 399 municipalities in the state from September to May were simulated for the 2011/2012, 2012/2013, and 2013/2014 harvests. Monthly precipitation data were obtained from the sum of the precipitation rates for each month. The average temperature and solar radiation for the period under study (Sep. 1st to May 31st) were calculated, and these values were assigned to each polygon that represented a municipality.

Figure 1 shows the 364 municipalities used in the study in grey, the 35 cities not considered in the spatial analysis in light blue, and the location of the municipalities that had a weather station surrounded by a dark blue line. A red dot indicates the location of the weather stations that had available data for solar radiation, average temperature and precipitation, and the black dots indicate stations that provided precipitation and average temperature data. The weather stations in Santa Catarina, Mato Grosso do Sul, and São Paulo were selected because they have influence on the weather values of some municipalities in the state of Paraná, considering the application of Thiessen polygons.

In the context of this study, ESDA (Exploratory Spatial Data Analysis) was used for the analysis of the spatial autocorrelation of the following variables: average corn yield, soil classification, altitude, average temperature, precipitation and solar radiation. The goal was to statistically determine whether similar values in the neighbouring regions were likely to occur and whether a given variable affected another.

The global Moran's index measures the spatial correlation of geographic locations (municipalities, in this study) using a given variable (for instance, corn yield). Given a set of characteristics and an associated attribute, the parameter determines if the pattern is clustered, scattered or random. The Moran's I statistic was used to represent the clustering degree to characterize the global spatial patterns of the variables in this study (Anselin, 1995; Al-ahmadi & Al-Zahrani, 2013).

$$I = \frac{n \sum_i W_{ij} (X_i - \bar{X}) (X_j - \bar{X})}{\sum_i \sum_j W_{ij} \sum_i (X_i - \bar{X})^2}$$

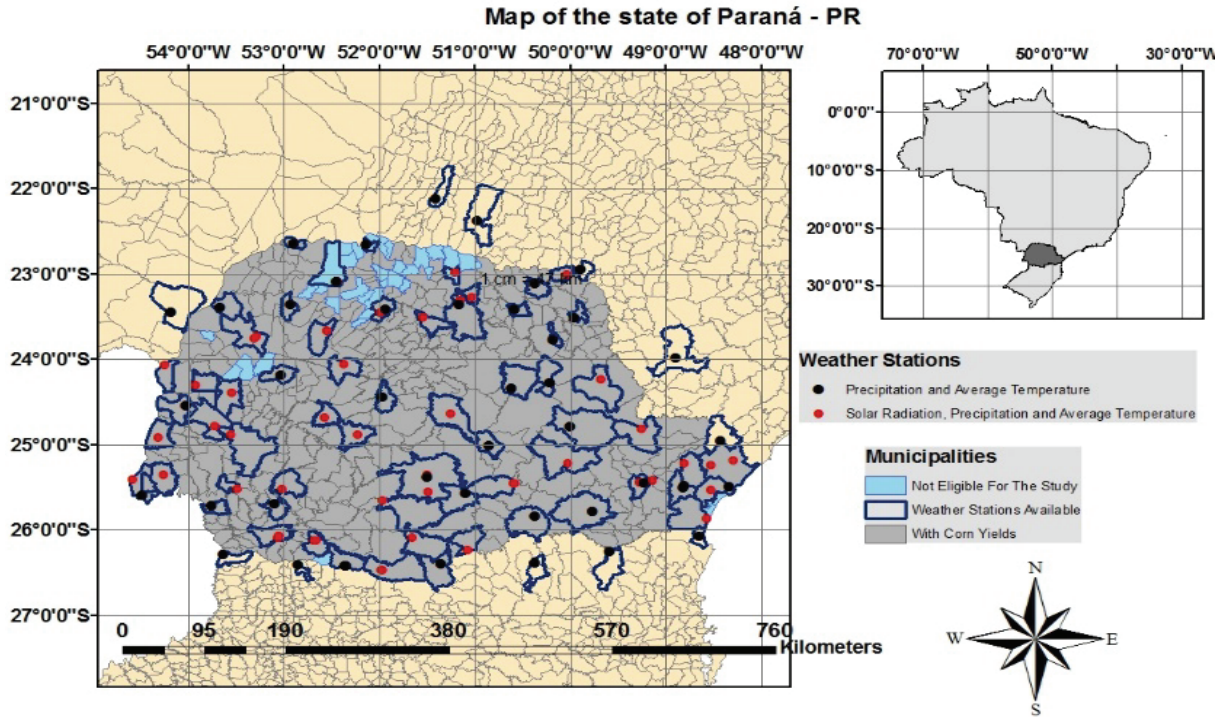

Figure 1. Map of the municipalities of the state of Paraná with their respective weather stations.

For corn yield as an example, X_i is the corn yield at position i in a municipality, $\bar{x}$ is the average corn yield for the municipalities of the state of Paraná, X_j is the corn yield at position j in the municipality, W_{ij} is the neighbourhood matrix for the municipalities i and j, which represent proximity, and n is the number of municipalities.

According to Anselin, Sridharan, and Gholston (2007), each element W_{ij} of the neighbourhood matrix W represents a proximity measurement between the municipalities (polygons) i and j, which can be calculated from one of the following criteria:

Distance between the centroid: $W_{ij} = X$, where X is the distance between i and j if the centroid of i is a certain distance from j; otherwise, $W_{ij} = 0$. For i ≠ j = 1,..., n.

Contiguity (Queen): $W_{ij} = 1$ if i it has a common side with j; otherwise, $W_{ij} = 0$. For i ≠ j = 1,..., n.

Nearest Neighbours: $W_{ij} = X$, where X is the distance between i and j to the n nearest neighbours; For i ≠ j = 1,..., n.

