EXPLORATORY AND CONFIRMATORY SPATIAL DATA ANALYSIS TOOLS IN TRANSPORT DEMAND MODELING

Simone Becker LOPES
PhD Candidate
Department of Transportation
School of Engineering of São Carlos
University of São Paulo
Av. Trabalhador São-carlense, 400
São Carlos 13566-590 Brazil
Tel: +55 16 3373 9613
Fax: +55 16 3373 9602
E-mail: simone@sc.usp.br

Nair Cristina Margarido BRONDINO
Lecturer
Department of Mathematics
Faculty of Science
São Paulo State University
Av. Luis Edmundo C. Coube, 14-01
Bauru 17033-360 Brazil
Tel: +55 14 3103 6086
Fax: +55 14 3103 6447
E-mail: brondino@fc.unesp.br

Antônio Nélson RODRIGUES DA SILVA
Associate Professor
Department of Transportation
School of Engineering of São Carlos
University of São Paulo
Av. Trabalhador São-carlense, 400
São Carlos 13566-590 Brazil
Tel: +55 16 3373 9595
Fax: +55 16 3373 9602
E-mail: anelson@sc.usp.br

Abstract: The main goal of this paper is to discuss a method for the definition of spatial dependence indicators and their inclusion as variables into transportation demand models. The method is based on ESDA (Exploratory Spatial Data Analyses) and CSDA (Confirmatory Spatial Data Analyses) tools, which have been used in two ways, both in a GIS (Geographic Information Systems) environment: i) to produce indicators of spatial dependence; ii) to evaluate the models estimations. The proposed method is applied in a case study in the city of Porto Alegre, State of Rio Grande do Sul, Brazil, based on origin-destination data obtained through household surveys. The results of this work show that ESDA and CSDA tools are very important for the definition of spatial dependency indicators, identification and selection of the most significant spatial variables, specification and evaluation of demand forecast models and evaluation of the results.

Keywords: transportation demand, spatial autocorrelation, GIS, ESDA, CSDA.

1. INTRODUCTION

Regression models are commonly used in the trip generation phase of transport planning, but they can result in some problems if the variables are spatially autocorrelated. Therefore, in this study the effects of spatial dependence on transport demand models are analyzed. That constitutes, without question, one of the main spatial analysis applications in transport planning.

There are various methods that take into account the spatial effects of regression models, but the identification of the best method is strongly influenced by the spatial characteristics of the phenomenon analyzed and by the nature of the independent variables considered. Therefore, the modeling process should start with the analysis of the intensity of spatial relationships among variables. If these relationships are statistically significant, they have to be incorporated into the models. Considering the importance of spatial analysis, the main objective of this study is to discuss a method for the definition of spatial dependence indicators and their inclusion as variables in transport demand models.

The method is based on exploratory and confirmatory spatial data analysis tools, ESDA (*Exploratory Spatial Data Analyses*) and CSDA (*Confirmatory Spatial Data Analyses*). They are used in two ways, both in a GIS environment: *i*) to produce indicators of spatial dependence; *ii*) to evaluate the model estimates. The method proposed is applied to a case study in the city of Porto Alegre, in the state of Rio Grande do Sul, Brazil, using household origin-destination data. The dependent variable studied was the number of *Home-Based Produced Trips* (HBPT) per Traffic Analysis Zone (TAZ).

It should be noted that previous studies conducted by the authors have shown a significant improvement in transportation demand forecasting models with the introduction of two types of spatial variables, Global and Local Indicators of spatial dependence (Lopes and Silva, 2004; Lopes, 2005; Lopes and Silva, 2005; Lopes *et al.*, 2005, and Lopes *et al.*, 2006).

2. ESDA & CSDA TOOLS

One of the objectives of spatial statistics is to characterize spatial patterns in the data, when the data is spatially dependent. That excludes the utilization of many traditional statistical models that require independence between observed events as a basic attribute. Models that do not consider spatial patterns, such as the multiple regression models, are denominated in this study as ***traditional models***. According to Anselin (1992), the spatial patterns can cause problems such as spatial dependence and heterogeneity, which affect the validity of traditional statistical methods. These problems can be identified and quantified by spatial statistics through the use of exploratory and confirmatory analysis tools, ESDA and CSDA, respectively.

2.1 ESDA Tools

The Exploratory Spatial Data Analysis tools (ESDA) have as their objective to visualize and describe spatial distributions, identify standards of spatial association (spatial agglomerations or *clusters*), identify atypical observations (extreme values or *outliers*) or the existence of spatial instabilities (non-stationarity). Wise *et al.* (1998) point out that the methods in this group are descriptive and not confirmatory, with the intent to detect patterns, elaborate hypotheses and estimate spatial models.

One of the functions used to estimate how much an observed attribute value in one region is dependent on the values of the same variable in neighborhood locations is spatial autocorrelation. The *Moran's I* Index indicates, through values that vary from -1 to +1, how similar each area is to its immediate neighbor. The closer to zero, the less the spatial autocorrelation. Values close to -1 or +1 indicate the presence of negative or positive autocorrelation. As such, Moran's I is very useful for the analysis of the initial stage of transport modeling, allowing the identification of characteristics of the dependent variable and possible independent variables.

