COMPARING DIFFERENT SPATIAL DATA ANALYSIS TO FORECAST TRIP GENERATION

Cira Souza Pitombo – Federal University of Bahia

Antonio Jorge de Sousa - Technical University of Lisbon (TULisbon)

José Alberto Quintanilha – University of Sao Paulo

Mark Birkin – University of Leeds

1. INTRODUCTION

One fundamental concept in geography is that nearby objects share more similarities than objects which are far apart (Tobler, 1979). As a consequence, similar values for a variable will tend to occur in nearby locations, as a low income county in a remote region may be neighboring other low income counties, for example. This spatial clustering implies that many samples of geographical data will no longer satisfy the usual statistical assumption of independence of observations. Thus, the localization of objects is so important for spatial data analysis (Anselin, 1992).

Spatial data analysis is a quantitative study of objects located in the space. Therefore, the formal techniques allow to find the spatial existing patterns and to measure relationships, considering the spatial localization of the objects. A large body of such techniques has been developed and is found in the literature. A useful classification for spatial data analysis was suggested by Cressie (1991). Cressie's sorting consists of lattice data (discrete variation over space, with observations associated with regular or irregular areal units), geostatistical data (observations associated with a continuous variation over space), and point patterns (occurrences of events at locations in space).

Studies in Transportation Planning field routinely employ data for which location attributes are an important source of information. These studies are associated to variables spatially positioned both in an absolute sense (coordinates) and in a relative sense (spatial arrangement, distance), such as: residential and socioeconomics activities densities, proximity between traffic zones, the transportation network, etc. Thus, the consideration of the spatial attribute in urban trips forecast models is a reasonable way.

The main proposal of this work is to compare two different formal techniques of spatial data analysis to forecast urban trip production and attraction for travel mode and trip purpose in Sao Paulo Metropolitan Area (SPMA), Brazil.

In the present work, two techniques of spatial data analysis were used. The first one belongs to geostatistical data class. It is a multivariate geostatic technique (kriging with external drift - KED). The second one, associated to discrete variation over space, is Geographically Weighed Regression (GWR). Moreover, a methodology based on the joint application of Principal Component analysis (PCA) and the two spatial data analysis techniques was proposed to extract the independent variables and prevent multicolinearity problems.

2. SPATIAL DATA ANALYSIS

Spatial data analysis could be defined as the statistical study of phenomena positioned in space. The focus of attention is essentially location and spatial arrangement. Thus, the observations are referenced in space and their locations are specified as points, lines or areal units (Anselin and Griffith, 1993).

Spatial data analysis aims at extracting knowledge such as spatial relations and patterns. Spatial statistics cross many fields, could be used for predicting ore quality in mining, examining high frequencies of disease events, to predict values at different location and thus produce a surface map of the variable under study (Kriging models), for example. There is an impressive array of sophisticated methods and techniques for visualization, exploration and modeling of spatial data. Some of them will be described here.

2.1 Geographically Weighed Regression

A major problem with regression when applied to spatial data is that the processes being examined are assumed to be constant over space. Geographically Weighted Regression (GWR) is a statistical technique that allows the modelling of processes that vary over space. GWR results in a set of local parameter estimates for each relationship which can be mapped to produce a parameter surface across the study region (Charlton et al, 2005).

GWR is a local multivariate regression function where the data samples are weighted by a function of their spatial proximity. It produces a separate set of regression parameters for every observation across the study area (Li et al, 2008).

The GWR model for each observation point g is:

$$y(g) = \beta_0(g) + \beta_1(g)x_1 + \varepsilon \qquad \text{(I)}$$

“g” represents the vector of co-ordinates of the location, which indicate that there is a separate set of parameters for each of the g observations. Using GWR the parameters can be estimated by solving:

$$\beta(g)=\left(X^T W(g) X\right)^{-1} X^T W(g) Y \qquad \text{(II)}$$

W(g) is the weight matrix which denotes connectivity between observations. The weight can be determined by several methods, as the Gaussian function.

2.2 Geostatistics

For a time, *geostatistics* meant statistics applied to geology or perhaps more generally to problems in the earth sciences. However, it is important to mention that nowadays, geostatistcs is also commonly applied to natural or social sciences (Goovaerts, 1997). Its is valid to highlight that Transportation Planning, specifically studies related to travel behaviour and forecast models of urban trips, can be defined as part of social sciences. These studies consider human and social aspects of individuals and groups and factors as socioeconomic characteristic and behavioural variables.

Geostatistical methods are used to assess the variability of a variable. In a field, the value of a variable generally varies in time and space and the values of a parameter, Z(u) vary with location within the same region. The spatial variability of field measured properties is usually described with autocorrelation or semivariogram (Prompong and Soralump, 2009).

2.2.1Semivariogram

(Semi)variogram analysis means the characterization of spatial correlation. It represents the variation between pairs of measurement as function of distance, defined as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [Z(u_\alpha + h) - Z(u_\alpha)]^2 \qquad \text{(III)}$$

γ (h) = semivariogram for lag distance h, N(h) = number of pairs for lag distance h and $Z(u_\alpha)$ is the value of the variable observed at the location u_α.

The semivariogram for lag distance h is defined as the average square difference of values separated approximately by h , and lag distance should coincide with data spacing, considerably, the variogram is only valid for a distance one half of the field site (Prompong and Soralump, 2009). Figure 1 shows the semivariogram structure and characteristics.

.

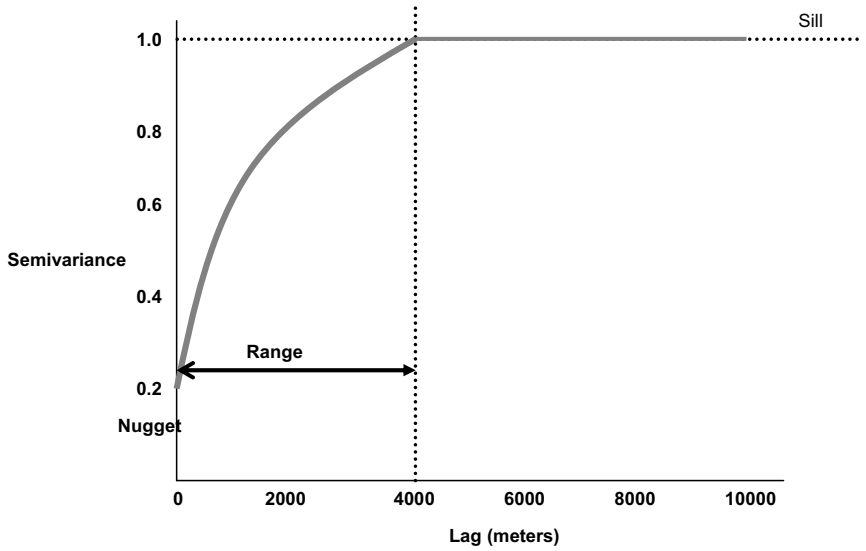


Figure 1: Variogram structure.

Sill: The semivariance value at which the variogram levels flats. **Range**: The lag distance at which the semivariogram reaches the sill value. Autocorrelation is essentially zero beyond the range. **Nugget**: represents variability at distances smaller than the typical sample spacing, including measurement error.

2.2.2 Modeling the semivariogram

For kriging (2.2.3 subsection) objectives, it is necessary to replace the empirical semivariogram with an acceptable semivariogram model. Considering that the semivariogram models used in the kriging process need to follow certain numerical properties to solve the kriging equations, geostatisticians choose from a palette of acceptable semivariogram models. Figure 2 illustrates three usual semivariogram models.

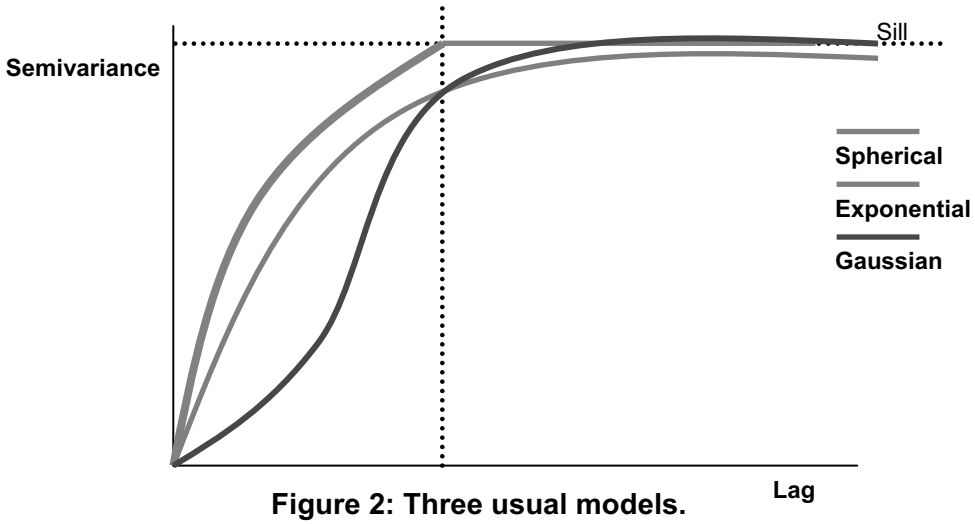


Figure 2: Three usual models.

