Metropolitan Analysis using Spatial Microsimulation combined with Skater Regionalization Methods: An Study for the Paraíba Valley and North Coast Metropolitan Region-SP

Gabriela C. Oliveira[1], Tathiane M. Anazawa[1], Antônio Miguel V. Monteiro[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE) Av. dos Astronautas, 1758 – 12227-010 – São José dos Campos – SP – Brasil

{gabriela.oliveira, tathiane.anazawa, miguel.monteiro}@inpe.br

Abstract. *This paper analyzes the distribution of socio-occupational groups in Subregion 4 of the Paraíba Valley and North Coast Metropolitan Region (in Portuguese: Região Metropolitana do Vale do Paraíba e Litoral Norte – RMVPLN) using spatial microsimulation techniques. To fulfill the proposed objective, the IPF technique was used to obtain, through the 2010 Demographic Census microdata, spatial microdata in the territorial unit of census tracts. After obtaining the data, the Skater regionalization technique was applied to obtain more homogeneous socio-occupational groups. It was possible to identify 15 homogeneous groups, five of them with larger numbers of census tracts. Overall, the proposed socio-occupational categories, studied at an intra-urban scale, allowed for highlighting the social structure on a subregion of the newest Metropolitan space in São Paulo. Unfortunately, although it is still a preliminary study, it points for degrees of inequalities consistently spatially segregating the less privileged socioeconomic groups of the population.*

1. Introduction

Focused studies of poverty identification and population classification by income range dominate much of Brazil's social policy discussions [Neri and Carvalhaes, 2008], [IPEA, 2008]. Although income plays an important role in the insertion of individuals in the market of goods and products, it cannot be seen as the sole delimiting factor of the position of individuals in the hierarchy of a society.

As an alternative to the stratification of the population according to income ranges, the literature proposes typologies based on broader concepts, which would be approximations more consistent with the class behavior of a Society [Rose e Harrison, 2007], [Rose e Pevalin, 2005], [Jannuzzi, 2003], [Quadros e Maia, 2010].

Occupations began to play an essential role in shaping the structure of modern capitalist societies. Identifying the socio-occupational structure of a society enriches social analyzes, whether related to exclusion, inequality, mobility, health, consumption, among others [Quadros e Maia, 2010].

Socio-occupational stratification is, however, a methodological challenge that is subject to the complexity of the theme and the limitations imposed by the data. To help understanding the complexity of social relations in Subregion 4 of the Paraíba Valley and North Coast Metropolitan Region (in Portuguese: Região Metropolitana do Vale do Paraíba e Litoral Norte – RMVPLN), this paper proposes to analyze the distribution and composition of its socio-occupational structure. The analyzes are supported by on a

proposal to stratify Brazilian society based on the structure of occupations of the labor market used by IBGE (2010).

These analyzes start from the premise that relatively homogeneous social groups can be obtained from the insertion of individuals into the labor market (occupational groups) and into individual income ranges (social strata). If socio-occupational stratification proposes to summarize the heterogeneity of the patterns of a society, it must be able to represent relatively homogeneous groups of the population according to characteristics associated with this concept. It is the type of analysis that the literature calls construct validity [Quadros and Maia, 2010], which was analyzed in this work according to the composition of the identified socio-occupational groups in relation to characteristics of sex, color, age, level of education, occupation, income and geographical region (census tracts) of its members.

To fulfill the proposed objective, the work steps consisted of: (i) spatial microsimulation to obtain the spatial microdata in the territorial unit of census tracts, since important variables for analysis (occupation and educational level) are only in the microdata; (ii) data regionalization by the Skater method, using as analysis variables in the territorial unit of census tracts: age, gender, color, income, occupation and education level. Individuals 10 years of age and over who performed paid work in the week or unpaid work for at least one hour a week, including self-consumption and self-building activities, were considered to be employed.