The *Moran Scatterplot*, utilized to obtain the *global spatial variables* (or global indicators of spatial dependence), is constructed using normalized values of the analysis variable (Z), which are compared with the average of the neighborhood values (W_z) in a two dimensional graph divided into 4 quadrants. *Moran's I* is equivalent to the coefficient that indicates the linear inclination of regression (α) of W_z in Z, such that the quadrants can be interpreted as:

- *Q1* (positive values and averages) and *Q2* (negative values and averages): indicate points of positive spatial association, signifying that a local has neighbors with similar values;
- *Q3* (negative values and positive averages) and *Q4* (positive values and negative averages): indicate points of negative spatial association, signifying that a local has neighbors with distinct values.

Moran's Scatterplot can also be presented in a two-dimensional map, the *Box Map*, in that each polygon is presented indicating its quadrant in the scatter diagram. While the global indicators, like *Moran's I*, provide a unique value as a measure of spatial association for the grouping of data, the local indicators produce a specific value for each area, allowing the identification of regions with similar attribute values (*clusters*), with outliers, and with more than a spatial regime. Anselin (1996) refers to the local indicators as LISA (*Local Indicators of Spatial Association*) statistics.

The statistic significance of the use of *Moran's* local indicator is computed in a form similar to the case of the global index. After calculating the index for each area, the values of other areas are randomly permuted until a pseudo distribution is obtained, for which significant parameters can be calculated. In this case, the maps (LISA *Map* and *Moran Map*) indicate the regions that present local correlation significantly different from the rest of the data, because they are areas with their own spatial dynamics, i.e., pockets of local non-stationarity, and require detailed analysis. Significant autocorrelations to a level of 5 % indicate very similar areas in relation to their neighbors.

It is through Local *Moran* LISA statistics that the *spatial variables*, which are obtained as local indicators of spatial dependence and denominated *local spatial variables*, are introduced into transport demand models in the present study. The visualization indices and tools of ESDA are also very useful in the diagnosis of the models performance, since they can be used in the analysis of the spatial distribution of residues.

2.2 CSDA Tools

Confirmatory Spatial Data Analysis (CSDA) tools group the quantitative processes of modeling, estimation and validation necessary for the analysis of spatial components. It can be highlight, in this group, the "toolkit" available for spatial statistics and spatial econometrics, as spatial regression, or the introduction of indicators of spatial autocorrelation as spatial variables in regression models.

Typically, when performing regression analysis, the aim is to find a good fit between predicted and observed values of the dependent variable in the model. In addition, it is important to find which of the variables significantly contribute to the linear relationship. The standard hypothesis is that the observations are not correlated and, as such, the residuals ε_i of the model, which follow a Normal Distribution with a zero average and constant variance, are independent and uncorrelated to the dependent variable. However, in the case of data that are spatially dependent, it is very unlikely that the standard hypothesis of uncorrelated observations is true. In the most common case, the residues continue to display spatial autocorrelation in the data that can be manifested in systematic regional differences, or even through a continuous spatial trend.

Regression analysis of spatial data incorporates, in the modeling, the spatial dependence between data, improving the predictive power of the model. Initially, an exploratory analysis is conducted with the aim of identifying the structure of dependence in the data. That is very important for the definition on how to incorporate this dependence into the regression model. There exist two basic types of modeling, named spatial regression, that allow the incorporation of the spatial effect: those of Global form and those of Local form (Anselin, 2002, and Fotheringham *et al*., 2000).

The Global models capture the spatial structure through a unique parameter that is added to the traditional regression model. The simplest spatial regression models, formally presented by Anselin (2002), are the *Spatial Auto Regressive* (SAR) or *Spatial Lag Model* and the *Conditional Auto Regressive* (CAR) or *Spatial Error Model*.

Another way to consider the spatial dependence in the regression models, which is called in the present study as the ***alternative*** *transport model*, consists in the introduction of indicators of spatial autocorrelation (Global and Local) as variables. They are added to the traditional variables in the multiple regression model, or ***traditional*** *model* (Lopes and Silva, 2004; Lopes, 2005; Lopes and Silva, 2005; Lopes *et al.*, 2005 and Lopes *et al.*, 2006). In this way, the global and local spatial variables are defined and obtained through spatial analysis of the socioeconomic variables with the use of ESDA tools through spatial statistics computer packages.

The *global spatial variables* are binary (dummy) variables associated to the quadrants of the *Moran Scatterplot* (Global indicator). For an independent variable "X" three variables (X_Q1, X_Q2 and X_Q3) are defined to represent the spatial regime of each TAZ. For the definition of the *local spatial variables* (LISA_X) the LISA indicators are considered.

In the existence of spatial dependence influencing the results of the ***traditional*** *models*, Lopes *et al*. (2006) proved that the ***alternative*** *models* were more efficient than the Global spatial regression models (SAR and CAR) in the prediction of HBPT for the data of Porto Alegre. However, it should be pointed out that, in the same way as the ***traditional*** *models*, the ***alternative*** *models* require rigorous analysis of the significance of the included variables, in order to avoid the addition of unnecessary variables. The previous study also showed that the most significant spatial variables were those that showed significant autocorrelation indices, and that the spatial effect was more evident in the variables expressed as rates (e. g., population, household and vehicle densities).