The variables' random spatial dependence when there is no spatial clustering is the null hypothesis (H0) for Moran's I. We simulated 999 data permutations for each result to determine whether the H0 hypothesis was accepted or rejected so that the null distribution (H0) could then be empirically estimated. Several autocorrelation tests are subsequently performed by permuting the values of the variables to be compared. The points' sites remain, and the values of a position are randomly replaced by a value from another point. This technique requires a high computational degree, and the null hypothesis (H0) is rejected when the p-value < 0.05 (Bazzi et al., 2013). Therefore, the p value of the test is given by the following equation:

$$p - value = \frac{ContIf\left(I^{(j)}\right) > I^{(1)}, j = 1, \ldots, n}{n + 1}$$

where: N is the number of permutations, and I: Moran's I.

Another test of significance used was the statistic $Z = \frac{I - E(I)}{\sqrt{VAR(I)}}$, where: E(I) is the expected value (theoretical mean), and VAR(I) is the variance of Moran's I. At significance levels of 1, 5 and 10%, if | Z | ≥ 2.58, | Z | ≥ 1.96 and | Z | ≥ 1.64, then it can be said that there is a significant spatial correlation between the samples. In addition, the Moran's I value varies from -1 to 1, which indicates a negative or positive spatial autocorrelation, respectively. A positive spatial autocorrelation indicates the predominance of similarity in the spatial values for the neighbouring regions, so these areas are classified in the Moran's Scatterplot Matrix as a high/high or a low/low clustering type, while a negative spatial autocorrelation indicates the predominance of dissimilarity between neighbouring regions, and the clustering in this case is the opposite in a Moran's scatterplot, i.e., the high/low or low/high type (Zhang, Hao, & Song, 2016).

Although the global spatial correlation tests identify general trends in all 364 municipalities, it is also necessary to know which cities in the state of Paraná have higher and lower spatial correlation. The LISA method quantify the presence of spatial correlation or clustering. The method identifies which municipalities in the study have similar characteristics (Anselin, 1995).

This statistical approach was developed by Anselin (1995) as a local indicator of spatial association (LISA), which, according to the author, suggests a significant increase of spatial clusters for each observation with similar values around them whose sum also allows the proportion of clustering in relation to the global spatial association to be determined. Taking that into account, the local Moran's I was calculated for all variables in all municipalities of the state of Paraná. The following formula shows the LISA developed by Anselin (1995), and the corn yield variable was used to exemplify the following equation:

$$I_i = \frac{(X_i - \bar{X}) \sum_j W_{ij} (X_j - \bar{X})}{S^2}$$

where: X_i is the corn yield at position i of a municipality, $\bar{x}$ is the mean corn yield in the municipalities of the Paraná State, X_j is the corn yield at position j of a municipality, W_{ij} is the Neighbourhood matrix for the i and j municipalities that represent proximity (Neighbourhood matrix used on equation LISA: distance-from-centroid), and S is the Standard deviation of the corn yield in the municipalities of the Paraná State. The null hypothesis (H0) of this local analysis is that no correlation exists between the variables, and the statistical significance is again estimated via 999 random permutations and a Z-test.

The bivariate Moran's I, expressed as I_{XY}, is used to consider spatially georeferenced variables for which x and y are the variables from n municipalities. The bivariate Moran's I_{XY} can be calculated with the following equation:

$$I_{xy} = \frac{\sum_{i-1}^{n}\sum_{j-1}^{n} u_i z_j w_{ij}}{S_0 \sqrt{S_u^2 S_z^2}}$$

where: *n* is the number of municipalities (populations), $Z_j = (x_j - \bar{x})$ and $u_i = (y_i - \bar{y})$ are the observed values centred on the mean values of *X* and *Y* under study, *Wij* is the element of the proximity matrix W (i.e., the Neighbourhood matrix used on bivariate Moran's I is the queen contiguity, the distance between the centroid and nearest neighbours) $S_0 = \sum_{i-1}^{n}\sum_{j-1}^{n} w_{ij}$; and $S_u^2 = \frac{\sum_{i-1}^{n}(y_i - \bar{y})^2}{n}$, $S_z^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n}$ are the respective variances of Y and X (Anselin, Syabri, & Smirnov, 2003).

The null hypothesis (H0) for statistical significance is estimated through the 999 random permutations and a Z-test, where H0 is that no correlation exists between the variables.

The regression analysis aims to interpret a dependent variable using a weighted linear combination of a set of independent variables. Because the majority of the geographic variables are spatially autocorrelated, spatial regression models are more appropriate than models that do not take spatial autocorrelation into consideration for analysing the relationships between dependent and independent variables, which the OLS model does. Spatial regression methods can be divided into SAR (the Spatial LAG Model) and CAR (the Spatial Error Model). The first method takes the spatial correlation of dependent variables into account, whereas the second considers the autocorrelation effects of random errors (Song et al., 2014). In this study, the dependent variable (Y) is the average corn yield, and the independent variables (X1, X2, X3, X4, and X5), are the soil classification, altitude, mean temperature, precipitation and solar radiation.

An OLS model adjusts the relationships between an independent variable and a set of dependent variables, as shown in the following equation:

$$Y = X\beta + \varepsilon$$

where: Y is the dependent variable, X is the matrix observed on the independent variables N, β, is coefficient of regression and ε is a vector in terms of random error with the distribution $N(0,\sigma^2 I)$.

In the SAR model, the ignored spatial autocorrelation is assigned to the dependent variable Y. The spatial dependence is considered by adding a new term in the form of a spatial relation for the dependent variable to the regression model. Anselin (2002) explains the SAR model according to the following equation:

$$Y = \rho WY + X\beta + \varepsilon$$

where: Y is the dependent variable, X is an independent variable, β is the coefficient of regression, ε represents random errors with a mean of 0 and a variance of σ^2, and W is the Spatial neighbourhood matrix. The null hypothesis to identify the non-occurrence of spatial autocorrelation is $\rho = 0$. The principal idea of this model is to incorporate the spatial autocorrelation as a part of the model.