2.2.3 Kriging

Kriging is an estimator based on regression against observed *z* values of surrounding data points, weighted according to spatial covariance values.Kriging predicts unknown values from data observed at known locations. This method uses variogram to express the spatial variation minimizing the error of predicted values.

Kriging allows deriving weights that result in optimal and unbiased estimates. Within a probabilistic framework, kriging attempts to: a. Minimize the error variance; and b. Systematically set the mean of the prediction errors to zero, so that there are no over – or under-estimates. There is different kind of kriging as simple kriging, ordinary kriging, universal kriging and cokriging. This study will focus on kriging with external drift that is described subsequently.

2.2.4 Kriging with External Drift

Kriging with External Drift (KED) allows the prediction of a variable, known only at small set of points of the study area, through another variable, exhaustively known in the same area. The two quantities are assumed to be linearly related.

KED is a non stationary geostatistical method. It is focused on the use of secondary information from a model to obtain better prediction. In the case of KED, predictions at new locations are made by:

$$Z_{KED}(s_o) = \sum_{i=1}^{n} w_i^{KED}(s_o) * Z(s_i) \qquad \text{(IV)}$$

$$\sum_{i=1}^{n} w_i^{KED}(s_o) * qk(s_i) = q_k(s_o); k = 1,.....p \qquad \text{(V)}$$

z is the target variable,
δ0 is the vector of KED weights (w*i* KED),
qk‘s are the values of the predictor variables at the primary locations s_i and at the location, where the value of the target variable is to be estimated (so).
p is the number of predictors and z is the vector of *n* observations at primary locations.

3. DATA BASE AND ORIGINAL VARIABLES

São Paulo is the largest and most important Brazilian metropolitan area, with a population of over than 17 million, distributed in 39 counties, including São Paulo city. The analysis was based on the origin-destination home-interview survey carried out by METRÔ-SP in SPMA, in 1997. The original sample contains 98,780 individuals. The area was split into 389 Traffic Zones (TZ) and seven regions. Figure 3 illustrated the study region.

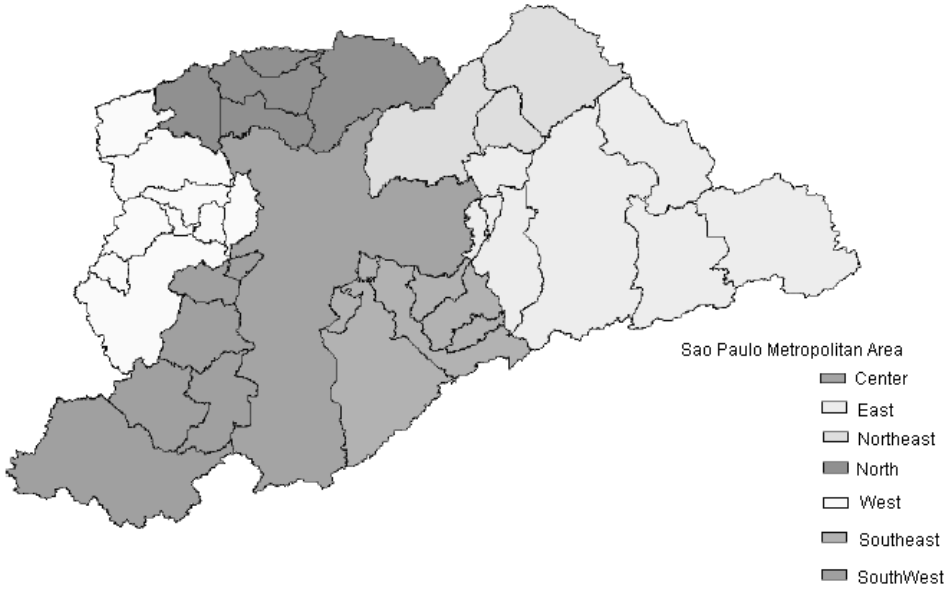


Figure 3 SPMA and seven subregions

For the application of spatial data analysis methods, the following dependent aggregated variables related to trip attraction and production rates are used:

(1) Total of trip production by transit per traffic zone per AREA;
(2) Total of trip production by car per traffic zone per AREA;
(3) Total of trip production non motorized per traffic zone per AREA;
(4) Total of trip production for industry per traffic zone per AREA;
(5) Total of trip production for commerce per traffic zone per AREA;
(6) Total of trip production for services per traffic zone per AREA;
(7) Total of trip attraction by transit per traffic zone per AREA;
(8) Total of trip attraction by car per traffic zone per AREA;
(9) Total of trip attraction by non motorized travel mode per traffic zone per AREA;
(10) Total of trip attraction in industry per traffic zone per AREA;
(11) Total of trip attraction in commerce per traffic zone per AREA;
(12) Total of trip attraction in services per traffic zone per AREA.

The study case presents 24 original socioeconomic variables that could be used as independent variables. However, there is a strong correlation between them. All the original variables are ratios (total/area) and continuous, including age and income. They are presented in Table 1.

Table 1: set of original independent variables

	Variables	Description
1	densemp	Total of Employment per Traffic Zone / Area
2	denspop	Total of Population per Traffic Zone / Area
3	densind	Total of Industry Employment per Traffic Zone / Area
4	denscom	Total of Commerce Employment per Traffic Zone / Area
5	densserv	Total of Services Employment per Traffic Zone / Area
6	densschool	Total of Schools per Traffic Zone / Area
7	densauto	Total of Cars per Traffic Zone / Area
8	DensMale	Total of Males per Traffic Zone / Area
9	DensFemale	Total of Females per Traffic Zone / Area
10	DenshouseholdInco	Average Household Income per Traffic Zone / Area
11	Untill 10 years	Total of People untiill 10 years old per Traffic Zone / Area
12	from 11 to 17 years	Total of People between 11 and 17 years old per Traffic Zone / Area
13	from 18 to 39 years	Total of People between 18 and 39 years old per Traffic Zone / Area
14	from 40 to 59 years	Total of People between 40 and 59 years old per Traffic Zone / Area
15	above 60 years	Total of People above 60 years old per Traffic Zone / Area
16	Income_untillU$130	Total of People with income untill U$130 per Traffic Zone / Area
17	Income_betweenU$130and260	Total of People with income between U$130 and U$260 per Traffic Zone / Area
18	Income_betweenU$260and525	Total of People with income between U$260 and U$525 per Traffic Zone / Area
19	Income_betweenU$525and945	Total of People with income between U$525 and U$945 per Traffic Zone / Area
20	Income_betweenU$945and1890	Total of People with income between U$945 and U$1890 per Traffic Zone / Area
21	Income_overU$1890	Total of People with income over U$1890 per Traffic Zone / Area
22	zero cars	Total of People with no cars per Traffic Zone / Area
23	1 car	Total of People with one cars per Traffic Zone / Area
24	2 or more cars	Total of People with two or more cars per Traffic Zone / Area

Considering spatial patterns at the study region, it is possible to recognize a spatial dependence between SMPA center (Sao Paulo city) and the periphery of the region. The values of the continuous original variables increase from the periphery to the center in general. Figure 4 shows a thematic map of the variables population density respectively.

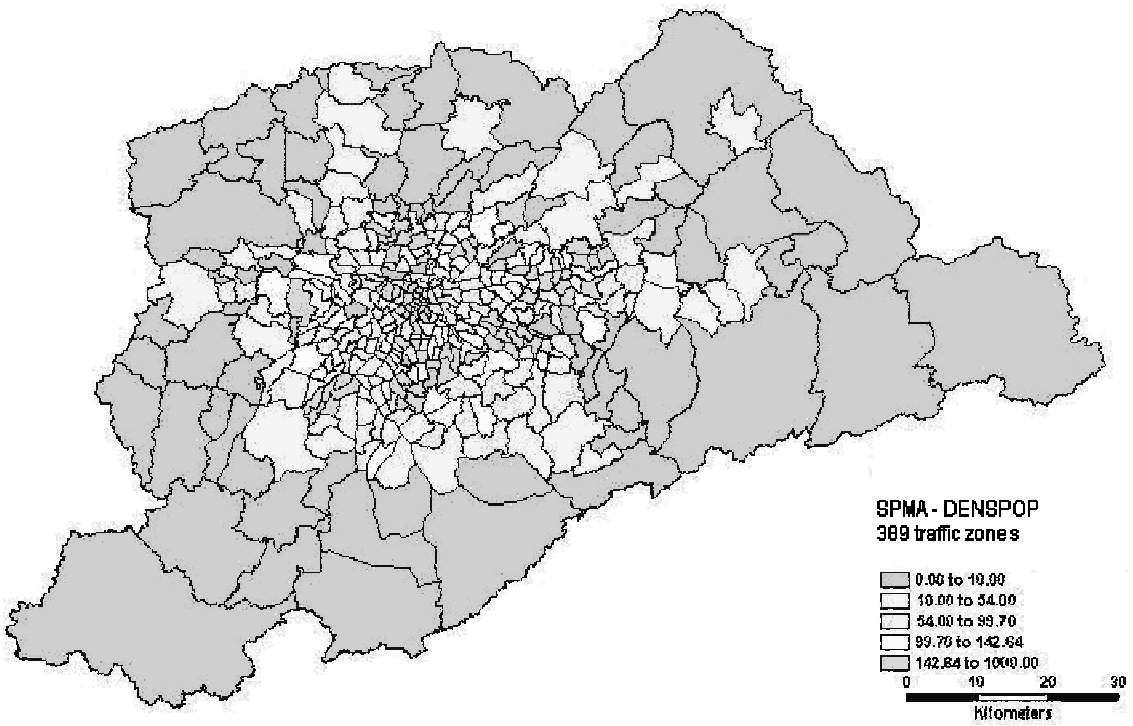


Figure 4 SPMA – Population density variable. Font: METRÔ-SP

4. PURPOSED APPROACH OF SPATIAL DATA ANALYSIS AND RESULTS

The new approach purposed in this paper is the joint application of Principal Component Analysis (PCA), Geographically Weighted Regression (GWR) and Kriging with External Drift (KED). The steps are illustrated bellow (Figure 5) and described in following sub-sections.