2.3 Evaluation of Spatial Models

The most common method to select regression models is based on the Maximum Likelihood values of the models, weighted by the difference in the number of estimated parameters. In the models with a dependence structure (spatial or temporal) the evaluation of the adjustment is penalized in function of the number of parameters. It is still necessary to consider the number of independent parameters to incorporate spatial functions into models. For each new variable in the regression model, a new parameter is added.

Usually, the comparison of models uses the Log Likelihood that represents the best adjustment to the observed data. The Akaike *Information Criterion* (AIC) is expressed in Equation (1).

$$AIC = -2 \times LIK + 2k \quad (1)$$

where:

LIK: is the Log likelihood;

k: is the number of regression coefficients.

The best model is the one that has the lowest AIC value. Many other information criteria are available in GIS with spatial statistics, through CSDA tools. Most of them are variations of AIC, with changes in the penalization of parameters or observations.

3. METHOD

The method of investigation of the spatial dependence effects in transport demand forecasting models that is based on Origen-Destination (O-D) surveys was applied through a case study in Porto Alegre, in the State of Rio Grande do Sul, Brazil, in two distinct periods, 1974 and 2003.

It is important to note that the analyses presented here should always be preceded with stages such as: analysis of available O-D survey data, the definition of the study area, preparation of the GIS database, definition of the dependent variable to be analyzed (*Home-Based Produced Trips,* in this study) and selection, in the database, of the candidate independent variables. Additionally, following traditional demand modeling processes, the traditional explanatory variables that are significant to the specification of the best possible model should be selected. It is important to have in mind that the variables are selected to be included in the "***traditional** model*", but it constitutes the base of the "***alternative** models*" subsequently evaluated. The analyses presented in this article are based on the application of ESDA and CSDA tools, occurring in three steps:

1. Verification of the necessity to introduce spatial dependence indicators into demand models: *are there spatial autocorrelation effects that negatively interfere in the performance of the **traditional** model?* (use of ESDA and CSDA tools);

2. Definition of the spatial dependence indicators highly significant as spatial variables, and their introduction in the demand models: *which variables could be perturbing the model performance*? *Which types (global or local) and which spatial variables could be significant for the model*? *Which candidate variables have the largest spatial correlation with the dependent variable*? (use of ESDA tools, *Box Maps* for *Global Spatial Variables* and *Moran Maps* for *Local Spatial Variables*; use of CSDA tools to specify the ***alternative** models*;

3. Comparative analysis of the ***traditional*** and ***alternative*** models: *did the introduction of spatial dependence indicators improve the predictive power of the demand model*? (use of CSDA tools, through statistical tests, and ESDA tools, to analyze the spatial distribution of the estimate residuals).

The GeoDa software (Anselin, 2003 and Anselin, 2004), which is utilized in the analysis presented here, contemplates both the ESDA and CSDA tools. In this way, as well as obtaining spatial variables, calibration and analysis of the models analyzed are also possible through the software. However, even though the GeoDa calculates the *Moran's I* Global index of spatial dependence, the generation of the quadrants of each area of analysis, such as the visualization of the *Box Map*, is not automatic.

The adjustment of the models can be evaluated through the analysis of the values of diverse statistical tests presented in the manuals provided by the software of each adapted model, such as the *Adjusted R-squared* and AIC, among others. For the variables, it was possible to verify the significance (*t-Student*) and the presence of multicollinearity (*Multicollinearity condition number*).

To verify the hypothesis of the existence of spatial autocorrelation affecting the model results, the program provides, from the outputs of each model specified through the CSDA tools, the following statistics: *Moran's I* (*error*), *Lagrange Multiplier* (lag), *Robust Lagrange Multiplier* (*lag*), *Lagrange Multiplier* (error), *Robust Lagrange Multiplier* (*error*), and *Lagrange Multiplier* (SARMA). For a better understanding of this effect, it was possible to visualize the dispersion of residuals through map generation (such as *Moran Maps* for the model residuals) with ESDA tools.

4. ANALYSIS OF THE RESULTS

The results obtained in the application of the method are presented and discussed in this section, following the sequence given in section 3, because the execution of each step is related to the results obtained in the previous step. Initially, in subsection 4.1, the verification of the necessity to include spatial dependence indicators into transport demand models is presented. The study follows with an analysis of which indicators of spatial dependence are more significant as spatial variables in the specification of ***alternative*** *demand models*, which is presented in subsection 4.2. Finally, in subsection 4.3, a comparative analysis of the model results is presented.

4.1 The Need of Spatial Variables in the Models

The ***Traditional*** *model*, which is a Multiple Regression model, was initially adapted through the use of GeoDa software. Standardized values of the independent variables of population (STD_POP) and fleet (STD_CAR), already confirmed as the most significant of the traditional explanatory variables for *Home Based Produced Trips* (HBPT) in 1974, were used in the model. Next, an analysis of the adjusted model was performed from CSDA model outputs. The *Adjusted R-squared* value obtained was 0.91, indicating that the model provides a good explanation of the variance of the dependent variable (HBPT). The *t-Student* tests realized for the model parameters reveal that all are significant to a significance level of 5 %.