According to Anselin (2002), the CAR global spatial regression model considers all spatial effects as noise, i.e., errors that must be removed. In this model, the spatial autocorrelation effects are associated with the error term ε, and the model can be expressed according to the following equation:

$$Y = X\beta + \varepsilon, \quad \varepsilon = \lambda W\varepsilon + \xi$$

where: $W\varepsilon$ = errors with spatial effect, ξ = random errors with mean 0 and variance σ^2, and λ = autoregressive coefficient. The null hypothesis for non-existence of spatial autocorrelation is λ = 0, i.e., the error term is not spatially correlated.

A decision process is conducted to determine the most appropriate model (OLS, SAR or CAR) to estimate the spatial data (Anselin, 2005; Song et al., 2014), which is illustrated in Figure 2.

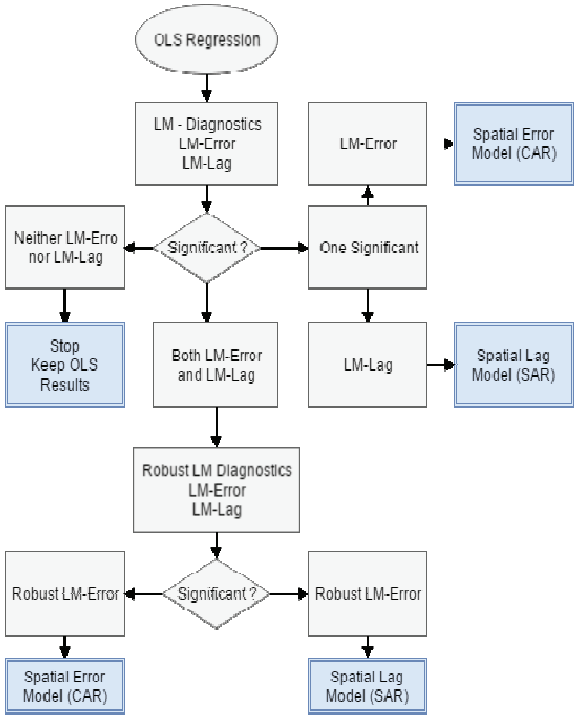

Figure 2. Decision of the Spatial Regression Process (Anselin, 2005; Song et al., 2014).

First, the OLS model is adjusted to obtain regression diagnostics for the spatial dependence of the residuals, and then, four statistical tests are carried out to detect the spatial dependence in linear models. The statistical tests refer to the LM Simple and LM Robust Lagrange Multiplier and include the following tests: LM Lag, LM Error, Robust LM Lag, and Robust LM Error. If the result of the LM Lag test is significant, the presence of spatial dependence is implicit, and the SAR model should be selected. If the result of the LM Error test is significant, the CAR model should be chosen. If the results of both tests (LM Lag or LM Error) are significant, then the spatial model that had a more significant LM Robust test result should be chosen (Anselin, 2005; Song et al., 2014).

Results and discussion

Table 1 describes the Global Moran's index for the variables selected for study, and all variables are indicated to have spatial autocorrelation at a significance level of 1% according to the neighbourhood matrix Queen Contiguity, the distance between the centroid and the nearest neighbours. The Z-score values were higher than 2.71 for all year and neighbourhood criteria, corroborating the spatial autocorrelation significance of the variables. Wrublack, Prudente, Mercante, and Machado Coelho (2013) used the Global Moran's I to determine the spatial dependence of canola yield in the state of Paraná and concluded that this crop also has a significant spatial autocorrelation at the 5% level. Bohórquez, Gómez, and Santa (2011) used the Moran's I to identify areas in which burnings that caused deforestation in the Tuparro National Park in Colombia possibly occurred. Zhang, Luo, Xu, and Ledwith (2008) used the Moran's I to determine the spatial autocorrelation of the presence of organic carbon in soil in southwestern Ireland and investigated critical points of soil contamination by lead (Pb) in Galway, Ireland. Therefore, the Global Moran's I has been used in studies that involve the identification of a spatial pattern related to the geographic location of events in areas of interest.

It was necessary to create cluster maps to show the spatial patterns to visualize the spatial autocorrelation of the corn yield in each municipality. A Moran's scattering map divided the municipalities according to their specific classifications, which are H-H (High-High), L-L (Low-Low) L-H (Low-High) and H-L (High-Low). According to Anselin (1995), HH clusters occur in areas that have a high value for a given indicator surrounded by municipalities that have high values for the same indicator; LL is for an area that has low values for a given indicator and is surrounded by areas that have low values for the same indicator; HL is for an area that has a high value for a particular indicator but is surrounded by neighbours that have low values for the same indicator; and finally, LH is for an areas with a low value for a given indicator that is surrounded by neighbours that have high values for the same indicator.