Spatial microsimulation techniques are then used in this work to combine the advantages of existing data and achieve the intended objective, both qualitatively and spatially detailed data [Feitosa, Jacovine e Rosemback, 2016].

2. Materials and Methods

2.1. Study area

The study area of the present work consists of RMVPLN, which was created by the Complementary Law n. 66 of 2011, and its effective creation in 2012 through Complementary Law no. 1166, 2012, is already born large and surrounded by conflicting interests [EMPLASA, 2018], [Maria, 2016]. Currently, there is wide intra-regional economic diversity, and the region has a great diversified economic activity, not yet explored correctly. Although municipalities have different scenarios ranging from forest formations to differences in the absolute number of population, all municipalities have been encompassed in a single Metropolitan Region (Figure 1), which, according to the state government, aims to join efforts to give more conditions to this region to better serve the State of São Paulo and the country, as well as to enable municipalities in less developed economies to have the opportunity to integrate into the regional development process [EMPLASA, 2018]. This vision makes small municipalities invisible in territorial planning.

The focus of this study was on subregion 4, due to its historical importance in the coffee period and currently invisible to the current metropolitan planning applied throughout the RMVPLN. The subregion comprises eight municipalities: Cruzeiro, Lavrinhas, Queluz, Silveiras, Areias, Sao Jose do Barreiro, Arapei and Bananal, these are shown in Figure 2.

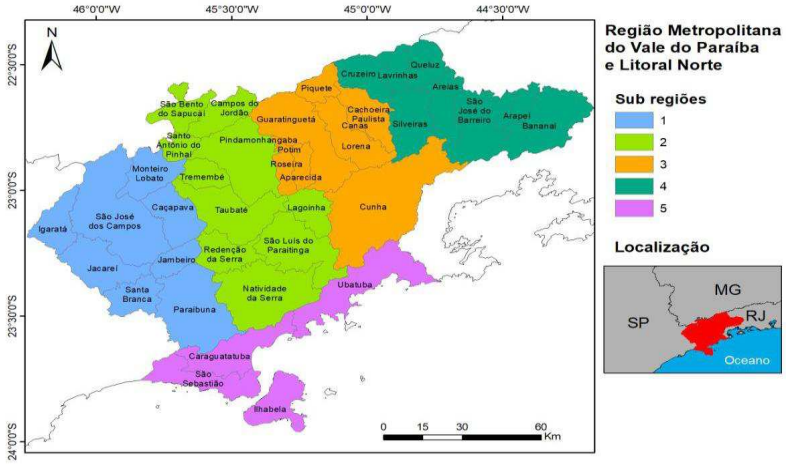

Figure 1. Location of the Paraíba Valley and North Coast Metropolitan Region. Source: Prepared by the author through the IBGE database (2010).

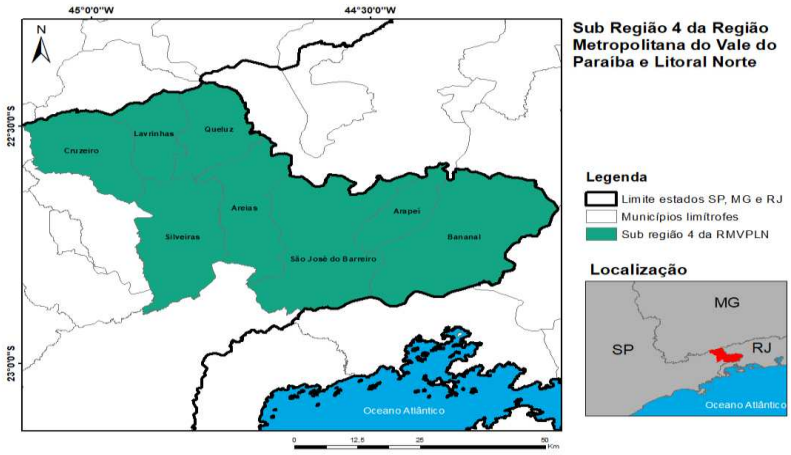

Figure 2. Location of RMVPLN Subregion 4, area of study of this work. Source: Prepared by the author through the IBGE database (2010).