The analysis of the model continues with a test to verify the existence of spatial dependence. The Moran's I value of the error (0.429) and its *p value* (0.000) indicate the presence of spatial autocorrelation, which can affect the model performance. This effect can be confirmed through the visualization of the spatial distribution of the residuals, which is presented in the *Moran Map* of Figure 1. It is possible to verify the existence of 7 TAZs in a significant grouping in quadrant 2 (Q2), in other words, low residuals (negative), in a region of the map that represents the centre of the city. Another significant grouping of 13 TAZs, this time in quadrant 1 (Q1), indicating high residuals (positive), appears in the south and southeast of the centre.

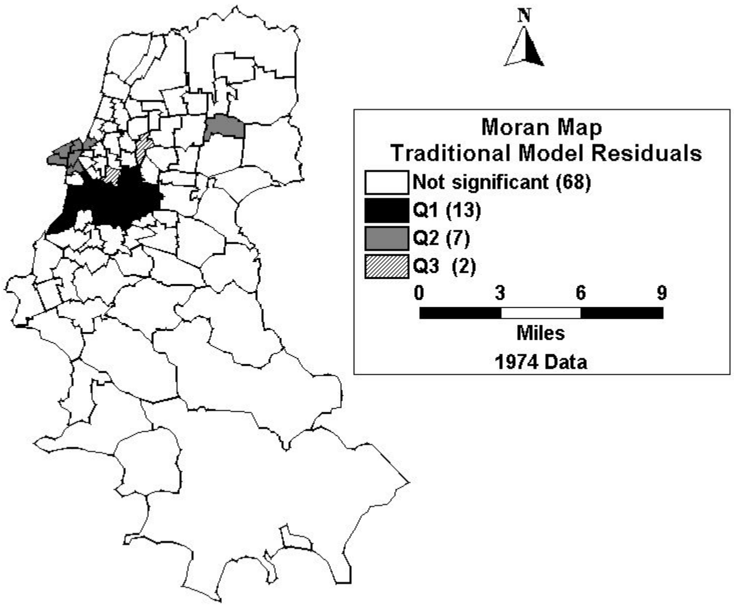


Figure 1 - *Moran Map* – ***Traditional*** *model residuals* – 1974 data

The same process was applied to the 2003 database. In that case, in addition to the variables population and fleet, the number of households and fleet density (number of cars/km^2) were also significant to HBPT. The ***traditional*** *model* adapted for 2003 with the standardized values of these four variables was verified to be suitable. The *Adjusted R-squared* value obtained was very high (0.981) and the *t-Student* tests realized for the model parameters also revealed that they are all significant to a significance level of 5 %. In addition to this, the statistical tests rejected the presence of spatial autocorrelation, which can be confirmed by the *Moran's I* value of the error (-0.002), by the *p value* (0.359), and through the *Moran Map* not indicating the presence of significant groupings of positive and/or negative residuals.

These results indicate that for the 1974 data, it is necessary to specify an ***alternative*** *model* in an attempt to correct the problems of spatial dependence presented. For the 2003 data, as such a problem was not identified; specification of a new model is not justified.

4.2 Definition of Spatial Variables

The definition of spatial variables for the specification of an ***alternative*** *model*, as was indicated in the previous step for the 1974 data begins with an exploratory analysis of the involved variables. The dependent variable (HBPT) and the candidate independent variables were examined to evaluate the presence of spatial autocorrelation. This study has the aim of initially discovering which variables could be affecting the model performance. That is a first step in the search for the best indicators for the model.

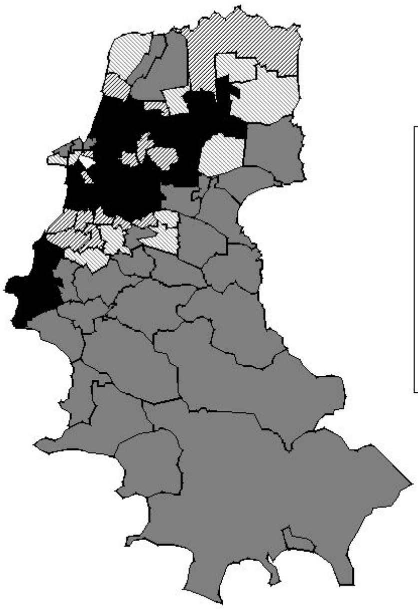
It can be verified in Table 1, for the two periods (1974 and 2003) that the variables that have the greatest spatial autocorrelation indices are those that have their values determined by area, which can be explained by the fact that, following the adopted criteria for division of the TAZs, the effects of the MAUP (Modifiable Areal Unit Problem) could hide the effects of existing spatial autocorrelation in the variables. Such effect is also more evident in the dependent variable (HBPT) when considered by area (D_HBPT).

