ESTIMATION OF POPULATION DENSITY USING HIGH RESOLUTION REMOTE SENSING DATA AND SPATIAL REGRESSION TECHNIQUES: A CASE STUDY IN SÃO PAULO CITY (BRAZIL)

TESSIO NOVACK
HERMANN J. H. KUX
ANTONIO MIGUEL V. MONTEIRO
CAROLINA PINHO

Instituto Nacional de Pesquisas Espaciais – INPE
Departamento de Sensoriamento Remoto, São José dos Campos – SP
{tessio, hermann}@dsr.inpe.br, {miguel, carolina}@dpi.inpe.br

RESUMO – Partindo da premissa de que o planejamento urbano visa o bem-estar dos cidadãos, questões como "quantas pessoas vivem na cidade?" e "onde elas moram?" tornam questões chaves. Neste trabalho utilizamos métricas de paisagem geradas pelo programa FragStats como variáveis independentes para a estimação da densidade populacional de setores censitários na megalópole de São Paulo, Brasil. As métricas foram calculadas sobre uma imagem do sensor QuickBird II classificada de forma supervisionada pelo classificador Maxver. A exatidão temática e geométrica da classificação foi avaliada qualitativamente por inspeção visual. Modelos de regressão simples foram gerados e testes estatísticos formais aplicados. A dependência espacial dos resíduos de cada modelo foi analisada através da visualização dos mapas LISA. Em seguida, modelos de regressão espacial foram testados e uma significativa melhora obtida em termos de redução da dependência espacial dos resíduos e de aumento do poder de estimação dos modelos. O uso de variáveis dummy em modelos de regressão linear simples também se mostrou como uma alternativa viável para se obter estes dois efeitos. Os resultados para a área teste dão evidências de que algumas métricas de paisagem obtidas sobre imagens classificadas de sensores de alta resolução quando usadas como variáveis independentes em modelos de regressão espacial podem estimar satisfatoriamente a densidade populacional.

ABSTRACT - Assuming that urban planning aims the optimization of urban functioning and the well-being of citizens, questions like "how many people are living in the city?" and "where do they live?" become key issues. In this work we utilized landscape metrics generated by the FragStats software for the estimation of population density out of census sectors in the mega city of São Paulo, Brazil. The metrics were calculated over an image from the QuickBird II sensor classified by the Maxver algorithm. The accuracy of the classified image was analyzed qualitatively. Ordinary linear regression models were generated and formal statistical tests applied. The residuals from each model had their spatial dependency analyzed by visualizing their LISA Maps. Afterwards, spatial regression models were tried and a significant improvement obtained in terms of spatial dependency reduction and augmentation of the prediction power of the models. For the sake of comparison, the use of dummy variables was also tried and has shown to be a suitable option for eliminating spatial dependency of the residuals as well. The results proved that some landscape metrics obtained over high resolution imagery classified by simple supervisioned methods can predict well the population density at the area under study when used as independent variables at spatial regression models.

1 INTRODUCTION

The estimated total population for the municipality of São Paulo (Brazil) on year 2007 was of 10,886,513 inhabitants, living in an area of 1,523 km^2 (IBGE, 2000). Considering that the majority of these people live in the urban area, which corresponds only to a fraction of the municipality, we get an idea on the size of the social and economic phenomenon of São Paulo city. Assuming that urban planning aims the optimization of urban functioning and the well-being of citizens, questions like "how many people are living in the city?" and "where do they live?" become key issues. The traditional approaches to estimate total population or population density are mainly based on field works. Those methods are labour-intensive, time-consuming, costly and also encounter difficulties in updating the database. Population projections for inner-city domains are more and more

required to allow the evaluation and monitoring of social programs, constituting the denominator of several social indicators established periodically (JANUZZI, 2004). Nevertheless the census in Brazil is decennial and presently there is no flexible and low cost methodology available to estimate the number and the density of inhabitants during an inter-census time frame. This problem is worsened by the fact that large cities in developing countries present an explosive growth dynamic with constant changes such as urban sprawl, increase on the number of vertical residential dwellings throughout the city, augmentation of the population living in slums and etc. In this context high resolution remote sensing data and spatial regression techniques are powerful tools, indicating simple solutions of relative easy implementation to overcome the problems mentioned.