2.2. Data base

The data used in this work come from the 2010 Demographic Census, conducted by IBGE. The Census is the most comprehensive statistical survey conducted in Brazil, collecting data on the composition and characteristics of the population, families, households and their surroundings and is available to all municipalities in the country [Feitosa, Jacovine e Rosemback, 2016], [IBGE, 2011], [IBGE,2010].

In performing the census, IBGE applies two types of questionnaires: the basic and the sample. The Basic Questionnaire (37 items) is applied to all households, except those selected for the sample, and contains the investigation of the characteristics of the household and residents. The Sample Questionnaire (108 items) is applied to all households selected for the sample, about 11% of the population. In addition to the research contained in the Basic Questionnaire, it covers other household characteristics and researches important social, economic and demographic information of its residents [Feitosa, Jacovine and Rosemback, 2016], [IBGE, 2011], [IBGE, 2010]. Table 1

summarizes the data used in this study.

Universe data from the Basic Questionnaire are available in tables and in a file aggregated by Census tracts. Census tracts according to IBGE (2011) are: "[...] the smallest territorial unit, formed by continuous area, entirely contained in urban or rural area, with adequate size for the operation of research and whose set exhausts the entire National Territory, which ensures full coverage of the country" [IBGE, 2011, p.4].

The microdata from the Sample Questionnaire are available in tables and in a territorial unit called Weighting Area. Weighting Area according to IBGE (2010) is: "[...] a geographical unit, formed by a grouping of census tracts, to apply the estimation calibration procedures with the known information for the population as a whole" [IBGE, 2010, p.14].

Table 1. Data used.

DATA	FORMAT	SPACE AGGREGATION UNIT	SOURCE	YEAR
Municipal Limits	Vector	MUN.	IBGE	2010
Census Tracts	Vector	MUN. e CT	IBGE	2010
Weighting Areas Aggregate	Vector	MUN. e WA	IBGE	2010
Demographic Census Data	Table	CT	IBGE	2010
Microdata Demographic Census	Table	WA	IBGE	2010

Legend: MUN.: Municipal; CT: Census Tracts; WA: Weighting Area. Source: Prepared by the author.

2.3. Spatial microsimulation and the IPF method

According to Lovelace and Dumont (2016), to understand the concept of spatial microsimulation it is important to look at the three parts that make up its nomenclature: spatial, micro and simulation. The first part, spatial, shows the intention to understand how what is being analyzed varies in space, and not (only) between individuals, thus distinguishing this approach from the field of microsimulation only. The second part, micro, shows the level of information and the degree of detail that can be worked with the technique. The third part, simulation, as in all modeling work, brings with it the idea of producing data estimates.

Thus, spatial microsimulation is understood in this paper as "the creation, analysis and modeling of data at the individual level allocated to geographical zones" [Lovelace and Dumont, 2016, p.7]. According to Jacovine (2017) it is important to emphasize that, strictly speaking, new individuals and information are not being created with spatial microsimulation. Both the "new" individuals and the information generated are, in fact, the result of the "reorganization" and "combination" of existing data in the bases worked.

Further explaining the process of spatial microsimulation, it is necessary to understand that initially there are two types of data: an aggregate in a certain spatial unit (in this case, aggregated by census tracts) and another disaggregated (called microdata that is in the spatial weighting area scale). In order to analyze socio-occupational groups,

the aim is to have a set of available information present in a smaller spatial aggregation unit than the municipality, such as the census tracts. By applying a spatial microsimulation technique to this data set, an estimate is generated, which is called spatial microdata, where there is a decrease in the spatial scale of the analyzed microdata set.