Table 1 – Spatial Autocorrelation – Moran's I for the variables of 1974 and 2003

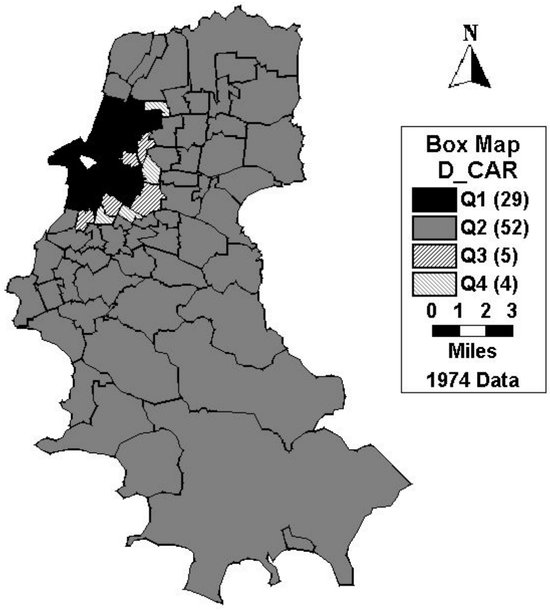
Variable	Description	*Moran's I* (1974)	*Moran's I* (2003)
POP	Total population per TAZ	0.207	0.199
HH	Total number of households per TAZ	0.277	0.139
CAR	Total number of cars per TAZ	0.436	0.234
D_POP	Population density (POP/km^2)	0.734	0.572
D_HH	Density of households (HH/km^2)	0.781	0.681
D_CAR	Density of cars (CAR/km^2)	0.762	0.639
HBPT	Home based produced trips	0.336	0.145
D_HBPT	Density of home based produced trips (HBPT/km^2)	0.754	0.664

From analysis of the data in Table 1, verification of the spatial dependence of the variables is more evident for the 1974 data, because *Moran's I* values are higher. The variable most affected is D_HH, followed by D_CAR and D_POP. It could be concluded that the variables are affected by spatial autocorrelation effects.

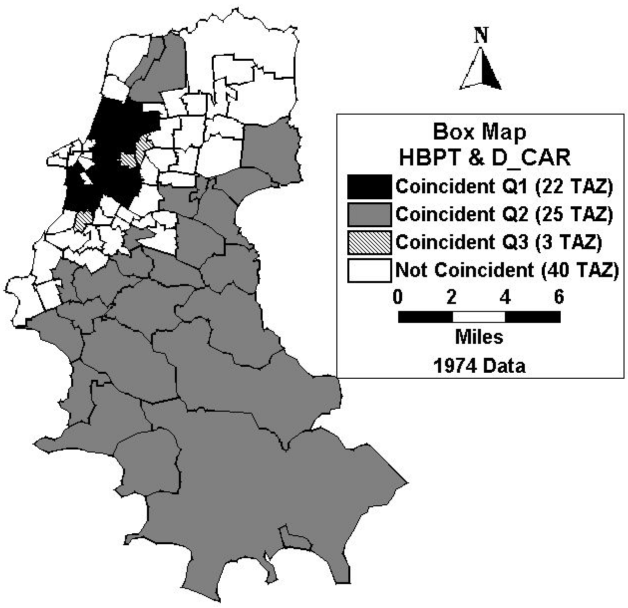
Following the analysis that indicated the variables D_POP, D_HH and D_CAR as those that have the greatest spatial autocorrelation indices, was verified, through a spatial correlation analysis of these variables with the dependent variable, which is the most recommended and which and what types of indicators (Global or Local) can contribute significantly to the ***alternative** model* of 1974. For this, the comparison of coincident elements of the *Box Maps* was done, for the definition of the global spatial variables, and *Moran Maps*, for the local spatial variables (Figures 2 and 3).