Figure 3 illustrates the local clusters for the corn yield as calculated by the LISA method, and the maps were generated with the distance-from-centroid neighbourhood criterion. The LISA cluster map represents Moran's scatterplot and shows the H-H (High-High), L-L (Low-Low), L-H (Low-High) and H-L (High-Low) clusters that are significant at 5%. In this case, it can be statistically inferred that they are regions that have local spatial association patterns and require special attention when analysed. An analysis of the clusters shown on the LISA map shows that the northwestern and eastern regions of the state of Paraná have a Low-Low cluster pattern for all years and are also locations that have low corn yields. Thus, the statistic corroborates that the terrain characteristics (i.e., an altitude close to sea level), the soil (which is limited for agricultural practices) and the precipitation (i.e., low rainfall in the west and excessive rainfall in the east) of the municipalities located in this region of the state that make it difficult to achieve a satisfactory average corn yield. In 2011/2012, the northeastern region did not show a Low-Low type cluster as it did in 2012/2013 and 2013/2014, and a determining factor might have been the higher rainfall rate in 2011/2012. The municipalities in the High-High type cluster (high corn yields) are located in the western region of Paraná, and the statistic confirms the good characteristics of this region for a corn crop for all years studied because the soil is the type recommended for agricultural activity, the altitude in some municipalities is considered high and rainfall frequency is good and well distributed. In 2011/2012, the southwestern region had a low corn yield, different than in 2012/2013 and 2013/2014, which had high yields, but the climate (i.e., the rainfall rate, average temperature and solar radiation) was similar, and the soil and altitude were the same. Therefore, a more detailed study of this region for 2011/2012 would be necessary to understand this difference. The central regions of the three maps did not exhibit a significant spatial association pattern. Some municipalities had different characteristics than their neighbours and were classified as Low–High for this reason, i.e., municipalities with low corn yield close to locations with high corn yield and municipalities with high corn yield surrounded by regions with low corn yield.

Table 1. Global Moran's Index applied to the study variables.

Variables	Global Moran's I								
	Contiguity (Queen)			Distance between centroids			Nearest Neighbors		
	2011/2012	2012/2013	2013/2014	2011/2012	2012/2013	2013/2014	2011/2012	2012/2013	2013/2014
Yield	0.6263**	0.6970**	0.7885**	0.6180**	0.6879**	0.7651**	0.6856**	0.7497**	0.8306**
	(18.87)	(20.37)	(24.36)	(21.93)	(25.49)	(27.42)	(13.87)	(14.85)	(17.37)
SPot	0.6757**	0.6757**	0.6757**	0.6000**	0.6000**	0.6000**	0.7315**	0.7315**	0.7315**
	(21.19)	(21.19)	(21.19)	(21.22)	(21.22)	(21.22)	(15.10)	(15.10)	(15.10)
Alt	0.6900**	0.6900**	0.6900**	0.6369**	0.6369**	0.6369**	0.6945**	0.6945**	0.6945**
	(20.32)	(20.32)	(20.32)	(22.74)	(22.74)	(22.74)	(13.86)	(13.86)	(13.86)
Prec	0.8880**	0.7898**	0.8423**	0.8579**	0.7599**	0.8029**	0.9088**	0.8691**	0.8927**
	(27.54)	(23.75)	(25.83)	(30.81)	(27.15)	(27.89)	(18.94)	(17.95)	(17.70)
AvgT	0.9364**	0.9299**	0.9261**	0.9091**	0.9027**	0.8997**	0.9533**	0.9500**	0.9501**
	(28.41)	(27.68)	(28.94)	(31.21)	(31.10)	(32.55)	(19.51)	(20.07)	(20.01)
SR	0.8656**	0.8827**	0.8904**	0.8500**	0.8732**	0.8994**	0.9043**	0.9257**	0.9297**
	(26.58)	(26.59)	(27.28)	(28.53)	(30.81)	(32.47)	(18.56)	(19.26)	(20.03)

Yield: (t ha^{-1}); SPot: soil agricultural potential; Alt: Altitude (m); Prec: precipitation (mm); AvgT: average air temperature (ºC); SR: global average solar radiation (KJ/M²). Between brackets is the Z-score * Significance at the level of 0.05. ** Significance at the level of 0.01.

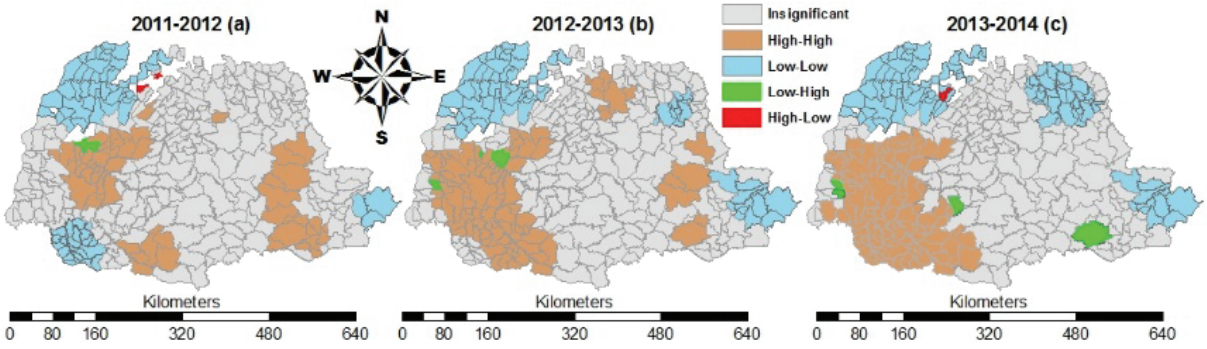

Figure 3. LISA cluster map for corn yields.

The spatial association of corn yield (t ha-1) with the other study variable can be seen in Table 2. According to the calculated value for the bivariate Moran's I, the variables Alt, Prec and AvgT had significant spatial correlation at the 1% level for all years studied (i.e., 2011/2012, 2012/2013, and 2013/2014) and the neighbourhood criteria (i.e., the queen contiguity, the distance between centroid and the nearest neighbours). The variable AvgT exhibited a negative spatial association, and the others were positive. The SPot variable exhibited a positive spatial association at respective significance levels of 1% and 5% for 2011/2012 and 2012/2013, except for the spatial proximity of the nearest neighbours, which were not significant at the 1% or 5% level. However, for 2013/2014, the SPot variable resulted in a negative and significant spatial correlation, different from the others that had a positive correlation. The variable SR indicated a negative and significant correlation at the 1% level in 2011/2012, and in 2012/2013 a positive and significant spatial correlation existed only for the criterion related to distance between the centroid and its neighbours and was non-significant for the other variables. In 2013/2014, a significant spatial association at the 1% level was present for the neighbourhood criteria, i.e., the queen contiguity and the distance from the centroid, but in the other years, these showed no significance.