For this to occur it is important that the data meet a set of requirements, something that varies between the various existing techniques of spatial microsimulation. But all methods have in common the requirement that both aggregate and microdata data must have variables in common, called constraint variables. In addition, databases must be organized and systematized in specific ways. After these requirements are met, the result of microsimulation is the allocation of individuals (new microdata) to the census tracts, thus bringing information that was present only in the coarser resolution spatial units to finer spatial resolution units, thus opening up many opportunities for analyzes and interpretations of territorial reality [Jacovine, 2017].

Because it has many applications in different contexts, there are numerous spatial microsimulation techniques available in the literature. In the present work, we chose to use the IPF ("*Iterative Proportional Fitting*") reweighting technique for the following reasons: the existence of publicly accessible 2010 IBGE Census microdata; Comparative studies published in the literature show that reweighting methods are the most efficient [Hermes and Poulsen, 2012]; and because it is more commonly used, being a simple technique, easy to understand and replicable [Feitosa, Jacovine e Rosemback, 2016], [Hermes e Poulsen, 2012].

The IPF method, like any other method of spatial microsimulation, consists of the estimation and allocation of microdata on spatial scales or geographic clippings of interest (census tracts, neighborhoods, etc.). For this, the method confronts different databases (microdata and aggregated data), but with common variables, seeking to compute the representativeness of individuals in each area of interest. The more representative an individual's characteristics are for a given area, the greater the weight given to him. In the opposite case, the rarer the characteristics of an individual, the lower their weight [Jacovine, 2017], [Lovelace and Dumont, 2016].

The IPF requires two types of aggregate data: data with spatial information that presents the number of individuals (total count) for each of its composing variables; and data at the individual level (microdata), which presents a greater richness of variables, besides allowing one or more characteristics to be associated to the same individual [Jacovine, 2017], [Lovelace and Dumont, 2016].

Regarding the variables used, they are subdivided into two groups by the IPF, based on the function they fulfill: restriction variables and variables of interest. Responsible for allowing the method to function properly, the presence of restriction variables on both bases is vital. This is because they enable the connection between these two universes, allowing estimates for the variables of interest to be generated. The variables of interest are those you want to know better, but do not present data or information at a given scalar level [Jacovine, 2017], [Lovelace and Dumont, 2016].

Regarding the variables selected to obtain socio-occupational groups from Subregion 4 of the RMVPLN, in the case of restriction variables, characteristics related to the household head were used, such as "race / color", "gender", "age" and "yield on all jobs". These variables have important characteristics that interfere with what is expected to be estimated, justifying their choice. Regarding the variables of interest, the chosen

ones were "occupation" and "educational level" both present only in microdata. This is because, with these two variables, one can obtain the main factors for the creation of socio-occupational groups. Once the restriction and interest variables are defined, the next step of the method is to define the initial weight to be assigned to each of the individuals involved in the process. Generally, the initial value assigned is the same, assuming that all should be treated the same at the beginning of the process [Jacovine, 2017], [Lovelace and Dumont, 2016].

Once the initial weight is set, the IPF can then be executed. For this, from Equation 1, the algorithm starts from the established initial weight and adjusts it for all households in the first census tract, for example. At the end of the first census sector, the algorithm will move to the second sector, using the weights obtained in the previous step. And so the process will go on, individual by individual, sector by sector. After all sectors are calculated for the first constraint variable, the algorithm will move to the next constraint variable and the same path will be taken. It is noteworthy that, in order to obtain a better fit, the algorithm, after computing the weights for all variables, returns to the first and restarts the calculations, using the final weight of the last restriction variable. This will end when the process is terminated using all constraint variables. What is verified, therefore, is that the procedure is made restriction variable by restriction variable, so that, at the end of the process, all individuals and their characteristics will have their weights computed for each of the census tracts analyzed [Jacovine, 2017], [Lovelace and Dumont, 2016].

$$Pn_i = \frac{P_i * Agreg_{var}}{Micro_{var}} \qquad \textbf{Equation 1}$$

Where,

Pn_i: New weight;

P_i: Initial or previous iteration weight;

$Agreg_{var}$: aggregated data for the census tract under analysis;

$Micro_{var}$: microdata for the same variable as the aggregate data.