1.1 The use of remote sensing data for total population and population density estimation

Many methods for population estimation have been reported in the GIS and remote sensing literatures. The statistical modeling approach is more interested in inferring the relationship between population and other variables for the purpose of estimating the total population for an area. This approach is originally designed to estimate the inter-census population or population of an area difficult to enumerate. Satellite images with medium spatial resolution (15 – 30 m) and linear regression techniques were already used for the construction of population estimation models using as explanatory variables (i.e. independent variables) the average spectral radiance (or reflectance) values associated to image pixels at several bands of a sensor (HARVEY, 2002; REIS, 2005). Lo et al. (2006) used an impervious surface fraction image obtained by spectral unmixing techniques, GIS data and regression techniques to estimate total population with ETM+/Landsat-7 data. In more contemporary days the increasing availability of high spatial resolution imagery represents a new paradigm for the estimation of population and population density, especially in urban areas. Liu et al. (2002) explored the relation between texture of an IKONOS image measured by spatial metrics and semi-variograms of population density in census sectors using linear regression methods. Herold (2006) used high resolution imagery and linear regression techniques to estimate the population density of Santa Barbara city (USA) using as independent variables spatial metrics calculated by the FragStat software. Despite the remarkable contribution of these works, there have been few discussions on the spatial dependency of linear regression residuals. This study is innovative in the sense that it intends to consider the spatial dependency structure of the regression model residuals.

1.2 Objectives

The objective of this paper was to construct potential models for the estimation of population density on 212 census sectors on the city of São Paulo using as independent variables spatial metrics obtained by the FragStats 3.3 software over a classified QuickBird image. The prediction power of the proposed models was compared on an index basis before and after considering the spatial dependency of the residuals.

2 TEST AREA

The test area has a size of 49 km^2 and is located at the border of districts Vila Sônia, Vila Andrade, Morumbi, Campo Limpo and Santo Amaro (figure 1). The site's central geographical coordinates are: W 46^o 43' 30''and S 23^o 36' 30''. This specific location was chosen as test area for the fact that it is very heterogeneous in terms of land cover classes. Different types and sizes of roofs and vegetation covered areas as well as parking lots, swimming pools, streets and avenues and bare soil areas are present at the site. These urban entities configure industrial and residential occupation of different structures, densities and social-economic levels. This area also contains the second largest slum of São Paulo (Paraisópolis slum) as well as vertical condominiums, and entire blocks occupied by houses of very high standards.

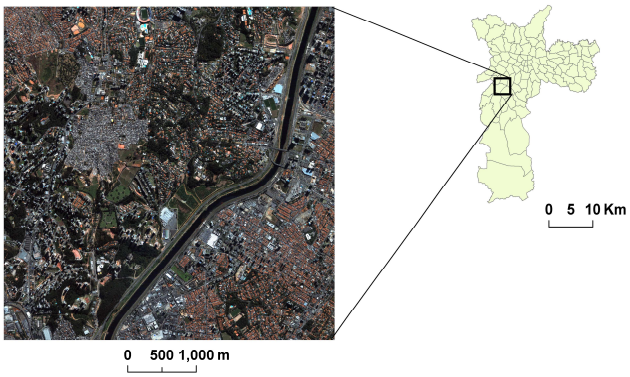

Figure 1 – QuickBird image of the area under study and the municipality of São Paulo on the upper right.

3 MATERIALS

The following materials were used for this study: 212 census sectors in vector format with its socio-economic and demographic data from the Demographic Census of IBGE for the year 2000, a QuickBird image from year 2002 (Table 1) and the following softwares: (1) SPRING 4.3.3 for the automatic classification of the QuickBird image; (2) FragStats 3.3 for the calculation of metrics for both the census sector and all the land cover classes in it; (3) Statistica 6.0 for the ordinary least square regression model construction and formal diagnostics

tests; (4) GeoDA 0.9.5.1 (Beta version) for the spatial regression model construction and diagnostics tests.

Table 1 – Resolutions of the QuickBird II sensor.

Mode	Spectral Resolution (nm)	Spatial Resolution (cm)	Radiometric Resolution
Panchromatic	450 – 900	62 to 71	11 bits
Multi-spectral	450 – 520	244 to 288	
	520 – 600		
	630 – 690		
	760 – 900		

4 METHODOLOGY

The methodological steps followed to establish spatial regression models for the estimation of population densities in São Paulo were: (1) thematic classification of the QuickBird image, (2) generation of independent variables using FragStats 3.3 software package, (3) logarithm transformation of the dependent variable population density (4) selection of independent variables, (5) selection of potential ordinary least square models, (6) tests of quality of these models (7) analysis of spatial dependency of residuals from the models, (8) running of the spatial regression models and (9) considerations on the constructed model and its appropriateness.

4.1 Image classification

Image classification is a process in which image pixels are grouped and associated with concepts based on their digital numbers for the *n* bands of a sensor (MATHER, 2004). The QuickBird image was classified by the supervised Maximum-Likelihood method (MAXVER) (RIBEIRO; CENTENO, 2001) considering the following classes: *ceramics roofs*, *dark roofs*, *vegetation*, *streets*, *shadow* and *non-classified*.

