Spatial clustering with different geographical scales

Alberto Augusto Eichman Jakob

Introduction

This paper aims to provide a clustering methodology which uses not only tabular information, but spatial information also in order to create more homogeneous areas in certain space. The local Moran's I can be used as a spatial autocorrelation indicator, and gives information about the spatial unit also in terms of its neighbors.

Statistical techniques are used here, like factor analysis, and spatial analysis techniques, like the Moran's I, to define homogeneous areas according to certain characteristics.

The smaller spatial unit, and the most easy to obtain and use for the whole country, is the census tract, available on the Demographic Censuses. The census tracts are then the better option to proceed with a clustering procedure, although frequently there isn't much tabular information, comparing with the big number of census variables available on the sample questionnaire. If the tracts have the more detailed geographical information, more tabular information are given by less detailed geographical information, like municipalities, counties or districts.

If the idea is to create segregation areas, and use these boundaries to study other kind of variables, the areas can be created with intra-municipality data or shapefile, and adjusted with an analysis for the tracts.

In order to do that, a study case was made with data from an emerging metropolitan area of Brazil, the Baixada Santista Metropolitan Area (BSMA), which includes nine municipalities and about 1.6 millions of inhabitants living in a coastal area of the state of São Paulo.

Methodology

First of all, some tabular data from the demographic census of 2000 was selected for census tracts of the BSMA, denoting educational and income characteristics of the people or household heads that lives in the tracts. Same information was selected for weighted areas, which represent aggregations of census tracts and are available for all variables of the Brazilian demographic census of 2000.

Table 1 shows these selected variables, as well as the result of the factor analysis made in order to reduce this five variables into one component or factor, which is the linear combination of the variables. This table denotes that almost 70% of the data was represented with this component for census tracts, and 88% for weighted areas, instead of five variables, which is good.

This table also shows that the years of education is the most important variable of the component, followed by the household heads with a primary education. Observing the signal of the component is possible to say that if the component is high, the education is poor and the income is low. The best situation is with low component values for each census tract.

Table 1
Component matrix of the factor analysis.
Census tracts and weighted areas, BSMA, 2000.

Variable	Census Tracts	Weighted Areas
% illitarate people (7 to 14 years-old)	0.5972	0.8366
% illitarate household heads	0.7870	0.9317
% household heads with primary education (until 4 years)	0.8680	0.9820
Mean years of study of the household heads	-0.9180	-0.9876
Mean monthly income of the household heads	-0.8379	-0.9346
% variance explained:	**69.2**	**87.6**

Source: FIBGE, Brazilian Denographic Census of 2000.

The factorial score, given by this analysis, was the variable used on the local Moran's I procedure for the different geographic scaled, which resulted on Maps 1 for the tracts and Map 2 for weighted areas.

Map 1 brings the census tracts allocated into categories according to the component of the factor analysis. The first class represents census tracts with high values surrounded by tracts with high values also (557 tracts). This is the worst condition in terms of education and income.

The second class shows 822 tracts with low values surrounded by tracts with low values (the best condition). The third group represents tracts with people in good condition with neighbors in not so good condition, and the forth tracts with persons with poor education and income between tracts with persons in good condition.