With the weights generated and expressed in integers, the next step performed is data expansion. This step consists of creating tables with individual records associated with certain portions of the territory. Thus, there is the spatial microdata [Jacovine, 2017], [Lovelace and Dumont, 2016].

2.4. Skater regionalization

Regionalization can be seen as a classification procedure applied to geo-objects with polygonal representation. It requires contiguity between geo-objects of the same class, where geo-objects members of the same class must form a single, homogeneous and spatially contiguous region. One tool that performs Regionalization is the Skater tool. It considers the spatial location of geo-objects (centroids) and is based on the neighborhood structure between geo-objects (graph: {nodes, edges}) [Assunção et al., 2006]. The neighborhood matrix considered in this study was the simplest one, which considers neighbors by contiguity criterion.

The Skater method performs regionalization via the Minimum Spanning Tree (MST) method, where the construction of the MST is based on measures of similarity

between geo objects, analyzing the "costs" of graph edges (between geo objects). Initially costs are calculated using a metric that assesses the similarity between two geo-objects. This metric is measured by the similarity coefficient, denoted by S, and these similarity coefficients across all geo-objects can be condensed into a S_{nxn} matrix [Assunção et al., 2006], as shown in Figure 3.

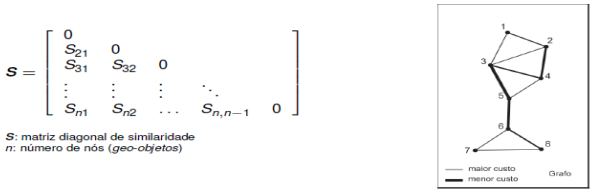

Figure 3. How similarity and its costs are measured. Source: Camargo and Monteiro (2010).

Similarly, the *p* attributes or variables associated with each of the *n* geo objects can also be represented by an X_{nxp} matrix (Figure 4).

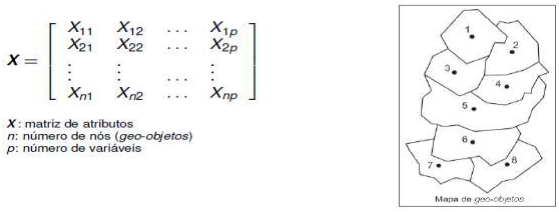

Figure 4. Representation of attributes in matrix form. Source: Camargo and Monteiro (2010).

The similarity coefficient is measured by the Minkowski metric, represented by Equation 2.

$$S_{ij}^{(\lambda)} = [\sum_{I=1}^{p} |X_{iI} - X_{jI}|^{\lambda}]^{1/\lambda} \qquad \lambda > 0$$

Equation 2

Where:

i e j: geo-object indexers;

I: variable indexer (attribute);

X_{iI} *e* X_{jI}: value of the ith variable associated to the i-th and j-th geo-object, respectively;

λ: is a parameter; higher values of λ => emphasize the variable with the greatest difference between X_{iI} and X_{jI}.

For λ=2, the similarity coefficient between two geo-objects is obtained through the calculated Euclidean distance over the attribute space. And it was with this case that the current work was performed.

Finally, there is the last step: the pruning of the MST. In this step of the procedure the way of assigning costs to edges is modified in order to obtain better results: more homogeneous regions, more balanced in terms of geo-object numbers per region, and finally, the lower cost edges are removed.

3. Results and Discussion

Figure 5 is the result after regionalization, forming 15 homogeneous socio-occupational groups considering the analysis variables. Spatial microsimulation expanded and allocated the original microdata into sectors and allowed for a much more detailed spatial distribution of occupations, the main variable for analysis of socio-occupational groups.