2a - Box *Map* – HBPT – 1974 data

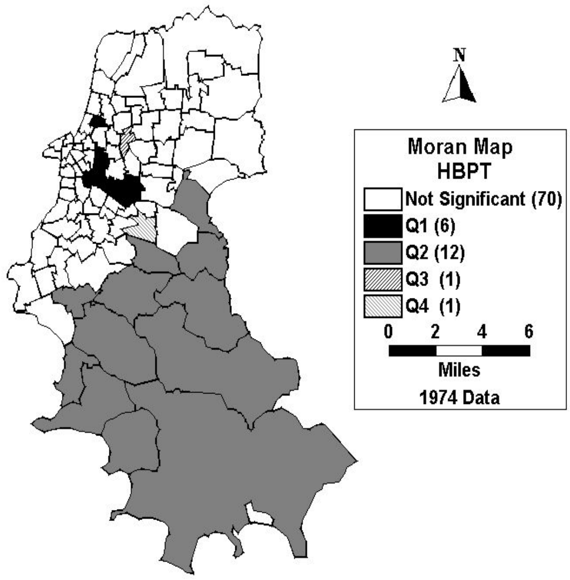


2b - Box *Map* – D_CAR – 1974 data

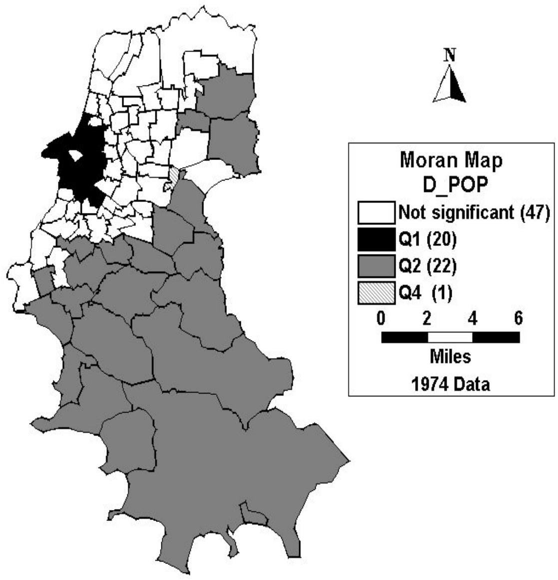


2c – Coincident TAZ analysis (*Box Maps* to HBPT and D_CAR) – 1974 data

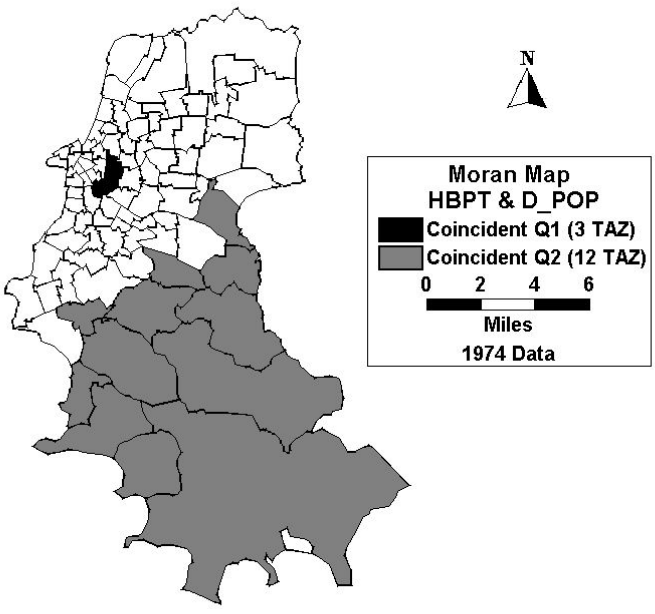
Figure 2 – Example of the spatial correlation analysis – identification of the most significant *Global spatial variables* – *Box Maps* to HBPT e D_CAR – 1974 Data



3a - Moran *Map* – HBPT – 1974 data



3b - Moran *Map* – D_POP – 1974 data



3c – Coincident TAZ analysis (*Moran Maps* to HBPT and D_POP) – 1974 data

Figure 3 – Example of the spatial correlation analysis – identification of the most significant *Local spatial variables* – *Moran Maps* to HBPT and D_POP – 1974 Data

Analyzing the maps by counting the coincident TAZs in the *Box Maps* (as in the example in Figure 2), the most spatially correlated variables to HBPT were verified as being D_CAR and D_POP, with 56 % of the coincident areas, while D_HH was 42 % coincident. Because the Global variables are defined by binary variables for the quadrants, it was necessary to analyze which quadrants are most correlated. Quadrant 2 (Q2) for the variable D_CAR gave the greatest percentage (28 %), followed by quadrant 2 (Q2) for the variable D_POP (27 %), quadrant 1 (Q1) for D_CAR (24 %) and quadrant 1 (Q1) for D_POP (23 %).

The analysis of the most significant local spatial dependence indicators was through a search of coincident zones in the *Moran Maps* (as in the example presented in Figure 3), that gave D_HH and D_POP at 63 % of coincident TAZs, followed by D_CAR with 61 %. Considering just the quadrants where local indicators are significant, the percentages of coincidence are 17 % for D_POP and D_CAR and 16 % for D_HH.

For the 2003 data, the D_POP variable that was most spatially correlated with HBPT had just 20 % of coincident TAZs in the *Box Map*. In the analysis of *Moran Maps*, no significant quadrant area was coincident. It was verified that the spatial correlation observed in the 1974 data did not exist in 2003. That was also indicated by the analysis of the residuals of the 2003 ***traditional** model* and in the analysis of the spatial autocorrelation of the variables presented in Table 1.

For the specification of the ***alternative** models*, through the use of CSDA tools present in the GeoDa software, the spatial variables (Local and Global) were included as per the degree of verified correlation. For each included variable, statistical tests provided by the program were used for analyzing the significance of the variables (*t-Student*), as well as the non-existence of multicollinearity, evaluated by the *Multicollinearity Condition Number*, that should be less than 30. The alternative models for 1974 are presented in Table 2.

Table 2: Summary of the *alternative model* adjusted for the 1974 data – included variables (coefficients and significance) and model evaluation (adjustment and predictive power)

Alternative Model 1974			
Included variables and diagnosis		**Coefficients**	**t-Statistics**
Traditional variables	**Constant**	13208.615	50.142
	STD_POP	4222.638	19.219
	STD_CAR	2121.465	8.424
Spatial variables	**D_CAR_Q2**	- 1753.661	-4.568
	STD_LISA_D_POP	1819.500	2.944
	STD_LISA_D_HH	- 2930.957	-4.742
Model evaluation (adjustment and predictive power)	**R^2 adjusted**	0.953	
	AIC	1558.4	
	Moran's I (error)	0.066 (p value = 0.087)	