This was done on the SPRING 4.3.3 image processing free software. We took care that the classes had low confusion among each other when collecting samples for the supervised classification. The classification accuracy was evaluated qualitatively by visual inspection.

4.2 Generation of independent variables

All independent variables were generated by the software FragStats 3.3. It generates spatial metrics for the landscape as a whole (i.e. in this case every census sector of the test area), for each class belonging to each census sector and for every patch in it. We did not use patch metrics because it could not be associated directly to population density of the sectors. At the end, sixty-four independent variables were obtained for both the census sectors and the thematic classes in it. A logarithm transformation on the dependent variable population density was applied to make its distribution closer to the normal one (the purpose of that was to enable eventual confidence intervals calculations for the predictions of the model). This operation was successful and has also augmented the correlation between the independent variables and the dependent one. All independent variables with a correlation below 0.4 with the dependent variable (i.e. population density) were discarded. Following we cared that, among the other ones, no pair of independent variables had a correlation above 0.7. This was done to avoid multi-colinearity problems (NETTER, 1974). That step eliminated most of the sixty-four variables available and made the selection of the model much simpler. Due to the fact that the metrics generated by FragStats were not made specifically for urban applications, all four of the selected variables were quite

Figure 2 – QuickBird image classified by the Maxver algorithm.

simple. These variables were: number of polygons classified as ceramics roofs on the census sector, percentage of class dark roof of the census sector, aggregation index (1) of class streets (i.e. aggregation index equals the number of like adjacencies involving the corresponding class, divided by the maximum possible number of like adjacencies involving the corresponding class, which is achieved when the class is maximally clumped into a single, compact patch) and the patch density of class vegetation (i.e. the number of patches of a certain class divided by the total sector area).

$$\text{Aggregation Index} = \left[\frac{g_{ii}}{\max. \rightarrow g_{ii}} \right] \quad (1)$$

where,
g_{ii} = number of like adjacencies (joins) between pixels of patch type (class) *i* based on the single-count method.
$\max. \rightarrow g_{ii}$ = maximum number of like adjacencies (joins) between pixels of patch type (class) *i* based on the single-count method.

4.3 Selection of the classic regression model

We judged that some of the variables could explain well the population density so we arbitrarily took to the quality tests step only the models that contained those variables and in which all variable coefficients were statistically significant. Table 2 shows the variables at both models with its respective R^2. The formal quality tests for the models showed that the normality of residuals was acceptable for both models (i.e. p-value below 0.05 for the Kolmogorov test) but non-constancy of variance was a problem for model M1 as showed the Breusch-Pagan test (i.e. calculated value higher than tabled value).

Table 2 – Models chosen for the classic linear regression.

Model	Variables	R^2
M1	• Number of polygons from class ceramics roofs; • Percentage of class dark roof; • Aggregation Index of class streets; • Patch Density of class vegetation.	0.75
M2	• Percentage of class dark roof; • Number of polygons from class ceramic roofs.	0.65

The area under study is very heterogeneous in terms of size of census sector and land cover classes in it. Despite of that we did not carried out outlier analysis which would certainly increase the R^2 but in the other hand would not consider the complexity of the test area reality.

4.4 Spatial analysis of the residuals

For the analysis of the spatial dependency of the residuals we set up a distance-based neighborhood matrix. The distance threshold was correctly defined when generating and fitting of the semi-variogram. The range detected in this operation was of 1,700 m. The Global Moran indexes calculated by the GeoDA software were low (0.0427 for M1 and 0.0825 for M2) but statistically significant for both models. That would already justify the use of a spatial regression model. Nevertheless, the strongest evidence of the existence of spatial dependency of the residuals for both models was the LISA maps which showed very clear spatial groupings at both models (Fig. 2). The next step was then to proceed to test which spatial regression model would be the most suitable for each model.