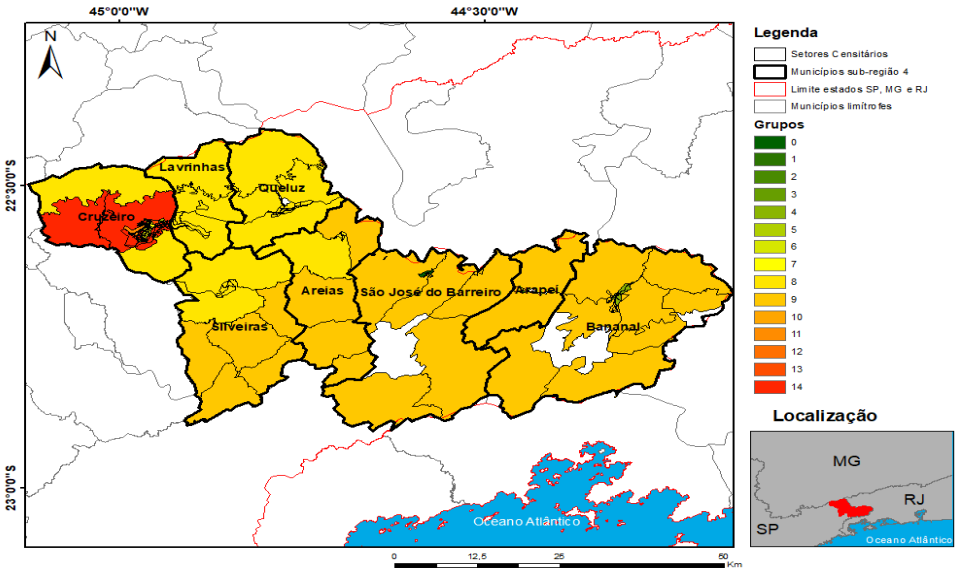

Figure 5. Regionalization result, totaling 15 socio-occupational groups. Source: Prepared by the author through the IBGE database (2010).

Figure 6 shows the main groups formed. Group 1 is a group that clashes with the whole region and is located in the municipality of Cruzeiro. This group contains 201 people, of which 45% are women and 55% men. 100% of the individuals declare themselves white, with an average age of 41 to 50 years. The average income is 5 to 10 minimum wages (which in 2010 was R $ 510.00), explained by the high number of individuals with complete high school and incomplete higher education (32%) and with complete higher education (58%). The occupations with the highest percentages are science and intellectual professionals (31%), directors and managers (13%), and technicians and middle level professionals (16%).

Groups 7, 8 and 9 were the groups that included more census tracts from different municipalities and portray the reality of the subregion as a whole. These groups contain on average 27% women and 73% men. 68% of the individuals declare themselves white, 26% brown and 5% black. The age groups with the largest numbers of people are 31 to 40 years (20%), 41 to 50 years (26%) and 51 to 60 years (29%). The average income in these groups is ½ to 2 minimum wages, explained by the high number of individuals with complete high school and incomplete higher education (22%) and without education and incomplete elementary school (48%). The occupations with the highest percentages are elementary occupations (28%), trade and market sales services workers (14%) and skilled workers and craftsmen of mechanical and another crafts construction (14%).

Finally, group 14 has the largest number of census tracts (65 sectors). This socio-occupational group contains 13,028 people, of which 17% are women and 83% men. 64% of people declare themselves white, 28% brown and 7% black. The average age is 51 to

60, with a low average income of ½ to 1 minimum wage. 58% of the individuals have no education and incomplete elementary school, where the occupations with the highest percentages are elementary occupations (26%), trade and market sales services workers (18%) and skilled workers and craftsmen of mechanical and mechanical arts construction; other holes (26%).

The other groups not mentioned have similar characteristics to groups 7, 8 and 9 and some were not grouped to them by contiguity criteria, or because they have one or more analysis variables with very different values. For example, group 8 and group 9 were not grouped by the large difference in numbers of people they contained in total. The creation of similar groups can be explained by the large difference in numbers of census tracts between municipalities. Most municipalities in subregion 4 have few census tracts because they are small towns, only Cruzeiro disagrees with this reality of the subregion. For this reason, most of the groups created, which were not mentioned, are within the municipality of Cruzeiro, as this is a municipality that has larger numbers of census tracts with greater differences between them (see Figure 6).