The last step, comparing kriging and GWR results, has as main objective the analysis of the best spatial data treatment concerning the trip rates estimation.

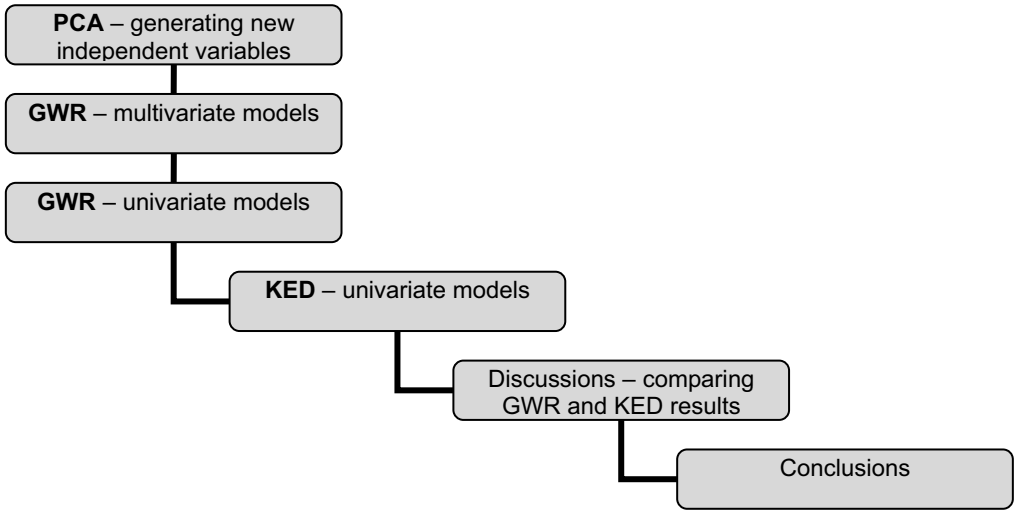


Figure 5 Purposed methodology

4.1 Principal Component Analysis

Principal Components Analysis (PCA) is an exploratory multivariate data analysis technique which the main objective is to detect the data structure (data patterns and relations) and to reduce multidimensional data sets to lower dimensions.

PCA plays an important complementary role with other multivariate techniques through data summarization. The insight provided by data summarization can be directly incorporated into other multivariate techniques. PCA creates a much smaller set of variables and the research can then use component scores, for example. Thus, problems associated with a high number of variables or high intercorrelations between variables (multicollinearity) can be reduced by the previous application of PCA (Hair et al, 1998).

In this work, a combined application of PCA and GWR and PCA and KED was used. The PCA usage has basically two steps, described below:

4.1.1 Extraction of the more significant components:

Taking into account the complete numerical variable set (Table1), it is adequate to reduce this multidimensional data sets to lower dimensions through PCA application. Considering

the latent root criterion for extraction of the components, four components with 88% of the data variability explained were extracted. Table 2 shows the components and the eigenvalues, as well as the percentage of variance and cumulative variance.

Table 2: Results of PCA application

	Initial Eigenvalues		
Component	Total	% of variance	Cumulative %
1	15.639	65.165	65.165
2	2.774	11.556	76.721
3	1.734	7.227	83.948
4	1.171	4.879	88.827

4.1.2 Interpreting and naming components:

In order to achieve a better interpretation regarding the role each variable plays in defining each component it was necessary to analyze the component scores. In general, the researcher attempts to assign some meaning to it, identifying the pattern of component scores, including signs, and trying to name each component. Variables with higher component scores influence the name or label selected to represent a component to a greater extent. The cutoff point for interpretation purposes in this research was all ±.50 or above. Considering the values of the loadings the following nomenclature for each one of the four components was proposed (Table 3).

Table 3: Components loadings and PCA interpretation

Original Variables	Component			
	1	2	3	4
densemp	0.47	**0.85**	0.07	0.09
denspop	0.48	0.01	-0.04	**0.80**
densind	-0.07	**0.80**	-0.07	-0.05
denscom	0.43	**0.83**	-0.04	0.08
densserv	0.46	**0.82**	0.11	0.10
densschool	0.46	0.25	0.00	**0.67**
densauto	**0.62**	-0.02	0.36	**0.60**
DensMale	**0.99**	0.03	0.02	0.02
DensFemale	**0.99**	0.03	0.06	0.04
DenshouseholdInco	**0.72**	0.40	0.42	-0.03
Untill 10 years	**0.93**	-0.04	-0.24	0.01
from 11 to 17 years	**0.95**	-0.04	-0.09	0.02
from 18 to 39 years	**0.98**	0.06	0.02	0.03
from 40 to 59 years	**0.97**	0.04	0.12	0.05
above 60 years	**0.92**	0.03	0.27	0.04
Income_untillU$130	**0.72**	-0.01	-0.41	-0.10
Income_betweenU$130and260	**0.80**	-0.02	-0.41	-0.04
Income_betweenU$260and525	**0.88**	0.00	-0.35	0.00
Income_betweenU$525and945	**0.94**	0.07	-0.19	0.02
Income_betweenU$945and1890	**0.95**	0.04	0.14	0.04
Income_overU$1890	**0.79**	0.01	**0.56**	0.06
zero cars	**0.88**	0.27	-0.17	0.02
1 car	**0.95**	0.06	0.19	0.07
2 or more cars	**0.83**	-0.10	**0.59**	0.06
Component 1 - Age/Socioeconomic Households characteristics				
Component 2 - Employment				
Component 3 - High Income				
Component 4 - Population				

4.2 Applying GWR – Multivariate models

Using GWR3 software (University of Newcastle upon Tyne), it was possible to run multivariate models (independent variables C1, C2, C3 and C4). The software produces a set of localized statistics that can be imported into other software (GIS) for mapping. This way, GWR provides valuable information on the nature of the processes being investigated and supersedes traditional global types of regression modelling.

There are several different varieties of regression model that can be run – here it is assumed that the authors wish to run a GWR with a Gaussian error term. The software generates an output file that contains location-specific parameter estimates and other diagnostics which can be read into a GIS (along with other spatially referenced data) for mapping.

Table 4 shows the models that have been calibrated. There are twelve models (twelve dependent aggregated variables related to trip attraction and production rates described above – section 3). The independent variables were the four components extract with PCA application (C1, C2, C3, C4). This output from GWR3 contains the parameter estimates from a global model fitted to the data. Some useful diagnostic information is printed which includes the coefficient of determination, the estimate of the parameters of the independent variables, the standard error of the parameter estimate and, the t statistic for the hypothesis that the true parameter value = 0.

Table 4: Results of global regression

GLOBAL REGRESSION PARAMETERS

Dependent variables		Coefficient of determination	Adjusted R-square	Intercept			C1		
				Estimate	Std Error	t	Estimate	Std Error	t
Trip Production	Transit	0.9	0.89	54.64	2.81	19.41	85.16	2.81	30.21
	Car	0.95	0.9	43.29	1.42	30.64	66.52	1.42	47.02
	Non-motorized	0.96	0.93	28.4	0.94	30.34	50.67	0.94	54.05
	Industry	0.81	0.66	2.13	0.13	16.95	3.24	0.13	25.72
	Commerce	0.92	0.84	5.74	0.29	19.75	9.33	0.29	32.06
	Services	0.96	0.93	15.81	0.52	30.66	28.23	0.52	54.66
Trip Atraction	Transit	0.95	0.9	55.41	2.92	19.01	86.78	2.92	29.73
	Car	0.95	0.9	43.63	1.44	30.43	67.06	1.44	46.71
	Non-motorized	0.96	0.92	28.18	0.94	29.99	49.82	0.94	52.95
	Industry	0.71	0.51	4.38	0.34	12.98	4.96	0.34	14.7
	Commerce	0.94	0.88	10.98	0.65	16.92	15.25	0.65	23.46
	Services	0.95	0.9	34.13	1.93	17.66	56.25	1.93	29.08