Table 2. Bivariate Global Moran's Index applied to corn yield with the other variables of the study.

Variables	Bivariate Global Moran's I								
	Contiguity (Queen)			Distance from centroids			Nearest neighbors		
	2011/2012	2012/2013	2013/2014	2011/2012	2012/2013	2013/2014	2011/2012	2012/2013	2013/2014
Spot	0.0841**	0.0590*	-0.0890**	0.0781**	0.0475*	-0.0968**	0.0784**	0.0499	-0.1141**
	(3.69)	(2.39)	(-3.67)	(3.81)	(2.35)	(-4.70)	(2.07)	(1.30)	(-3.07)
Alt	0.3043**	0.2225**	0.2654**	0.2915**	0.2134**	0.2532**	0.3072**	0.2200**	0.2722**
	(11.63)	(9.33)	(10.27)	(12.95)	(9.80)	(11.88)	(7.60)	(5.70)	(6.98)
Prec	0.1752**	0.2838**	0.4796**	0.1794**	0.2842**	0.4611**	0.1705**	0.2950**	0.4712**
	(7.04)	(11.96)	(18.67)	(8.03)	(12.57)	(20.26)	(4.40)	(7.74)	(11.40)
AvgT	-0.3049**	-0.2450**	-0.3713**	-0.3078**	-0.2486**	-0.3673**	-0.3183**	-0.2558**	-0.3746**
	(-11.83)	(-9.89)	(-14.22)	(-14.31)	(-11.39)	(-16.99)	(-8.46)	(-6.76)	(-9.79)
SR	-0.1421**	0.0234	-0.0754**	-0.1183**	0.0463*	-0.0659**	-0.1267**	0.0385	-0.0507
	(-5.66)	(0.99)	(-3.17)	(-5.65)	(2.28)	(-3.20)	(-3.49)	(1.09)	(-1.37)

SPot: Soil agricultural potential; Alt: Altitude (m); Prec: precipitation (mm); AvgT: average air temperature (ºC); SR: overall average solar radiation (KJ/M²). Between brackets is the Z-value. * Significant at 0.05 level. ** Significant at 0.01 level.

After confirming spatial correlation between the variables, the next step was to find an explanation for the average corn yield with the other variables.

Table 3 shows the OLS diagnostics to indicate the most highly recommended regression model to predict the dependent variable $\widehat{Yield}$ (corn yield), according to the flow diagram illustrated in Figure 2, which was estimated using the independent variables P_{prec}, P_{AvgT}, P_{SR}, P_{SPot}, and P_{Alt}. The OLS model did not provide a satisfactory result because the independent variables did not provide an acceptable prediction of the corn yield. This can be explained because this model did not incorporate the spatial dependence between the variables; the R^2 value is lower than 50% for the three years, and the BIC and MVLF criteria improved, which also applies to the estimation of the independent variables in the spatial regression models in Table 4.

The LM Lag, LM Error, Robust LM Lag, and Robust LM Error statistical tests indicated the best spatial regression model for each year. The SAR model was recommended for 2011/2012 and 2012/2013 (Robust LM Error: 0.89 > 0.05; Robust LM Error: 0.051 > 0.05). For 2013/2014, both the SAR and the CAR models were recommended, and both models were used in this case. This criterion for the selection of the best regression model is also discussed in the studies conducted by Anselin (2005) and Song et al. (2014).

Table 4 describes the results of the SAR model for 2011/2012, 2012/2013 and 2013/2014, which were estimated by $\widehat{Yield} = \hat{\beta}_0 + \hat{\beta}_{1\,PREC} + \hat{\beta}_{2\,avgT} + \hat{\beta}_{3\,SR} + \hat{\beta}_{4\,Spot} + \hat{\beta}_{5\,Alt} + \hat{p}\,_{Yield}$, and the CAR model for 2013/2014, explained by $\widehat{Yield} = \hat{\beta}_0 + \hat{\beta}_{1\,PREC} + \hat{\beta}_{2\,AvgT} + \hat{\beta}_{3\,SR} + \hat{\beta}_{4\,SPot} + \hat{\beta}_{5\,Alt} + \hat{\lambda}_{W\varepsilon}$. The responses obtained by the spatial regression models indicate a significant improvement; the coefficient of determination (R^2) for 2011/2012 (64.06%), 2012/2013 (71.94%), and 2013/2014 (CAR: 79.63% and SAR: 79.54%) indicate a better prediction.

The value of the Bayesian information criterion (BIC) and the maximum value of the logarithm of the likelihood function were improved, which are the parameters recommended for spatial regression studies. According to Anselin (2005) and Song et al. (2014), the coefficient of regression R^2 is not appropriate for a spatial regression model. The value listed in the output is not a true R^2 but can be called a pseudo-R^2, and it cannot be compared with the result obtained in the OLS regression.

An analysis of the prediction of independent variables indicated that the average air temperature was the variable that had the greatest influence on the estimated corn yield. Its coefficient value is negative, which indicates that a higher average temperature indicates a lower corn yield. The other variables have positive coefficients, indicating an opposite effect from the average temperature, i.e., a higher corn yield as the variable increases. The auto-regressive spatial coefficient (CAR: $\hat{\lambda}$ and SAR: $\hat{p}$) was significant at the 1% level in all years, and the highest value occurred in 2013/2014 for the CAR model (0.89) and the lowest in 2011/2012 (0.81).

The corn yield prediction for 2011/2012 showed the worst R^2 coefficient, BIC and Log-likelihood. The Global Moran's Index for corn yield in Table 1 suggests a reason: when compared to the other years studied, 2011/2012 showed the lowest spatial autocorrelation, but the highest spatial autocorrelation occurred in 2013/2014, resulting in the best responses in the SAR and CAR spatial regression models.