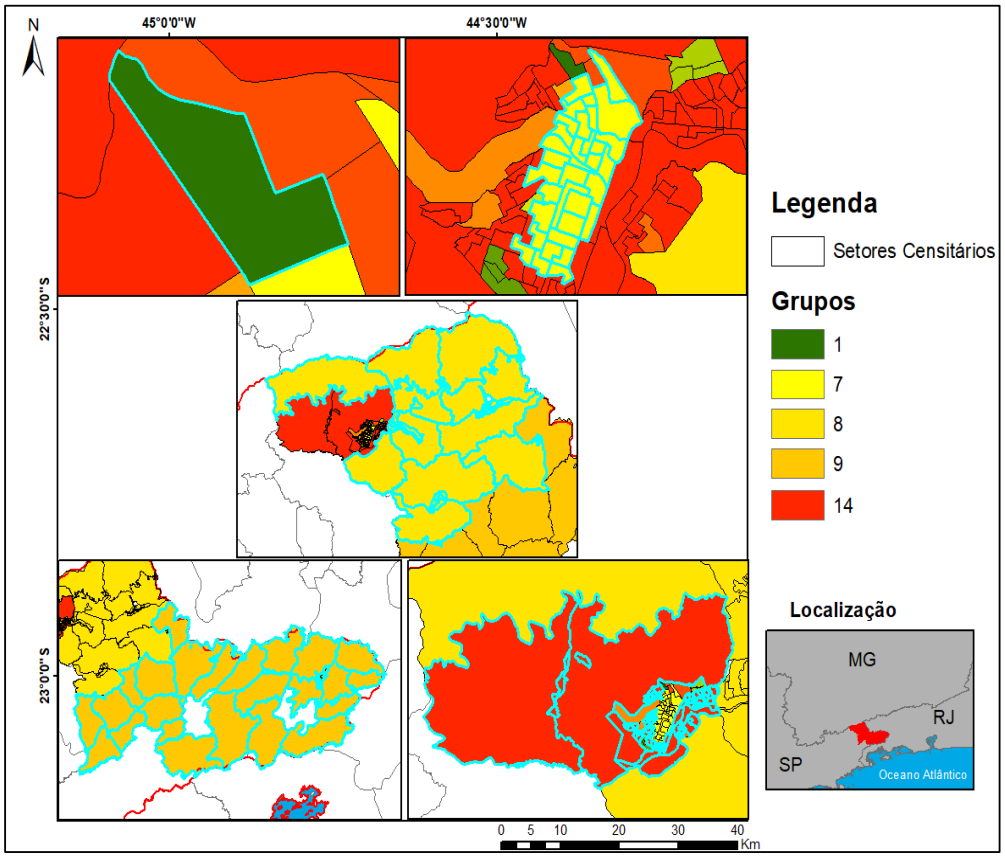

Figure 6. Socio-occupational groups 1, 7, 8, 9 and 14. Source: Prepared by the author through the IBGE database (2010).

4. Conclusions

This work showed how spatial microsimulation techniques introduce new possibilities for studies of socio-occupational groups in more detailed spatial units. While data aggregated by census tracts show fine spatial resolution, there is no possibility of having variables

such as “occupation” at this level due to the confidentiality of data presented by the IBGE Demographic Census. Census sample data (microdata) have a more detailed data set that is suitable for analyzing and proposing a socio-occupational structure, but lacks detailed spatial information, and the IPF method was able to merge the two qualifications of the two data.

The Skater Regionalization method allowed to analyze and to join in homogeneous groups the studied variables, that is, it was possible to propose a socio-occupational structure for the RMVPLN subregion 4. The grouping allowed highlighting the high degree of inequality within the subregion and consistently discriminating important socioeconomic groups of the population.