Dependent variables		C2			C3			C4		
		Estimate	Std Error	t	Estimate	Std Error	t	Estimate	Std Error	t
Trip Production	Transit	135.9	2.81	48.21	5.8	2.81	2.05	-12.88	2.81	-11
	Car	34.66	1.42	24.5	22.98	1.42	16.24	5.79	1.42	5.27
	Non-motorized	37.88	0.94	40.41	-16.34	0.94	-16.77	10.27	0.94	10.28
	Industry	0.44	0.13	3.46	-0.4	0.13	-3.16	-1.09	0.13	-1.74
	Commerce	8.09	0.29	27.79	-1.62	0.29	-5.57	-1.4	0.29	-2.36
	Services	10.01	0.52	38.74	3.62	0.52	7.01	8.73	0.52	8.41
Trip Atraction	Transit	138.85	2.92	47.56	4.76	2.92	1.63	-2.85	2.92	-0.98
	Car	36.14	1.44	25.17	23.43	1.44	16.32	2.14	1.44	1.49
	Non-motorized	36.23	0.94	38.5	-15.7	0.94	-16.05	-0.11	0.94	-0.12
	Industry	13.76	0.34	31.13	-0.83	0.34	-2.46	-1.38	0.34	-4.08
	Commerce	29.1	0.65	44.76	-0.65	0.65	-0.99	-2.14	0.65	-3.29
	Services	94.3	1.93	48.75	8.56	1.93	4.42	-1.91	1.93	-0.99

The output from GWR can be voluminous. At every regression point there will be a set of parameter estimates, a set of associated standard errors, and some diagnostic statistics. For this reason here, only some main results will be presented. Table 5 and 6 present results for the model of TRIP PRODUCTION BY TRANSIT for five observations (centroids of traffic zones). Table 5 presents: (1) values of the estimates of the parameters at each regression

point for the intercept and also the four independent variables; (2) Values of the estimates of the standard errors of the parameters at each regression point; (3) Pseudo-t values. Table 6 shows: (1) Observed y variable value; (2) Predicted y variable value; (3) Residuals values and (4) R^2 values.

Table 5: Casewise diagnostics (part 1) – estimates of parameters of variables

ID	LONGY	LATITX	Estimates of the parameters					Standard errors of the parameters					Pseudo-t values				
			Intercept	C1	C2	C3	C4	Intercept	C1	C2	C3	C4	Intercept	C1	C2	C3	C4
1	-46633691	-23548404	63.01	62.82	152.18	11.39	13.68	9.03	5.83	3.21	3.57	6.94	6.98	10.78	47.47	3.19	1.97
2	-46629525	-23543746	66.47	61.57	152.05	11.28	15.34	9.38	5.96	3.19	3.58	7.03	7.09	10.33	47.68	3.15	2.18
3	-46632324	-23553721	60.05	64.74	152.37	11.04	12.56	8.85	5.74	3.21	3.56	6.85	6.79	11.28	47.52	3.10	1.84
4	-46641581	-23551130	58.77	64.63	151.63	11.80	10.94	8.53	5.59	3.23	3.59	6.79	6.89	11.56	46.96	3.29	1.61
5	-46642640	-23543989	62.59	62.71	151.18	12.18	12.09	8.75	5.67	3.22	3.60	6.87	7.15	11.06	46.91	3.39	1.76

Table 6: Casewise diagnostics (part 2)

ID	LONGY	LATITX	OBS	PRED	RESID	R2
1	-46633691	-23548404	2161.26	1653.21	508.05	0.92
2	-46629525	-23543746	635.94	590.56	45.38	0.92
3	-46632324	-23553721	498.92	356.53	142.39	0.92
4	-46641581	-23551130	560.87	557.35	3.52	0.92
5	-46642640	-23543989	1405.90	1588.41	-182.51	0.92

It is probably a little more useful to be able to map the statistics presented above. Figure 6 and Figure 7 show the map of the observed and predicted values of the dependent variable (Trip Production by Transit). Both maps illustrate a spatial pattern: great concentration of high values of the variable in central area and some points of high values of trip production by transit distributed in region. Some of these points (highlighted in pictures) are traffic zones with high economic activities. Moreover, low values of the variable at periphery traffic zones are found.