The Breusch-Pagan test was not significant at the 1% level in any year, which confirms the rejected hypothesis of heteroscedasticity in the data analysed and suggests spatial autocorrelation. The Likelihood Ratio test was significant at the 1% level, confirming the strong importance of using the spatial autoregressive coefficient for the study data for all years. Figure 4 shows the maps of the standardized residual models of spatial regression generated by the SAR and CAR models for 2011/2012, 2012/2013, and 2013/2014.

For 2011/2012: (I) = -0.0031, E(I) = -0.0028, P-value = 0.498 and z-score = 0.0042; for 2012/2013: (I) = 0.0020, E(I) = -0.0028, P-value = 0.442 and Z-score = 0.1784; and for the 2013/2014 SAR model: (I) = 0.0192, E(I) = -0.0028, P-value = 0.202 and Z-score = 0.8302. For the 2013/2014 CAR model: (I) = -0.0051, E(I) = -0.0028, P-value = 0.471 and Z-score = -0.0917; the Moran's I indicated that the residuals are randomly distributed in space, and the z-score is not significant at the 5% level, so the null hypothesis related to the residuals random distribution is accepted. This indicates that the inclusion of the WY component to the models almost eliminated the spatial dependence, meaning that the angle of inclination of the straight line that represents Moran's I in the scatter diagram was very small. Therefore, the regression models generated residuals that were randomly distributed over the study area, as seen in Figure 4, which represents the map of the standardized residuals generated by the standard-deviation method, resulting from the application of the SAR model (2011/2012, 2012/2013, and 2013/2014) and CAR (2013/2014).

Table 3. OLS diagnostics to find out the appropriate model for the application of regression, according to the flow diagram illustrated in Figure 1.

Crop years	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	LM-Error	LM-Lag	R LM-Error	R LM-Lag	R^2	MVLF	BIC
2011/12	8060.49	0.57	-320.4	0.99	258.15	2.02	0.00	0.00	0.89	0.00	20.61	-3246.4	6528.29
2012/13	8808.30	2.33	-503.6	3.04	229.29	1.01	0.00	0.00	0.05	0.00	28.00	-3231	6497.38
2013/14	6979.09	3.84	-771.2	7.07	173.66	0.17	0.00	0.00	0.00	0.00	40.55	-3266.5	6568.38

$\hat{\beta}_0$: linear coefficient estimation; $\hat{\beta}_1$: estimation of the precipitation-related parameter (mm); $\hat{\beta}_2$: estimation of the parameter related to average air temperature (°C); $\hat{\beta}_3$:estimation of the parameter related to the average solar radiation (KJ/M²); $\hat{\beta}_4$: estimation of the parameter related to the soil agricultural potential; $\hat{\beta}_5$: estimation of the altitude-related parameter (m); LM-Error: Significance Lagrange Multiplier (error); LM-Lag: Significance Lagrange Multiplier (lag); R LM-Error: Significance Robust LM (error) ; R LM-Lag: Significance Robust LM (lag); R2: coefficient of determination; MVLF: maximum value of the log-likelihood function = (Log likelihood); BIC: Bayesian information criterion (Schwarz criterion).

Table 4. Spatial Regression Model used to estimate corn yields.

Model	Crop years	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	CAR: $\hat{\lambda}$ / SAR: $\hat{\rho}$	R^2	BP	LR	MVLF	BIC
SAR	2011/12	1204	0.02	-77.82	0.43	74.10	0.87	0.81	64.06	0.89	0.00	-3128.31	6297.9
SAR	2012/13	714	0.63	-107.8	0.74	83.47	0.69	0.84	72.94	0.84	0.00	-3081.91	6205.1
CAR	2013/14	6016.6	1.51	-314.5	3.48	63.90	0.76	0.89	79.63	0.47	0.00	-3106.15	6247.6
SAR	2013/14	-608.3	0.87	-139.4	1.92	65.56	0.63	0.84	79.54	0.22	0.00	-3101.78	6244.8

$\hat{\beta}_0$: linear coefficient estimation; $\hat{\beta}_1$: estimation of the precipitation-related parameter (mm); $\hat{\beta}_2$: estimation of the parameter related to average air temperature (°C); $\hat{\beta}_3$:estimation of the parameter related to the average solar radiation (KJ/M²); $\hat{\beta}_4$: estimation of the parameter related to the soil agricultural potential; $\hat{\beta}_5$: estimation of the altitude-related parameter (m); $\hat{\lambda}$: estimation of the exponential coefficient of the conditional auto-regressive model (CAR); $\hat{\lambda}$: estimation of the exponential coefficient of the simultaneous autoregressive (SAR); R^2: coefficient of determination BP: Breusch-Pagan's test significance; LR: Likelihood Ratio's test significance; MVLF: maximum value of the log-likelihood function (Log likelihood); BIC: Bayesian information criterion (Schwarz criterion).