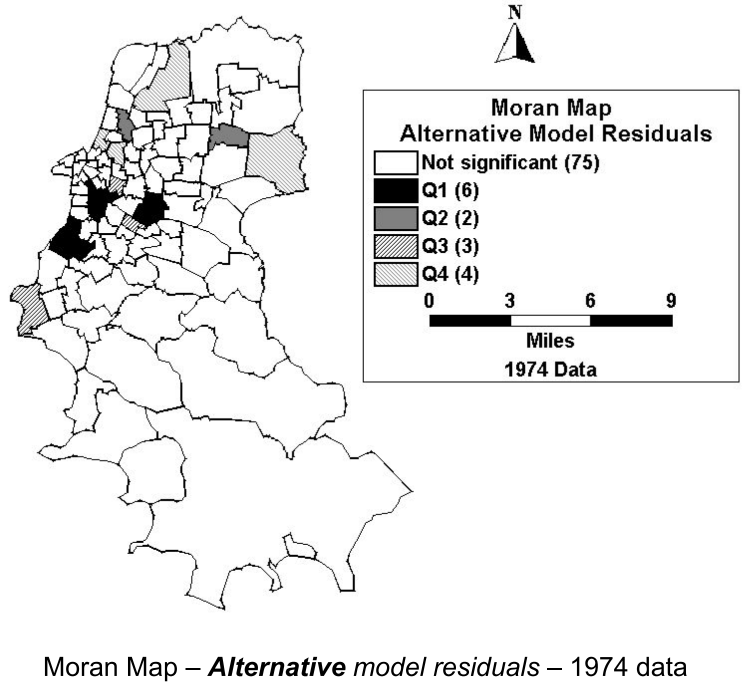
For the ***alternative** model*, adapted for 1974, several variables have shown statistical significance. That is the case of the two traditional socioeconomic variables (STD_POP and STD_CAR), a global spatial variable (D_CAR_Q2) and two local spatial variables, which also has their values standardized (STD_LISA_D_HH and STD_LISA_D_POP). The three spatial variables had a degree of significance compatible with the degree of spatial correlation of the variables D_CAR, D_HH and D_POP with the dependent variable (HBPT).

For the 2003 data, even without having the necessity to specify an ***alternative** model*, but in order to prove the relationship between spatial correlation with the dependent variable and the significance of the spatial variables, the same analysis was performed. Only one global spatial variable has statistical significance (D_POP_Q2), which also corresponded to the indicator most spatially correlated with HBPT.

4.3 Comparative analysis between Traditional and Alternative models

The following stage involved the analysis of the results provided in the outputs of the program for the 1974 *alternative model* (Table 2) and comparison with the results (see section 4.1) previously obtained with the 1974 ***traditional*** *model*. The Adjusted R-squared value obtained with the inclusion of the indicators of spatial dependence in the model increased to 0.953, indicating that the ***alternative*** *model* better explains the variance of the dependent variable (HBPT). A reduction in the AIC statistics from 1610.58 (obtained with the ***traditional*** *model*) to 1558.40 was also observed, what reaffirms the superiority of the ***alternative*** *model* over the ***traditional*** *model*.

The predictive power of the models was compared through the verification of the *Moran's I* of errors and the analysis of the dispersion of residuals, which can be visually conducted through the *Moran Maps*, as presented in Figures 1 and 4, for the residuals of the ***traditional*** *model* and the ***alternative*** *model*, respectively.



Moran Map – ***Alternative*** *model residuals* – 1974 data

Figure 4 - *Moran Map* – ***Alternative*** *model residuals* – 1974 data

As was presented in section 4.1, the ***traditional*** *model* adapted to 1974 data, resulted in a *Moran's I* index of 0.429 for the residuals and the hypothesis of existence of spatial autocorrelation was not rejected by the statistical tests of the model generated by GeoDa software. After the inclusion of spatial dependence indicators (spatial variables), all of the statistical tests, of the denominated "1974 ***alternative*** *model*" (Table 2), rejected the hypothesis of the existence of spatial autocorrelation for the residuals, as *Moran's I* index reduced dramatically to 0.066, confirmed by the *p value* of the test (0.087).

This result could be better understood with the use of ESDA spatial visualization tools. In the *Moran Map* of Figure 4, for example, it is possible to see that the residuals of the 1974 ***alternative*** *model* does not have so many significant groupings of high (Q1) and low (Q2) values as those of the 1974 ***traditional*** *model*, shown in Figure 1.

The same analysis was done for the 2003 models (although the results are not presented in this paper). The analysis has shown a small difference in the dispersion of residuals in the 2003 ***alternative*** *model* when compared with the residuals of the 2003 ***traditional*** *model*. This result was already indicated through the analysis of the 2003 traditional model, described in section 4.1 and confirmed in the analysis of spatial correlation for the 2003 data, commented in item 4.2. The analysis of the *Box Maps* confirmed, in addition to the variable D_POP that presented the greatest coincidence with the spatial characteristics of HBPT, that the spatial correlation was not as high as observed in the 1974 variables.

5. CONCLUSIONS

With the results and analyses presented in this article, the objective of this study, which was to determine the contribution of statistical spatial tools to transport demand modeling, was reached. We initially found that the effects of spatial dependence present in the correlated variables with the specific phenomenon being studied can affect the prediction of this phenomenon if such effects are not considered in the modeling. This problem, which was manifested in the 1974 data by the unrandom spatial distribution of the residuals, could be detected and visualized through the use of CSDA and ESDA tools. Significant spatial autocorrelation indices, such as *Moran's I* (error), and the verification of the presence of significant groupings of positive and/or negative residuals are indications that a new model should be specified in order to take such effects into consideration.

It was also discovered, with the analyses presented in Section 4.2, that the variables that serve the base for obtaining the spatial dependence indicators (*spatial variables*) should result in significant indexes of spatial autocorrelation. The phenomenon was clearly seen in the variables related to area, such as D_POP, D_HH and D_CAR, which was observed for both the 1974 and 2003 data. This relationship reduces the possible effects of MAUP, which can hide the existence of spatial autocorrelation in the variables with absolute values.