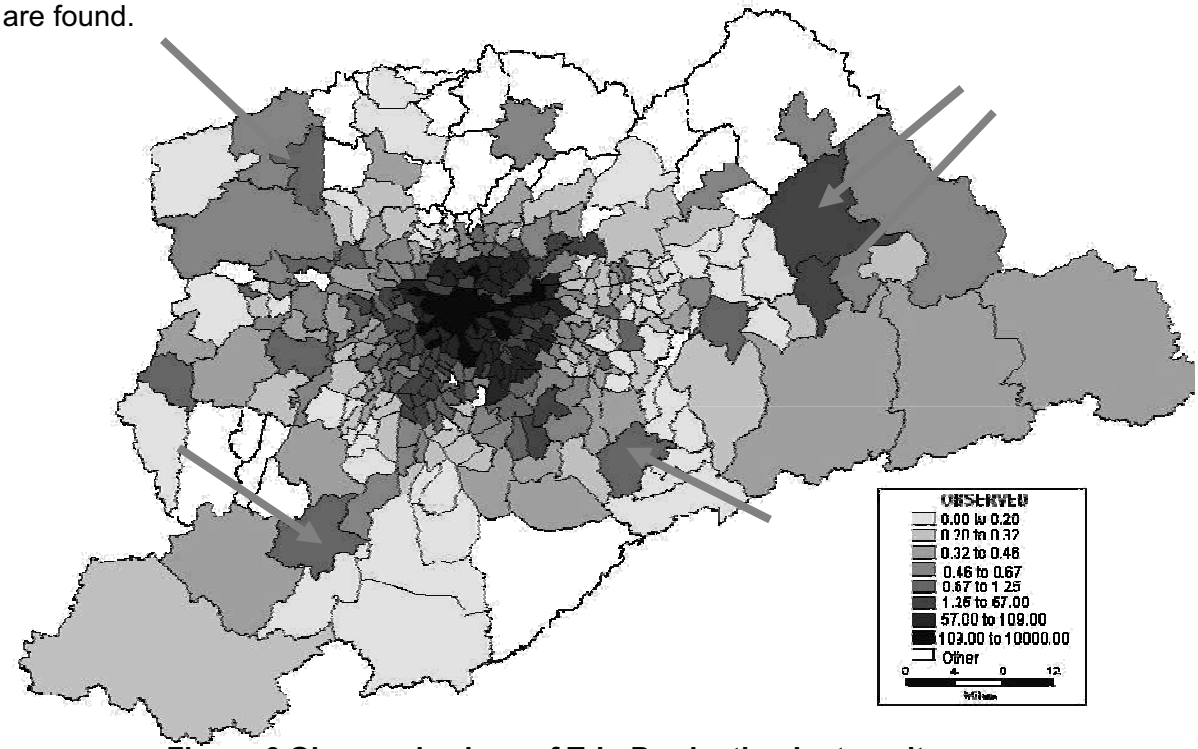


Figure 6 Observed values of Trip Production by transit per area

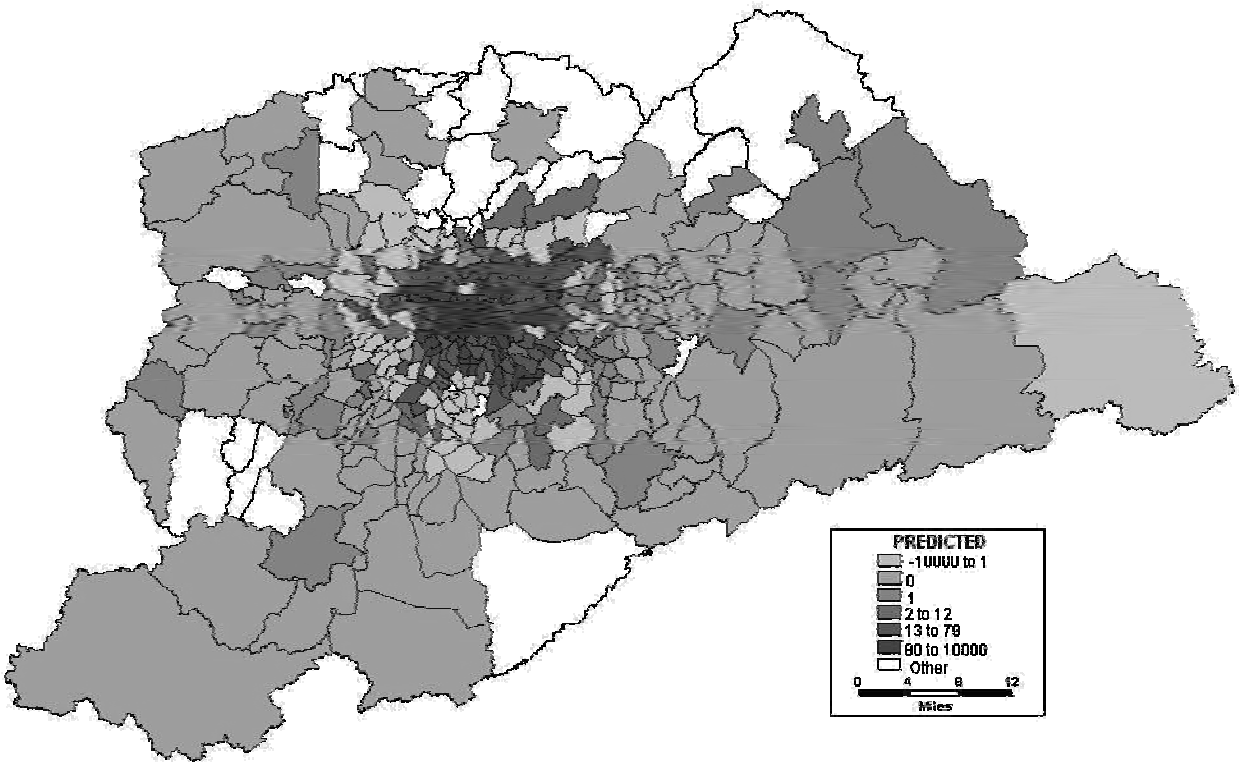


Figure 7 Predicted values of Trip Production by transit per area

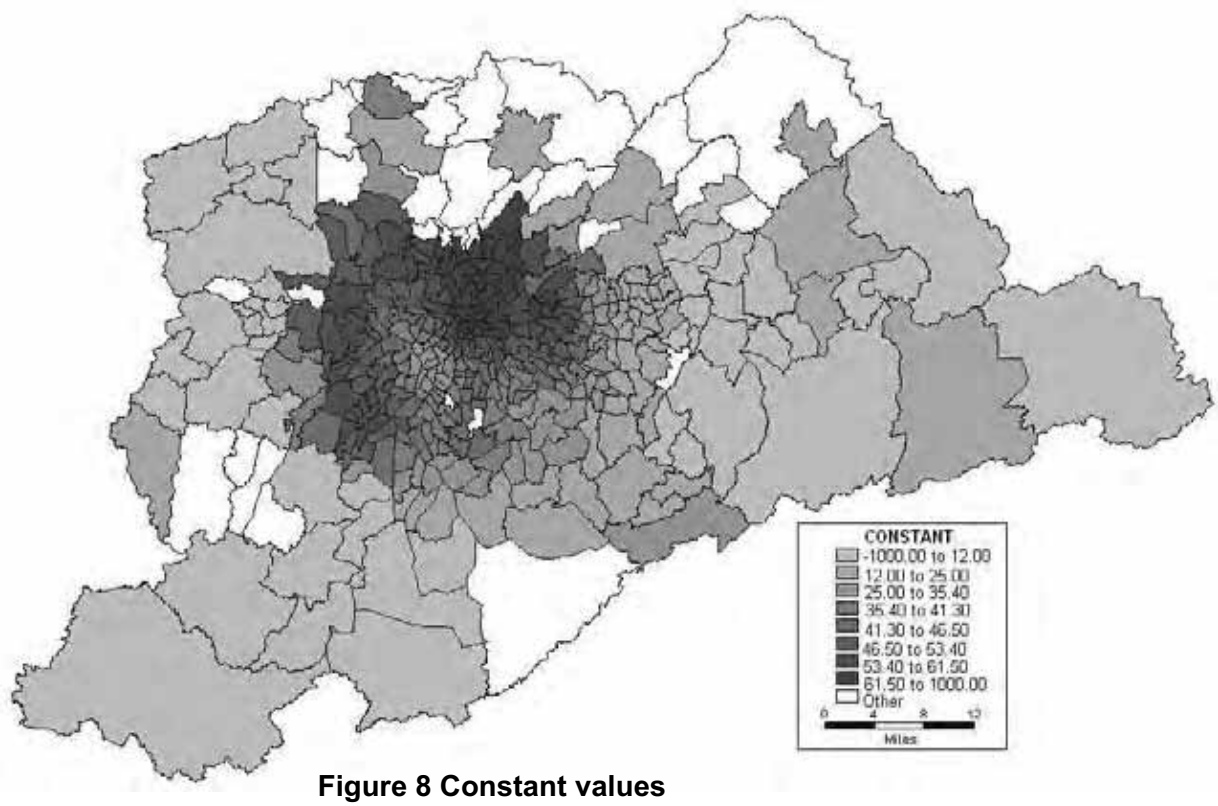


Figure 8 Constant values

Figure 8 illustrates the constant values of GWR. The Figure 9 shows the parameters values of component 1. Component 1 represents Age/Socioeconomic household characteristics and, in general, positive and high values for parameters for this independent variable are indentified. Considering the global regression model, the map shows similar values for this parameter in the center (70.70 to 84.10). Furthermore, the values increase with a buffer/ radial pattern in the center region.

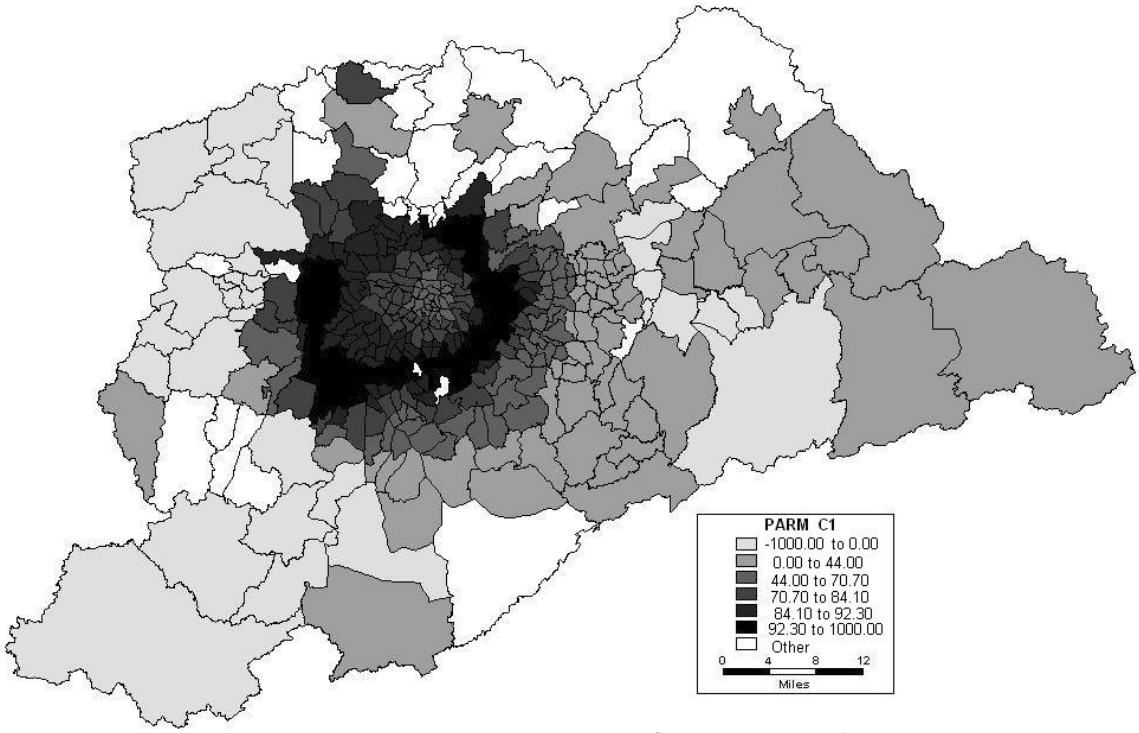


Figure 9 Parameters values of component 1

The Figure 10 shows the parameters values of Component 2. Component 2 represents Employment. Employment is highly correlated to trip production, so it is possible to see a high and positive value for the parameter of Component 2 in global model (135.9). Positive and high values for parameters for this independent variable are observed. Considering the global regression model, the map shows similar values for this parameter in the center (138.30 to 1000.00). The parameter values decrease with a buffer/ radial pattern in the center region.

The following map (Figure 11) illustrates the parameter values of Component 3. Usually this variable presents a low coefficient value. Component 3 represents high income. As expected, this variable is not so highly correlated to trip production by transit. Maybe, the coefficients values for this variable are higher considering trip production by car, for example. High income is high correlated to car ownership and, consequently, to travel mode choice. Figure 12 shows the values of the parameters estimated for Component 4 (population).