Additional testing should be performed to ensure that the resulting spatial microdata is as representative as possible within the limitations of the data. This requires exploring the choice of different constraint variables and validating the resulting estimates. It is also important to test and compare different methods of spatial microsimulation, exploring their main characteristics, variability and validity against the resulting external data sets, in order to arrive at a better estimate. In addition, modifying the neighborhood criteria for the neighborhood matrix considered in this paper to apply the Skater method may also offer improvements to the proposed socio-occupational structure.

References

Assunção, Renato, Neves, M. C., Câmara, G., Freitas, C. (2006) “Efficient regionalisation techniques for socio-economic geographical units using minimum spanning trees”. International Journal of Geographical Information Science (Print), Inglaterra, v. 20, n.7, p. 797-812.

Camargo, E. C. G., Monteiro, A. M. V. (2010) “Regionalização via Skater”. SER-301 Análise Espacial de Dados Geográficos, Instituto Nacional de Pesquisas Espaciais Divisão de Processamento de Imagens.

EMPLASA. (2018) “Região Metropolitana do Vale do Paraíba e Litoral Norte. São Paulo: 2018”. Disponível em:

<https://bibliotecavirtual.emplasa.sp.gov.br/ExibirDetalhes.aspx?funcao=kcDocumentos&id=2715&lingua=PT>.

Feitosa, F., Jacovine, T. C., Rosemback, R. G. (2016) “Small Area Housing Deficit Estimation : A Spatial Microsimulation Approach”. Brazilian Journal of Cartography (2016), Nº 68/6, Special Issue GEOINFO 2015: 1157-1169 Brazilian Society of Cartography, Geodesy, Photgrammetry and Remote Sense ISSN: 1808-0936.

Hermes, K., Poulsen, M. (2012) “A review of current methods to generate synthetic spatial microdata using reweighting and future directions”. Computers, Environment and Urban Systems, v. 36, n. 4, p. 281–290.

IBGE. Censo Demográfico: Notas Metodológicas. 2010.

___. Base de informações do Censo Demográfico 2010: Resultados do Universo por setor censitário. 2011.

IPEA. (2008) “Pobreza e mudança social”. v. 1, Brasília, Instituto de Pesquisas Econômicas Aplicadas, Comunicado da Presidência, n. 9. Disponível em: <

http://www.ipea.gov.br/sites/000/2/pdf/Pnad_2007_AnalisesPobreza.pdf>

Jacovine, T. C. (2017) “Estimativas de Deficit Habitacional para Pequenas Áreas: Uma Proposta de Abordagem Baseada em Microssimulação Espacial”. São Bernardo do Campo.

Jannuzzi, P. M. (2003) “Estratificação socioocupacional para estudos de mercado e pesquisa social no Brasil”. São Paulo em Perspectiva, v. 17, n. 3-4, p. 247-254.

Lovelace, R., Dumont, M. (2016) “Spatial Microsimulation with R”. Chapman & Hall/CRC The R Series.

Maria, J. M. (2016) “Região e regionalização: estudo da região metropolitana do Vale do Paraíba e Litoral Norte”. Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, Rio Claro – SP.

Neri, M., Caravalhaes, L. (Coords.). (2008) “Miséria e a nova classe média na década da igualdade”. Rio de Janeiro: FGV/IBRE, CPS.

Quadros, W. J., Maia, A. G. (2010) “Estrutura sócio-ocupacional no Brasil”. R. Econ. contemp., Rio de Janeiro, v. 14, n. 3, p. 443-468.

Rose, D.; Harrison, E. (2007) “The European socio-economic classifi cation: a new social class schema for comparative European research”. European Societies, v. 9, n. 3, p. 459-490.

Rose, D.; Pevalin, D. J. (2005) “The NS-SEC: origins, development and use”. Basingstoke: Palgrave Macmillan.