The analysis of spatial autocorrelation with the dependent variable, also described in Section 4.2, was a significant contribution for the definition of the type of *spatial variables* (Global or Local) and for the identification of which indicators of spatial autocorrelation are the most significant. We highlight the correlation analysis of the *Box Maps* for the identification of the most significant *Global spatial variables* (such as D_CAR_Q2 for 1974 and D_POP_Q2 for 2003) and the correlation analysis of the *Moran Maps* for the identification of the most significant *Local spatial variables* (such as STD_LISA_D_POP and STD_LISA_D_HH for 1974). If there is no spatial correlation, the variable does not have statistical significance in the model. That was verified in the analysis of *Moran Maps* for 2003, although they were not presented in this article.

This study showed, through the analyses presented in Section 4.3, that the introduction of spatial variables can improve the predictive power of transport demand models (1974 ***alternative*** *model*). This happens, however, only when the spatial autocorrelation effect significantly affects the explicit variables involved (as in the 1974 ***traditional** model*). In the case of the 2003 ***alternative** model*, even though the analyses were not presented in this article, it was verified that the introduction of spatial variables did not improve the predictive power of the model, it actually deteriorated. ESDA and CSDA tools are indispensable for the detection of such spatial effects, and through these tools effective confirmation of the best model design can be determined.

Finally, it is important to note that independently of how spatial effects are considered or even without the necessity to consider such effects, the use of ESDA and CSDA tools constitutes an advance in the analysis of transport models. Their inclusion guarantees a more efficient evaluation of the phenomenon when compared with traditionally utilized techniques, because traditional regression estimators and model evaluation methods do not take into account spatial effects.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the Brazilian agencies CAPES (Post-Graduate Federal Agency), FAPESP (Foundation for the Promotion of Science of the State of São Paulo), and CNPq (Brazilian National Council for Scientific and Technological Development), which have supported our efforts for the development of this work in different ways and periods.

REFERENCES

Anselin, L. (1992) Spatial data analysis with GIS: An introduction to application in the Social Sciences National Center for Geographic Information and Analysis – University of California – California, EUA – In: http://www.ncgia.ucsb.edu/Publication/Tech_Reports/92/92-10.pdf Access in: 04/25/2004.

Anselin, L. (1996) *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. In: *Spatial Analytical Perspectives on GIS,* eds. Fischer, M.; Scholten, H.; Unwin, D. Taylor & Francis, London, p. 111-125.

Anselin, L. (2002) *Under the hood: issues in the specification and interpretation of spatial regression models*, In: <http://agec221.agecon.uiuc.edu/users/anselin/papers/hood.pdf> Access in 03/03/2004.

Anselin, L. (2003) GeoDa 0.9 User's Guide, In: <http://sal.agecon.uiuc.edu> Access in 04/16/2005.

Anselin, L. (2004) GeoDa 0.9.5-i Release Notes, In: <http://sal.agecon.uiuc.edu/> Access in 04/16/2005.

Fotheringham, A. S.; Brunsdon, C. and Charlton, M. (2000) *Quantitative Geography - Perspectives on Spatial Data Analysis*. Sage, London.

Lopes, S. B; Bondino, N. C. M.; Rodrigues da Silva, A. N. (2006) Análise do Desempenho de Modelos de Regressão Espacial na Previsão de Demanda por Transportes. In: XIV Congreso Panamericano de Ingeniería de Tránsito y Transporte, 2006, Las Palmas de Gran Canaria, Spain, 2006.

Lopes, S. B. (2005) Efeitos da Dependência Espacial em Modelos de Previsão de Demanda por Transporte. Master's Thesis, University of São Paulo, São Carlos, SP, Brazil.

Lopes, S. B. and Rodrigues da Silva, A. N. (2004) *An Assessment Study of The Spatial Dependence in Transportation Demand Models*. In: XIII Pan-American Conference in Traffic and Transportation Engineering, Albany, NY, EUA, September 26-29, 2004.

Lopes, S. B. and Rodrigues da Silva, A. N. (2005) Modelos de Previsão de Demanda por Transportes Empregando Análise de Dependência Espacial. In: 19th ANPET Annual Conference of Education and Research in Transportation, Recife – PE, Brazil, 7-11, November, 2005. (v. I. p. 374-385)

Lopes, S. B.; Brondino, N. C. B.; Rodrigues da Silva, A. N. (2005) Um estudo da dependência espacial em modelos de previsão de demanda por transportes no caso de Porto Alegre. In: Antonio Nelson Rodrigues da Silva; Lea Cristina Lucas de Souza; Jose Fernando Gomes Mendes. (Org.). Planejamento urbano, regional, integrado e sustentável: Desenvolvimentos recentes no Brasil e em Portugal. School of Engineering of São Carlos, University of São Paulo, São Carlos, 2005, p. 173-190.

Wise, S.; Haining, R. and Signoretta, C. (1998) The Role of Visualization in the Exploratory Spatial Data Analysis of Area-based Data – Proceedings of the 3rd International Conference on GeoComputational – University of Bristol, UK. In: <http://www.geocomputational.org/1998/> Access in: 09/20/2003.