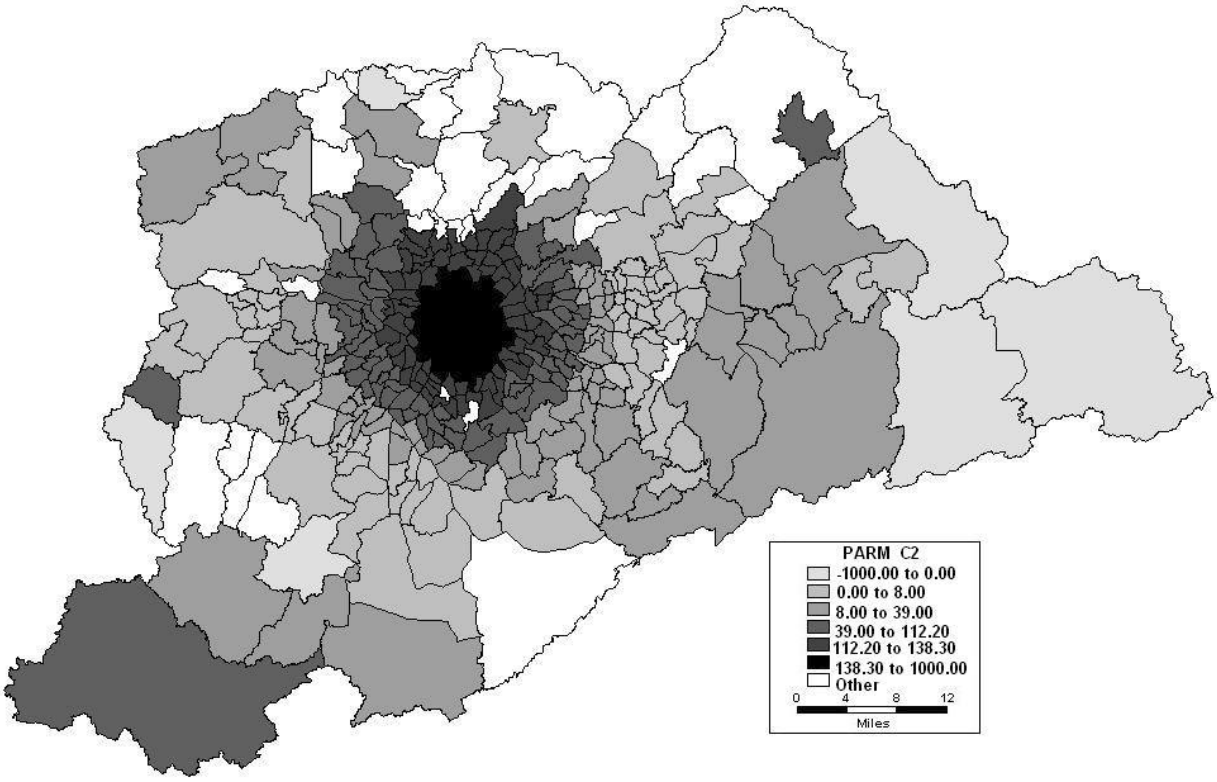


Figure 10 Parameters values of component 2

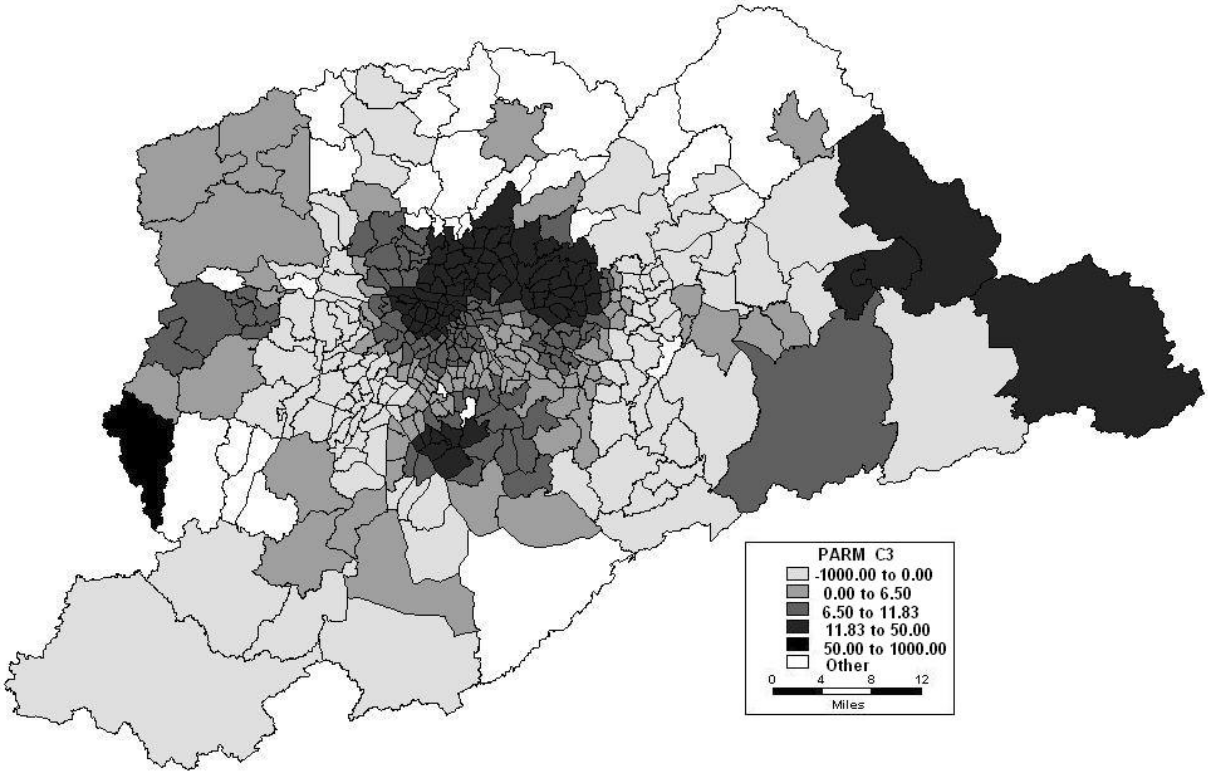


Figure 11 Parameters values of component 3

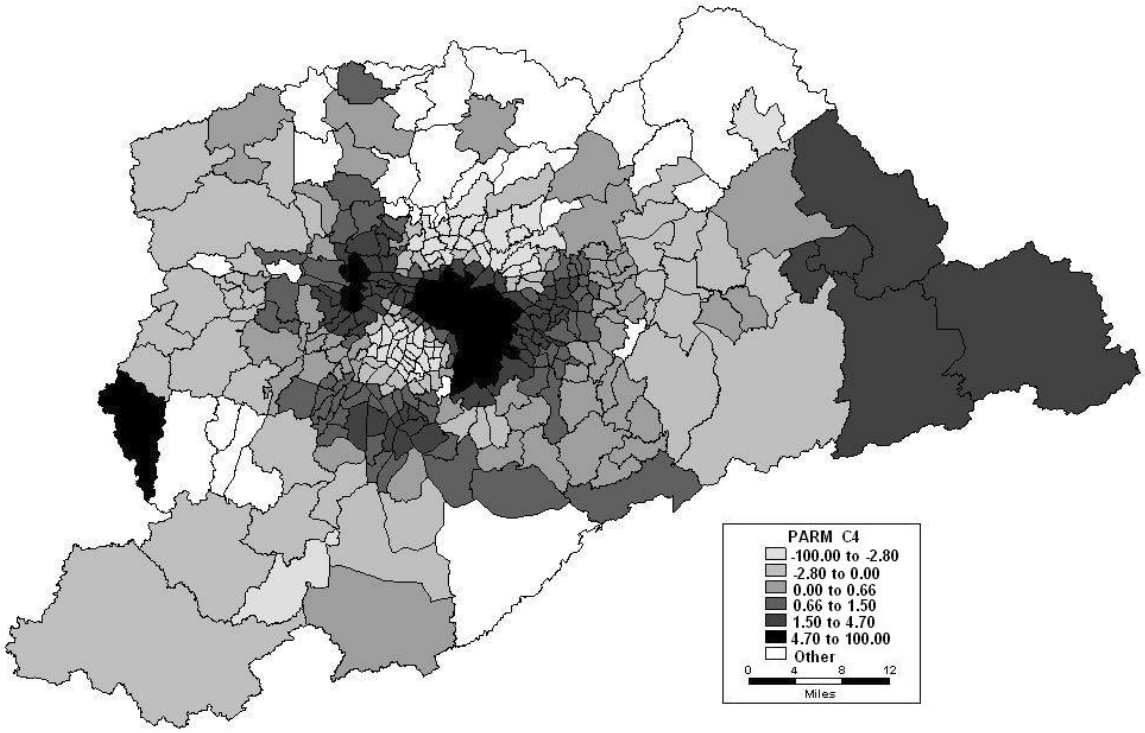


Figure 12 Parameters values of component 4

4.3 Applying GWR – Univariate models

For comparing the two spatial data analysis techniques (KED and GWR), it was necessary to choose only one of the components as independent variable to integrate the GWR model. Even if it implied loosing some accuracy, the authors had to generate univariate models taking into account the KED application. For KED application it is possible to use only one variable as secondary variable (GeoMs software allows only one secondary variable).

Thus, the components were selected as independent variables in GWR considering the highest correlation with the dependent variables. The component used in univariate models varied between Component 1 and Component 2. Component 3 and Component 4 did not enter in the models. Table 7 summarises the main results of the global univariate models.

Table 7: Results of global regression univariate models

GLOBAL REGRESSION PARAMETERS												
Dependent variables		Coefficient determination	Adjusted R-square	Intercept			C1			C2		
				Estimate	Std Error	t	Estimate	Std Error	t	Estimate	Std Error	t
Trip	Transit	0.84	0.65	54.65	5.31	10.29				135.85	5.32	25.61
Production	Car	0.83	0.64	43.29	2.61	16.59	66.52	2.61	25.46			
	N-motorized	0.77	0.59	28.4	2.23	12.73	50.67	2.23	22.68			
	Industry	0.8	0.64	2.13	0.129	16.5	3.24	0.129	25.04			
	Commerce	0.69	0.47	5.74	0.52	10.97	9.33	0.52	17.81			
	Services	0.78	0.61	15.81	1.19	13.28	28.23	1.19	23.68			
Trip	Transit	0.83	0.65	55.41	5.43	10.2				138.85	5.43	25.53
Atraction	Car	0.8	0.63	43.63	2.68	16.24	67.06	2.69	24.92			
	N-motorized	0.77	0.6	28.17	2.15	13.11	49.82	2.15	23.14			
	Industry	0.55	0.3	4.37	0.4	10.96	4.96	0.4	12.42			
	Commerce	0.83	0.69	10.98	1.04	10.57				29.09	1.04	27.96
	Services	0.81	0.66	34.13	3.57	9.57				94.31	3.57	26.39

4.4 Kriging with External Drift (KED)

KED application – spatial interpolation - was done considering the following primary and secondary variables:

- Model 1 : **Primary** – Trip Production by transit; **Secondary** – Component 2
- Model 2 : **Primary** – Trip Production by car; **Secondary** – Component 1
- Model3 : **Primary** – Trip Production with non-motorized travel mode; **Secondary** – Component 1
- Model 4 : **Primary** - Trip production for industry ; **Secondary –** Component 1
- Model 5 : **Primary** - Trip production for commerce ; **Secondary –** Component 1
- Model 6 : **Primary** - Trip production for services ; **Secondary –** Component 1
- Model 7 : **Primary** – Trip Attraction by transit; **Secondary** – Component 2
- Model 8: **Primary** – Trip Attraction by car; **Secondary** – Component 1
- Model 9: **Primary** – Trip Attraction with non-motorized travel mode; **Secondary** – Component 1
- Model 10 : **Primary** - Trip Attraction for industry ; **Secondary –** Component 1
- Model 11 : **Primary** - Trip Attraction for commerce ; **Secondary –** Component 2
- Model 12 : **Primary** - Trip Attraction for services ; **Secondary –** Component 2

In the present work, the data had been considered for geostatistics interpretation, using the package software geoMS (Geostatistical Modelling Software), developed for *the Instituto Superior Técnico* of the Technical University of Lisbon.