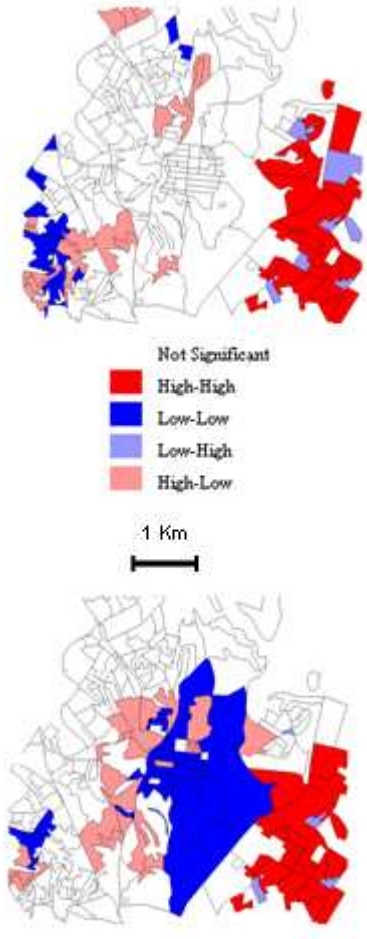

Figure 3 – LISA maps of models M1 (above) and M2 (below) for ordinary least square regression.

4.5 Spatial regression model construction

The criterion to choose the spatial regression model presented by Anselin (2004) which considers de Lagrange Multiplier Test indicated as the most adequate the Spatial Lag Model for M1 and the Spatial Error Model for M2.

The diagnostics for these spatial regression models indicated that the non-constancy of variance was solved for model M1 and the reduction of index Akaike, Log Likelihood and Schwarz attested the improvement of the models which justifies the inclusion of the new variable, i.e. the spatial auto-correlation parameter (ANSELIN, 2004) (Table 3).

The global Moran indexes, although still statistically significant, reduced at both models (-0.017 for M1 and 0.0075 for M2) and the LISA maps showed that the spatial dependency at model M1 was almost eliminated, while it was clearly reduced at M2 (figure 4). As the model M2 still presented some spatial dependency and non-constancy of variance problems, we decided to try the inclusion of dummy variables and compare it with the spatial regression model in terms of spatial dependency elimination. Two dummy variables were generated. This operation gives the model three different

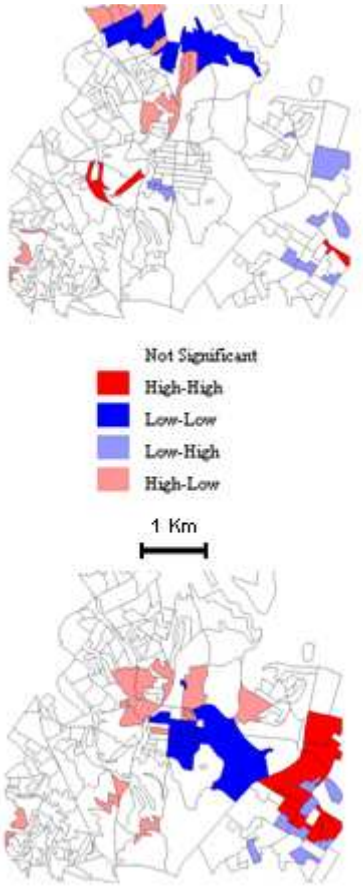

Figure 4 - LISA maps for the two spatial regression models – M1 above and M2 below.

coefficient subsets depending on the spatial domain the cases are located (defined by the dummy variables).

The cases associated with values 1 and 0 at both dummy variables were selected over the classical regression LISA map (figure 5). The ordinary linear regression model was again applied but this time for local conditions.

Table 3 – Results obtained by classic global regression and by local conditions for the model M.

Regression Model Type (M1)	Indexes R^2	Akaike	Global Moran	Breusch-Pagan
OLS	0.74	366	0.042	11.45
SLM	0.75	359	-0.017	8.39

Table 4 – Results obtained by classic and spatial regression for model M2.

Regression Model Type (M2)	Indexes R^2	Akaike	Global Moran	Breusch-Pagan
OLS	0.65	423	0.08	14.13
SEM	0.68	411	0.007	11.41
OLS - LR	0.77	339	0.01	17.64

We could as well have multiplied the beta coefficients on all possible combination of two coefficients for the improving of the prediction power of the model with dummy variables. This would imply in different offsets and gains for each spatial domain (NETTER, 1974). But as the aim was to come up with a model as simpler as possible (for operational purposes and for the exporting of the model to other areas) this was not done. The analysis of the prediction power of the ordinary least square regression model with dummy variables had the purpose of evaluating the necessity of automatic selection of dummy variables (yet to be implemented) for the case of the city of São Paulo in areas with complex spatial dependency.

5 RESULTS

An increase of the prediction power on model M1 was achieved by the insertion of the Spatial Lag parameter. The non-constancy of variance which was a problem for M1 was solved when using the Spatial Lag regression as it shows the calculated Breusch-Pagan values on table 4. On model M2 the Spatial Error Model has slightly improved the R^2 and decreased the Akaike index. On the other hand, the calculated Breusch-Pagan values show that non-constancy of variance is still a problem even at the model with the dummy variables. The insertion of these two variables has enhanced the prediction power as the R^2 and the Akaike indexes show on table 5.