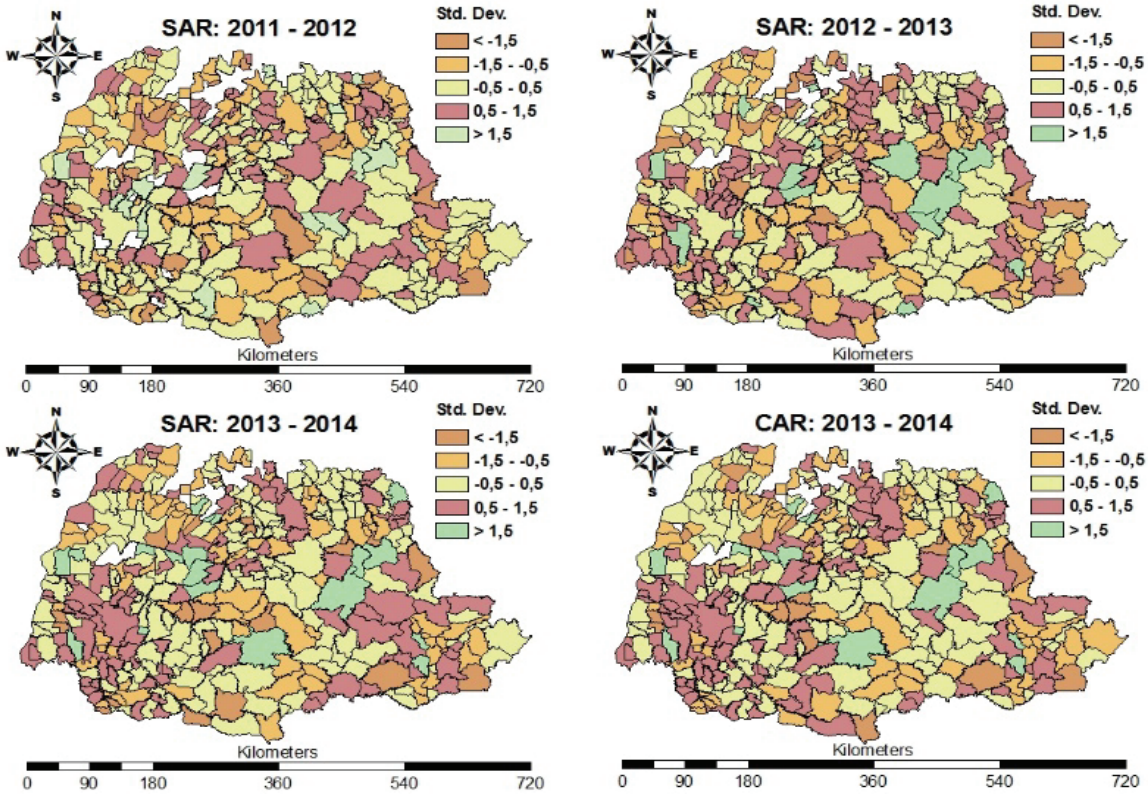

Figure 4. Map of the standardized residuals of spatial regression generated by the Spatial Error and Spatial Lag considering the standard-deviation method.

Conclusion

The global Moran's I indicated spatial autocorrelation for all years, variables and neighbourhood criteria. The LISA method showed that local patterns of spatial association existed, highlighting the regions in the state of Paraná that had similar characteristics. The independent variables, i.e., the average air temperature, precipitation and altitude, had significant spatial association ($p < 0.05$) with the corn yield for all years and criteria.

Application of the SAR regression model for 2011/2012, 2012/2013, and 2013/2014 and the CAR model for 2013/2014 was the most appropriate compared to the OLS regression model. The

coefficient of determination (R^2), the Bayesian information criteria (BIC) and the maximum value of the logarithm of likelihood function proved to be better for the estimation of corn yield.

In this study, the area spatial statistics proved to be effective concerning the application of techniques to investigate spatial patterns. The statistical methods employed in this study proved to be effective for the identification of area patterns, the quantification of spatial autocorrelation and the application of spatial regression.

Acknowledgements

The authors would like to thank SIMEPAR, IAPAR, INMET, IBGE, IPARDES and the Federal Technological University of Paraná for support during this study.

References

Al-Ahmadi, K., & Al-Zahrani, A. (2013). Spatial autocorrelation of cancer incidence in Saudi Arabia. *International Journal of Environmental Research and Public Health*, *10*(12), 7207-7228. doi: 10.3390/ijerph10127207

Alegana, V. A., Atkinson, P. M., Wright, J. A., Kamwi, R., Uusiku, P., Katokele, S., … Noor, A. M. (2013). Estimation of malaria incidence in northern Namibia in 2009 using Bayesian conditional-autoregressive spatial-temporal models. *Spatial and Spatio-Temporal Epidemiology*, 7(suppl), 25-36. doi: 10.1016/j.sste.2013.09.001

Anselin, L. (1995). Local indicators of spatial association – LISA.pdf. *Geographical Analysis*, *27*(2), 94-115.

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247-267. doi: 10.1016/S0169-5150(02)00077-4

Anselin, L., Syabri, I., & Smirnov, O. (2003). Visualizing multivariate spatial correlation with dynamically linked windows. *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, 1-20. Retrieved on Jan. 12, 2016 from http://geodacenter.asu.edu/pdf/multi_lisa.pdf

Anselin, L. (2005). *Exploring Spatial Data with GeoDa: A Workbook*. *Spatial Analysis Laboratory. Department of Geography*. Urbana, IL: University of Illinois. Revised Version, Retrieved on Jan. 12, 2016 from http://www.csiss.org/clearinghouse/GeoDa/geodaworkbook.pdf

Anselin, L., Sridharan, S., & Gholston, S. (2007). Using Exploratory Spatial Data Analysis to Leverage Social Indicator Databases: The Discovery of Interesting Patterns. *Social Indicators Research*, *82*(2), 287-309. doi: 10.1007/s11205-006-9034-x

Ahn, K. H., & Palmer, R. (2016). Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique. *Journal of Hydrology*, *540*(suppl), 515-526. doi: 10.1016/j.jhydrol.2016.06.047

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. New York, NY: Longman Scientific & Technical.