The geoMs is software directed to analyze geostatistical data that presents spatial dependence. It is composed for some modules that are necessary to aim the considered objectives. The geoestatistical procedures used will be described in the next sub-section.

4.4.1 Data treatment and spatial data description:

For data preparation of input, the variables mentioned above were considered. All variables are aggregated by centroid of traffic zone (389 in the total). The input presents a total of 389 Ids (traffic zones), co-ordinated (in meters) and the variables.

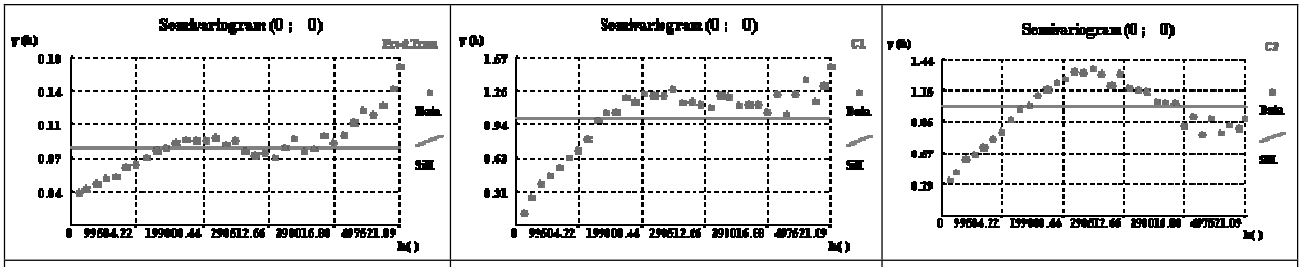
The spatial description of data has as objective to visualize the manner as each variable is dispersed in the space. This procedure is important to define the main characteristics of each variable: anisotropies e discontinuities. Through the visualization of the spatial arrangement of the experimental data, it is possible to be familiar with the basic parameters for the calculation of the variograms: directions, classes of angles and distances.

4.4.2 Generating experimental variograms:

The calculation of the variograms for the primary and secondary variables was effected according to five directions: (0º, 45º, 90º, -45 and omnidirectional one). In the ominidirecional variogram, the same weighting scheme is attributed to samples that are at the same distance

(h) of the point, even so in different directions. Variograms according to determined direction use the pairs of samples lined up in the corresponding directions.

The experimental variogram allows characterizing the spatial behavior of the variables, identifying preferential directions in the space (anisotropy), or not (isotropy). For the case of the variables analyzed in the present work, a similar spatial behavior is observed for all directions (isotropy). From any analyzed direction, the variables relative to trip production and trip attraction as well as components 1 and 2 increase in value of the periphery in relation to the center. A total of 70 variograms was calculated. Figure 14 illustrates the ominidirecional variograms for the variables TRIP PRODUCTION BY TRANSIT, COMPONENT 1 and COMPONENT 2. Axis y represents the semivariance whereas axis x represents the distances. The points represent the data and the straight line represents the variance of the data of the considered population (Sill).



Trip Production by Transit	Component 1	Component 2

Figure 14 Experimental variograms

The calculation of the experimental variograms not only for the primary variable but also for secondary variables is necessary for: (1) modelling variograms of primary variables; (2) modelling variograms of secondary variables; (3) using the variograms parameters of primary variable for KED; (4) using the variograms parameters of secondary variable for ordinary kriging – the ordinary kriging of secondary variable is to achieve the values of this variable at co-ordenates that will be used for KED application.

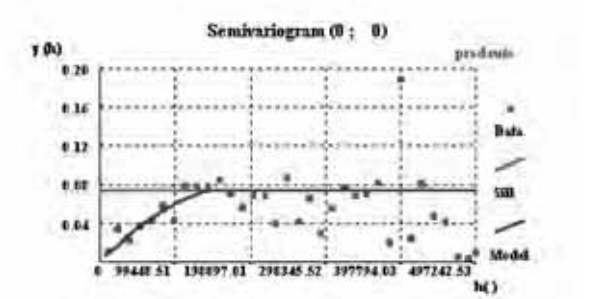
4.4.3 Modelling variograms

After the choice of parameters and the directions, all this information needs to be transposed for a representative general function. Thus, the adjustment of the experimental variograms to a general function is necessary. In such a way, for kriging (next steps), the parameters of adjusted curves of the variograms are considered.

For all the variables, the adjustment of the theoretical variogram was realized. The function was spherical type and its structure was defined: nugget effect or nugget (C0), Sill (C1) and range (a).

Concluded the adjustment of the theoretical variograms for all variables and directions, the main direction and the secondary direction could be selected. For the case of the variables analyzed in this work, it is possible to affirm that all of them are isotropic. Figure 15 presents

the theoretical ominidirencional variograms for the variables TRIP PRODUCTION BY CAR and COMPONENT 1.



Trip Production by Car	Component 1

Figure 15 Theoretical variograms

4.4.4 Ordinary Kriging – secondary variable

This step of the method has the objective of forecasting the values of the secondary variables (component 1 and component 2) in diverse unknown coordinates, considering the known values of these variables in centroids of Traffic zone of SPMA. Through the Ordinary Kriging, was possible to estimate different values in *n* coordinate (depending on the parameters adopted in kriging) for such variables.

Interpolation for Ordinary Kriging (Davis, 1987) was used to estimate unknown points, allowing to generate maps of spatial distribution of secondary variables (component 1 and component 2) considering SPMA. Table 8 shows some values found for Component 1 and Component 2 for 10 points (5 known co-ordinated and 5 unknown coordinates). A total of 40,000 values of points was found taking account the adopted grid. The interpolated values of the secondary variables in 40,000 points will be used in the subsequent step - application of the KED.

Table 8 – Interpolated values –Ordinary Kriging

Y	X	Component 1	Component 2	
-23548404	-46633691	-0.83782	1.5426	known
-23546230	**-46630342**	**-0.64392**	**0.90987**	**unknown**
-23543746	-46629525	-0.76221	0.76777	known
-23530932	**-46624387**	**-0.77226**	**0.66409**	**unknown**
-23503721	-46602324	-0.78698	0.56286	known
-23546585	**-46620484**	**-0.60186**	**0.6676**	**unknown**
-23551130	-46641581	-0.49554	0.56414	known
-23560147	**-46615414**	**-0.62083**	**0.37611**	**unknown**
-23543989	-46642640	-0.55879	1.6804	known
-23570800	**-46619454**	**-0.66575**	**0.20243**	**unknown**

4.4.5 Kriging with external drift (KED)

With KED application, it is possible to estimate the values of the primary variables (trip production and trip attraction) in 40,000 points based on the values of the secondary variables. Defining the grid (200 x 200) and using data of the theoretical variograms of the

primary variables and data of the secondary variables obtained with ordinary kriging, the interpolated values of primary variables were found.

Table 9 presents values of the variable TRIP PROCTION BY TRANSIT in known and unknown coordinates.

Table 9 – Interpolated values –KED

ID	Y	X	TriprodTRANSIT
Known	-46633691	-23548404	2,161.26
Unknown	**-46630492**	**-23547103**	**1,814.54**
Known	-46629525	-23543746	635.94
Unknown	**-46631520**	**-23543948**	**590.32**
Known	-46632324	-23553721	498.92
Unknown	**-46640381**	**-23551544**	**502.34**
Known	-46641581	-23551130	560.87
Unknown	**-46642031**	**-23542331**	**914,43**
Known	-46642640	-23543989	1405.9

4.4.6 Cross validation

The cross validation is a technique where for each point the value of the variable is sequentially estimated, considered unknown, calculating the error of real estimation for each point. Thus, a table of results of known points (374 centroids of traffic zones - some absent values had been excluded) with observed and estimated values, coordinated of the 374 zones and variance.