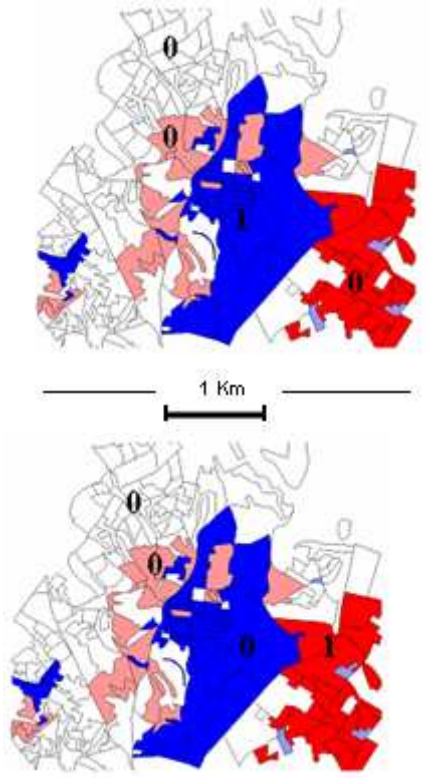

Figure 5 – Selection of cases associated with the value 1 at the two dummy variables.

6 CONCLUSION

It was proven that population density can be relatively well estimated by the use of spatial metrics calculated over a classified high resolution image. More accurate results obtained in the classification procedure could have considerably augmented the prediction power of the models proposed. An object-oriented classification approach should be tried in future works.

The analysis of the ordinary least square model residuals has shown that although numerically subtle the spatial dependency of the residuals is clearly observed at the LISA maps on the two proposed models. The spatial regression approach was numerically justified and compensated as the R^2 and Akaike index values show. The use of dummy variables has also proved to be potential for population density estimation at the test area. When choosing a regression model to be used periodically for population density estimation by the urban planning organisms one must take into account limitations and advantages of each of the potential models. For this reason, the aim here was to create a model simple but accurate enough which could be applied on other areas and validated in other occasions in the future.

REFERENCES

ANSELIN, J. L. **GeoDA User Guide**, 2004. Available at <https://www.geoda.uiuc.edu/support/help/references.html>. Accessed on 18th December 2007

HARVEY, J. T. Estimating Census District Populations from Satellite Imagery: Some Approaches and Limitations. **International Journal of Remote Sensing**, V. 23, n. 10, p. 2071-2095, 2002.

HEROLD, M. Population Density and Urban Texture: A Comparison Study, **Photogrammetric Engineering and Remote Sensing**, 2006.

IBGE. Fundação Instituto Brasileiro de Geografia e Estatística. **Censo demográfico de 2000**. Rio de Janeiro: IBGE, 2002b. CD-ROM.

JANNUZZI P. M. 2004, **Projeções Populacionais para Pequenas Areas: Métodos e Aplicações**. Availble at <http://www.abep.nepo.unicamp.br/docs/anais/pdf/2000/Todos/prot20_1.pdf>. Accessed on 18th 2007.

LIU, X. H.; CLARKE, K. Estimation of Residential Population Using High-Resolution Satellite Imagery. In: **International Symposium on Remote Sensing of Urban Areas**, 3, 2002, Istanbul, Turkey. Proceedings… Istanbul, Turkey, 11-13 june, 2002.

LO, D.; WENG, Q.; LI, G.; Residential Population Estimation Using a Remote Sensing Derived Impervious Surface Approach. **International Journal of Remote Sensing** V. 27, No. 16, p. 3553–3570, 2006.

MATHER, P. M. **Computer processing of remotely-sensed images: an introduction**. 3 ed. Chichester: John Wiley & Sons, 324 p. 2004.

NETTER J., WASSERMAM W., **Applied Linear Statistical Models**. Irwin, Homewood, 1974.

REIS, I. A. Estimação da População dos Setores Censitários de Belo Horizonte Usando Imagens de Satélite. **Anais XII Simpósio Brasileiro de Sensoriamento Remoto**, Goiânia, Brasil, INPE, p. 2741-2748, 2005.

RIBEIRO, S. R. A.; CENTENO, J. S. Classificação do uso do solo utilizando redes neurais e o algoritmo MAXVER. In: **Simpósio Brasileiro de Sensoriamento Remoto**, 10., 2001, Foz do Iguaçu. **Anais...** São José dos Campos: INPE, 2001. Available at <http://urlib.net/dpi.inpe.br/lise/2001/09.20.17.56> Accessed on 18th December 2007