Bazzi, C. L., Souza, E. G., Uribe-Opazo, M. A., Nóbrega, L. H. P., & Rocha, D. M. (2013). Management Zones Definition Using Soil Chemical and Physical Attributes in a Soybean Area. *Engenharia Agrícola*, *3434*(55), 13-17. doi: 10.1590/S0100-69162013000500007

Bohórquez, L., Gómez, I., & Santa, F. (2011). Methodology for the discrimination of areas affected by forest fires using satellite images and spatial statistics. *Procedia Environmental Sciences*, 7(suppl), 389-394. doi: 10.1016/j.proenv.2011.07.067

Charras-Garrido, M., Azizi, L., Forbes, F., Doyle, S., Peyrard, N., & Abrial, D. (2013). On the difficulty to delimit disease risk hot spots. *International Journal of Applied Earth Observation and Geoinformation*, *22*(1), 99-105. doi: 10.1016/j.jag.2012.04.005

Chen, T. D., Wang, Y., & Kockelman, K. M. (2015). Where are the electric vehicles? A spatial model for vehicle-choice count data. *Journal of Transport Geography*, *43*(suppl), 181-188. doi: 10.1016/j.jtrangeo.2015.02.005

Companhia Nacional de Abastecimento [CONAB]. (2016). Acompanhamento da Safra Brasileira - Grãos. *Observatório Agrícola*, *2*(4), 1-60. doi: 2318-7921

ESRI. (2011). *ArcGIS Spatial Analyst*. Retrieved on Jan. 12, 2016 from: http://www.esri.com/software/arcgis/extensions/spatialanalyst/surface.html

Fang, C., Liu, H., Li, G., Sun, D., & Miao, Z. (2015). Estimating the Impact of Urbanization on Air Quality in China Using Spatial Regression Models. *Sustainability*, 7(11), 15570-15592. doi: 10.3390/su71115570

Guo, L., Zhao, C., Zhang, H., Chen, Y., Linderman, M., Zhang, Q., & Liu, Y. (2017). Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. *Geoderma*, *285*(suppl), 280-292. doi: 10.1016/j.geoderma.2016.10.010

Hoogenboom, G. (2000). Contribution of agrometeorolgy to simulation of crop production and its applications. *Agricultural and Forest Meteorology*, *103*(1), 137-157.

Ku, C. A. (2016). Incorporating spatial regression model into cellular automata for simulating land use change. *Applied Geography*, *69*(suppl), 1-9. doi: 10.1016/j.apgeog.2016.02.005

Lou, M., Zhang, H., Lei, X., Li, C., & Zang, H. (2016). Spatial autoregressive models for stand top and stand mean height relationship in mixed Quercus mongolica broadleaved natural stands of northeast China. *Forests*, 7(2), 7-43. doi: 10.3390/f7020043

Lu, C., Zhou, Z., Zhu, Q., Lai, X., & Liao, K. (2017). Using residual analysis in electromagnetic induction data interpretation to improve the prediction of soil

properties. *Catena*, *149*(Part 1), 176-184. doi: 10.1016/j.catena.2016.09.018

Meyfroidt, P., & Lambin, E. F. (2008). The causes of the reforestation in Vietnam. *Land Use Policy*, *25*(2), 182–197. doi: 10.1016/j.landusepol.2007.06.001

OPENGEODA. (2016). *GeoDa center for geospatial analysis and computation*. Retrieved on Jan. 12, 2016 from: http://geodacenter.asu.edu/about

Ponciano, P. F., & Scalon, J. D. (2010). Spatial analysis of the dairy yield using a conditional autoregressive model. *Semina: Ciências Agrárias*, *31*(2), 487-496.

POSTGREE. (2016). PostgreSQL. Retrieved on Jan. 12, 2016 from: https://www.postgresql.org/about/

POSTGIS. (2016). *PostGis. Spatial and Geographic objects for PostgreSQL*. Retrieved on Jan. 12, 2016 from: http://postgis.net/documentation

Simionescu, M. (2015). Predicting the National Unemployment Rate in Romania using a spatial auto-regressive model that includes random effects. *Procedia Economics and Finance*, *22*(suppl), 663-671. doi: 10.1016/S2212-5671(15)00281-6

Song, J., Du, S., Feng, X., & Guo, L. (2014). The relationships between landscape compositions and land surface temperature: Quantifying their resolution sensitivity with spatial regression models. *Landscape and Urban Planning*, *123*(suppl), 145-157. doi: 10.1016/j.landurbplan.2013.11.014

Tian, W., Song, J., & Li, Z. (2014). Spatial regression analysis of domestic energy in urban areas. *Energy*, *76*(suppl), 629-640. doi: 10.1016/ j.energy.2014.08.057

Unwin A., & Unwin D. (1998). Spatial Data Analysis with Local Statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(3), 415-421.

Wrublack, S. C., Prudente, V. H. R., Mercante, E., & Machado Coelho, S. R. (2013). Spatial distribution of Canola culture in the State of Paraná (Brazil) between the agricultural years of 2005 and 2009. *Ciencia y Investigación Agraria*, *40*(3), 523-535.

Yoo, J., & Ready, R. (2016). The impact of agricultural conservation easement on nearby house prices: Incorporating spatial autocorrelation and spatial heterogeneity. *Journal of Forest Economics*, *25*(suppl), 78-93. doi: 10.1016/j.jfe.2016.09.001

Zhang, Y. J., Hao, J. F., & Song, J. (2016). The CO2 emission efficiency, reduction potential and spatial clustering in China's industry: Evidence from the regional level. *Applied Energy*, *174*(suppl), 213-223. doi: 10.1016/j.apenergy.2016.04.109

Zhang, C., Luo, L., Xu, W., & Ledwith, V. (2008). Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the Total Environment*, *398*(1–3), 212-221. doi: 10.1016/j.scitotenv.2008.03.011

Zheng, B., Myint, S. W., & Fan, C. (2014). Spatial configuration of anthropogenic land cover impacts on urban warming. *Landscape and Urban Planning*, *130*(1), 104-111. doi: 10.1016/ j.landurbplan.2014.07.001

Received on March 29, 2017.
Accepted on June 30, 2017.