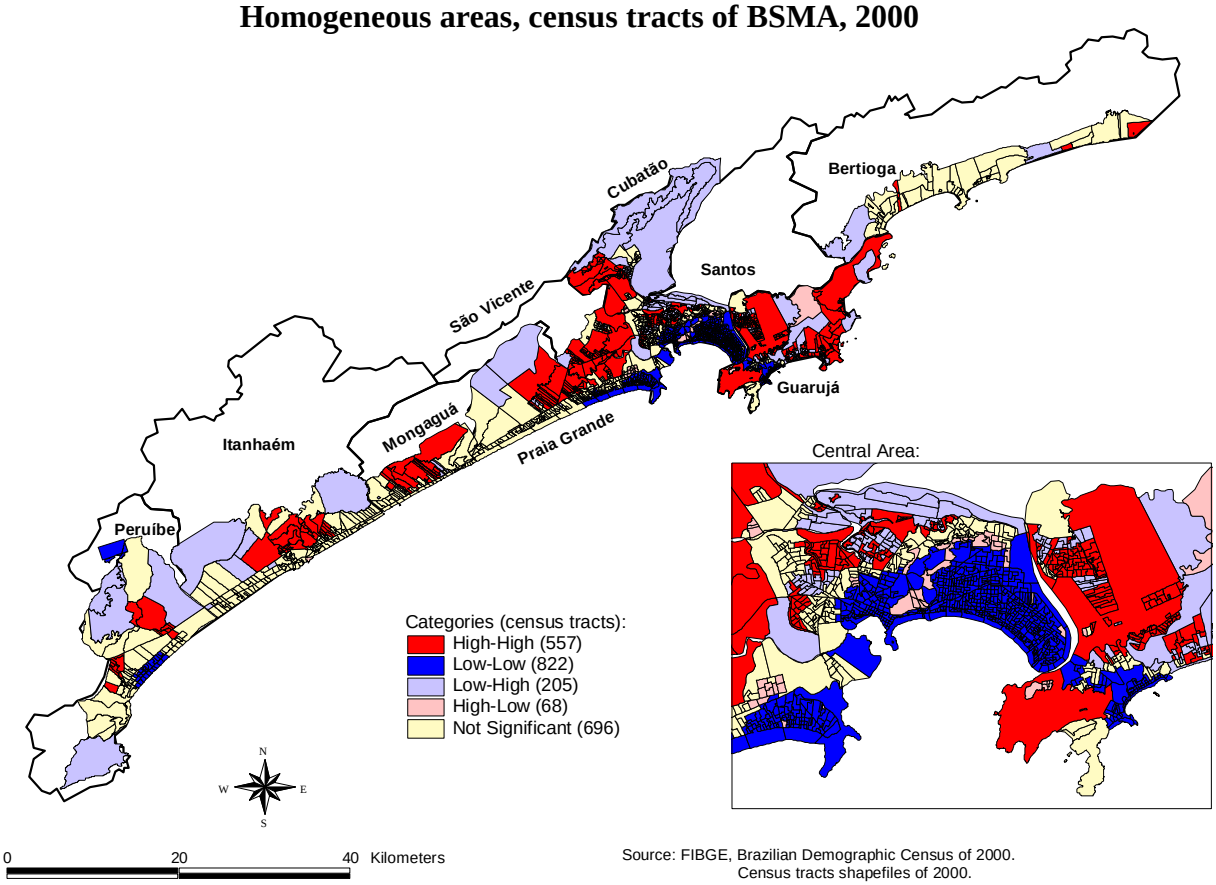
Finally, the last class (not significant) represents tracts with much heterogeneity, and is very difficult to analyze this data in the model. The data is not statistically significant.

It is also very important to denote that only the urban census tracts were used in this analysis (colored areas). Rural areas were not analyzed since have only 0.5% of the population of the BSMA, and they are basically protected areas near mountains and state parks.

The analysis for weighted areas is expressed on Map 2. In order to adjust this result with the tracts, considered the best division, the classification of the tracts was correlated with the classification of the weighted areas.