With the purpose of measuring the accuracy of the models, the following parameters had been calculated: correlation coefficient; Mean absolute error; Variance of error. Table 10 summarises the results of cross validation for all variables (trip production and attraction rates). For the purpose of comparison, the same adopted parameters had been calculated for the GWR application and are represented in the Table 10. Through these parameters, the authors could compare both techniques taking account trip generation analysis. The results show that GWR produces better correlations while KED minimize the mean and variance of estimation errors. However only KED permits densify, by interpolation,the values in the node of a designed network.

Table 10 – Parameters for comparing KED and GWR

		GWR	KED	GWR	KED	GWR	KED
Models		Coefficient of correlation		Mean absolute error		Variance of erros	
Trip Production	Transit	0.89	0.81	0.01	0.0005	0.02	0.0033
	Car	0.9	0.78	0.008	0.00001	0.001	0.0004
	Non-motorized	0.93	0.76	0.02	0.0004	0.03	0.003
	Industry	0.66	0.55	0.04	0.0003	0.05	0.001
	Commerce	0.84	0.79	0.009	0.0007	0.01	0
	Services	0.93	0.9	0.02	0.0001	0.001	0.01
Trip Atraction	Transit	0.9	0.8	0.01	0.0005	0.02	0.002
	Car	0.9	0.77	0.07	0.0002	0.001	0.01
	Non-motorized	0.92	0.75	0.01	0.006	0.004	0.002
	Industry	0.51	0.44	0.03	0.0004	0.01	0.003
	Commerce	0.88	0.76	0.008	0.001	0.01	0.0002
	Services	0.9	0.86	0.01	0.0009	0.02	0.0001

5. DISCUSSIONS AND CONCLUSIONS

The approach purposed by the authors - PCA + GWR and PCA + KED – has a lot of objectives: (1) to summarize the data; (2) to avoid multicollinearity problems for GWR application; (3) to combine different variables through one component, so the univariate analysis actually is a combination of original variables (components); (4) to purpose two different spatial ways to forecast trip generation; (5) to use a spatial interpolation that associate a **secondary information** (*component 1 – Age/socioeconomic household characteristics and component 2 – employment*) that represents variables that influence travel behavior; and (5) to adopt a set of procedures that allows the consideration of socioeconomic variables such as employment, household characteristics, income for traffic zone, etc. (represented through the components), urban trip production and attraction (dependent variables) and spatial correlation.

Analyzing the results of the global regression of the multivariate models, considering the parameters for measuring the accuracy of the models, all the models are considered well adjusted to forecast trip generation. Moreover, the variables (components) are considered statistically significant. Searching for some meaning in the global multivariate model results, one can mention that they corroborate with the literature on travel behavior/trip generation. A lively and diverse literature continues to investigate the complexity and variety of travel patterns in accordance with individual and household characteristics. The findings show, among other things, the important relations between the household structure, gender, car ownership, household income, and travel analysis (Mcguckin and Murakami 1999; Sarmiento 1996; Golob and McNally 1997; Bhat and Koppelman 1991).

Thus, Component 1 that represents Age/socioeconomic household characteristics is considered important or significant for almost all twelve models. Component 2 is important for trip production and trip attraction by travel mode. However, considering trip production and attraction for trip purpose, it is possible to note that the variable is not important to trip production but is significant for trip attraction (industry, commerce and service). These findings are relevant regarding the fact that employment is more related to trip attraction. Traffic Zones with high level of jobs and economic activities in general are centers of work-trip attraction.

As expected, Component 3, that represents High income, influences Trip production and attraction by car and by non-motorized travel mode. Trip Production/Atraction by car is directly proportional to Component 3. On the other hand, Trip Production/Atraction by non-motorized travel mode is inversely proportional to Component 3. The literature confirms that higher income households have higher car ownership rates and, consequently, generate more motorized trips by individual travel mode, less non-motorized trips and longer travel distances (Hanson and Hanson, 1981; Mitchell and Town, 1977; Zegras and Srinivasan, 2007). The statistically findings of Component 4 corroborate the fact that population is highly related to trip production rates.

Nevertheless, The GWR application is more useful because it allows us to map the statistics presented. So, the main advantage is related to the visualization of spatial travel patterns of each one of the concerned variables. Regarding the spatial distribution of the parameters of the variables it is possible to observe a common spatial pattern: the values are higher at the center and they decrease in the periphery. Nonetheless, each component presents a specific pattern. Component 1, for example, shows the following spatial pattern: the values increase with a buffer/radial pattern in the center region.

For comparing the two spatial data analysis techniques (KED and GWR), it was necessary to choose only one of the components as independent variable to integrate the GWR model. So, the subsequent analysis was the univariate models of GWR. Even if it implied loosing some accuracy, the authors had to generate univariate models taking into account the KED application. For KED application it is possible to use only one variable as secondary variable (GeoMs software allows only one secondary variable). It is important to mention that each component is a combination of many original variables; therefore, this analysis was essentially not univariate.

The twelve models in this analysis are related only to Component 1 or Component 2. As mentioned before, Component 2 has an affect on trip attraction per work-trips attraction. Table 7 shows these statistics. The values of the coefficient of correlation were used later to compare the quality of forecasting of both techniques.

With KED application, it is possible to estimate the values of the primary variables (trip production and trip attraction), not only in known centroids but also in 40,000 points based on the values of the secondary variables. Some adjustments have to be considered regarding the fact that the data used here was not geostatistical data (observations associated with a continuous variation over space).

Comparing the two techniques, considering trip production by transit, for example, the GWR presented greater value for the correlation coefficient (0.89 for GWR and 0.81 for the KED). The KED presented relatively lesser value for the average of absolute error (0.01 for GWR and 0.005 for KED). Both techniques could be considered well adjusted to forecast trip generation. They presented good results measurements and models related to socio-economic factors (Component 1 or Component 2), trip variables, and spatial correlation.

For both techniques, trip production and attraction in industry have the lower accuracy levels. This could be justified by the independent variables used. It is possible that another set of variables could be more correlated with these two dependent/primary variables. Land use variables or High level of industry jobs traffic zones or economic activity distribution could be more efficient.

The main benefit concerning KED application was not the estimation but the spatial interpolation. Even so the KED estimations could be considered adequate (high values of coefficient correlation and low values of errors), the advantage of KED in detriment of GWR was the possibility of estimating values in diverse unknown coordinates. This could be very

useful, for example, to generate data related to trip generation for later analysis (modal choice, for example). Besides, the use of KED + PCA is interesting because the technique can accomplish the interpolation of trip related variables regarding a secondary information (components) that combines original variables that influence travel forecast models.

The benefit of GWR in detriment of KED was the possibility of visualizing the findings in the study area. Thus, it becomes possible to see spatial patterns or to conclude about differences of variables influence at center or in some locations, to interpret the spatial distribution of values of the intercept terms, to observe local parameter estimates for each surface across the study region.

6. REFERENCES

Anselin, L. (1992) Space and applied econometrics: introduction, Regional Science and Urban Economics 22, 307-16.

Anselin, Luc and Daniel A. Griffith (1993). Operational Methods of Spatial Data analysis (Oxford, Oxford University Press)

Charlton, M.; Fotheringham, S. and C. Brunsdon (2005) Geographically Weighted Regression. ESRC National Centre for Research Methods NCRM Methods Review Papers NCRM/006. http://eprints.ncrm.ac.uk/90/1/MethodsReviewPaperNCRM-006.pdf. (February, 1st).

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

Davis, B. (1987). Uses and Abuses of Cross-Validation in Geostatistics. 1997, Mathematical Geology, vol. 19, nº 3, pp 249-258.

Golob, T.F., McNally, M.G., (1997). A model of household interactions in activity participation and the derived demand for travel. Transportation Research 31B, 177–194.

Goovaerts, P (1997). Geostatistics for natural resources evaluation. Oxford University Press, New York, 512 p., 1997.

Hair, J.F.; Anderson, R.E; Tatham, R.L and W.C. Black (1998). Multivariate Data Analysis. 5ª ed. Prentice-Hall. Upper Saddle River, New Jersey, 730p.

Hanson, S. and Hanson, P. (1981) The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic Geography*, v. 57, p. 332-347

Li, T.; Corcoran, J.; Pullar, D; Robson, A. and Stimson, R. (2008) A Geographically Weighted Regression Method to Spatially Disaggregate Regional Employment Forecasts for South East Queensland. Appl. Spatial Analysis (2009) 2:147–175.

Mcguckin N. and Murakami E. (1999) Examining trip-chaining behavior: Comparison of travel by men and women. *Transportation Research Record*, ISSN 0361-1981, n 1693, p. 79-85.

Mitchell, C. G. B. and Town S.W. (1977) Accessibility of various social groups to different activities**.** *Transport and Road Research Laboratory, Report 258*, England.

Prompong, L and S. Soralump (2009) Spatial data analysis by using geostatistics evaluate the undrained shear strength of soft bangkok clay. Geotechnical Infrastructure Asset Management. EIT-JSCE Joint International Symposium. Thailand.

Sarmiento, S. (1996) Household, Gender, and Travel. Women's Travel Issues. Proceedings from the Second National Conference, Baltimore.

Tobler, W. (1979). Cellular geography. In S. Gale and G. Olsson (Eds.), Philosophy in Geography, pp. 379-86 (Dordrecht, Reidel).

Zegras, P.C. and Srinivasan, S. (2007) Household Income, Travel Behavior, Location and Accessibility: Sketches From Two Different Developing Contexts. T*ransportation Research Record No. 2038.*