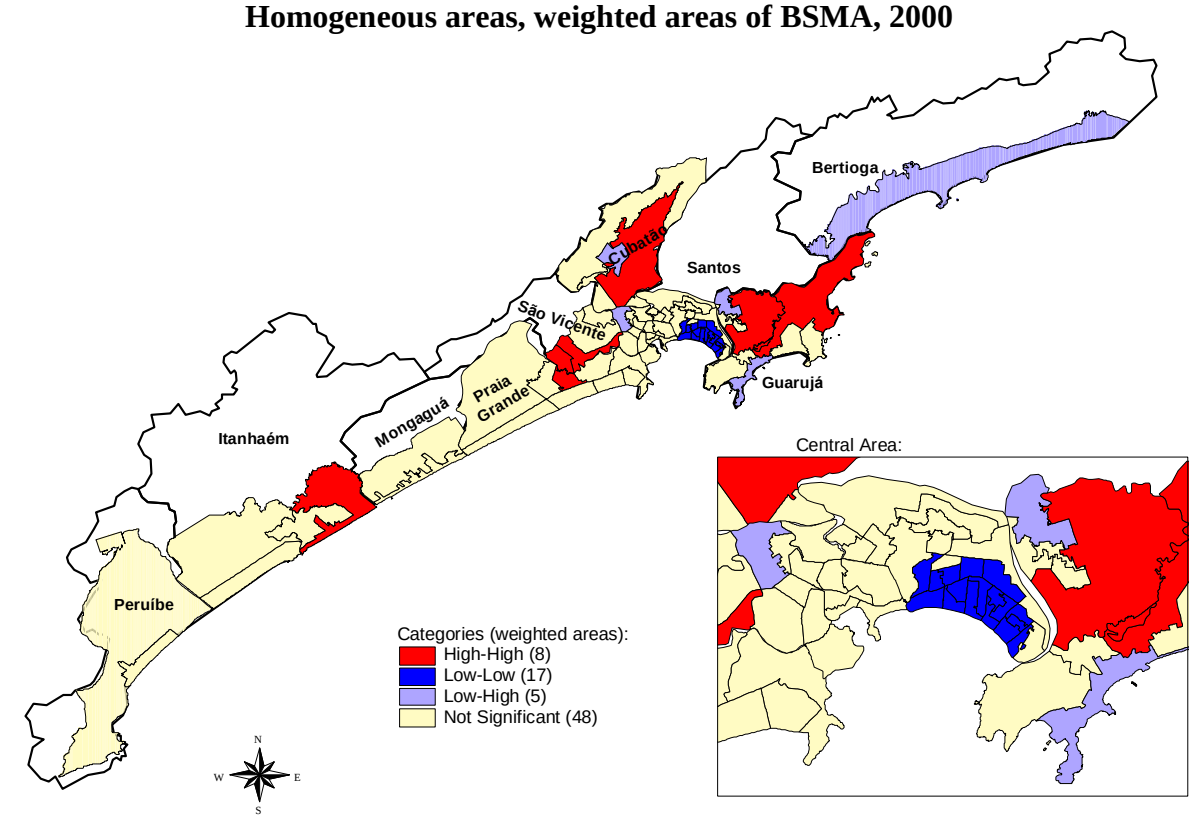
The population living on the tracts and weighted areas are related on a table like Table 2, where each weighted area is classified according to its homogeneous area in terms of tracts analysis also. The percentage of population in each weighted area for each category are compared in terms of both scale and adjusts are made if necessary.

Map 1
Homogeneous areas, census tracts of BSMA, 2000



Source: FIBGE, Brazilian Demographic Census of 2000.
Census tracts shapefiles of 2000.

Map 2
Homogeneous areas, weighted areas of BSMA, 2000



Source: FIBGE, Brazilian Demographic Census of 2000.
Census tracts shapefiles of 2000.

Table 2
Correlation between census tracts and weighted areas

Weighted Areas (WA)		Census Tracts			
		High-High	**Low-Low**	**Low-High**	**High-Low**
High-High	**$WA_1 \ldots WA_n$**	%...%	%...%	%...%	%...%
Low-Low	**$WA_1 \ldots WA_n$**	%...%	%...%	%...%	%...%
Low-High	**$WA_1 \ldots WA_n$**	%...%	%...%	%...%	%...%

Conclusions

It is very difficult to discuss something in an extended abstract due to the restrict number of pages. So, a broader analysis will be presented on a full version of this paper if approved. But it is important to show that is possible to delineate segregation areas by these homogeneous areas, and to measure them, in terms of known measures, like kilometers or miles.

The most interesting thing of this spatial analysis is that it is possible not only to show information about the census tract, but also for its neighbors. The clustering procedures with this spatial variable become much more accurate.

These techniques were tested for other regions with very good results also. And the final homogeneous areas can be used on other statistical models as a spatial variable. Many studies deal with the place of residence assuming important roles in terms of labor market and access to public services.

In a possible acceptance of this paper, much more can be say in terms of the techniques, as well as analyzed, like the almost concentric zones (in terms of census tracts) and some poor residences of the central area, the concentration of economic activities and services in central areas, the impacts of the regional geography on the analysis (rivers, flooded areas, the Atlantic Ocean, mountains, protected areas) among others.

References

CUNHA, J.M.P. da; JAKOB, A.A.E. Socio-spatial segregation and the labor market in emerging metropolitan areas in Brazil: the case of Campinas, State of São Paulo. IN: Urban Segregation and Labor in the Americas. Seminário organizado por Organizado por Llilas Cluster on Social Policy, University of Texas at Austin, 14-15 de Fevereiro, 2008.

JAKOB, A.A.E. The unequal spaces beyond the official limits: The intra-urban dynamic of São Vicente Island in 1990s. IN: Conferência Internacional de População - IUSSP, 25, 2005. Tours, França. **Anais...** Tours: IUSSP, 2005.

JAKOB, A.A.E. The Intra-Urban Dynamic based on Spatial Statistics:A case study of a Brazilian Municipality in 1990s. IN: 2004 Annual Meeting of the Population Association of America – PAA. Boston, EUA. **Anais…** Boston: PAA, 2004.