Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Spatial Microsimulation Combined with Skater Regionalization Methods: A Study for the Paraíba Valley and North Coast Metropolitan Region Subregion 4 in the São Paulo State

Microssimulação Espacial Combinada com Método de Regionalização Skater: Um Estudo para a Sub-região 4 da Região Metropolitana do Vale do Paraíba e Litoral Norte no Estado de São Paulo

Gabriela Carvalho de Oliveira [1], Tathiane Mayumi Anazawa [2] and Antonio Miguel Vieira Monteiro [3]

1 Instituto Nacional de Pesquisas Espaciais (INPE), Observação da Terra (OBT), São José dos Campos, Brasil.
gabriela.oliveira@inpe.br.
ORCID: https://orcid.org/0000-0003-4784-6620
2 Instituto Nacional de Pesquisas Espaciais (INPE), Observação da Terra (OBT), São José dos Campos, Brasil.
tathiane.anazawa@inpe.br.
ORCID: https://orcid.org/0000-0003-2675-0566
3 Instituto Nacional de Pesquisas Espaciais (INPE), Observação da Terra (OBT), São José dos Campos, Brasil.
miguel.monteiro@inpe.br.
ORCID: https://orcid.org/0000-0003-1477-1749

Recebido: 03.2020 | Aceito: 08.2020

Abstract: This study analyzes the socio-occupational distribution in the Paraíba Valley and North Coast Metropolitan Region (in Portuguese: Região Metropolitana do Vale do Paraíba e Litoral Norte – RMVPLN) Subregion 4 using spatial microsimulation techniques. To fulfill the proposed objective, the Iterative Proportional Fitting (IPF) technique was used to obtain spatial microdata in the territorial census tracts unit through the 2010 Brazilian Demographic Census. After the Skater regionalization technique was applied, eight homogeneous socio-occupational groups were found. 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 the São Paulo state. Although this is a preliminary study, it is already capable to identify inequalities degrees that consistently spatially segregate and the less privileged population socioeconomic groups.
Keywords: Metropolitan Analysis. Spatial Microsimulation. IPF. Skater.

Resumo: Este estudo analisa a distribuição sócio-ocupacional na Sub-região 4 da Região Metropolitana do Vale do Paraíba e Litoral Norte (RMVPLN), utilizando técnicas de microssimulação espacial. Para cumprir o objetivo proposto, foi empregada a técnica de Ajuste Proporcional Iterativo (IPF) para obtenção de microdados espaciais na unidade territorial de setores censitários utilizando dados do Censo Demográfico Brasileiro de 2010. Após a aplicação da técnica de regionalização Skater, foram encontrados oito grupos sócio-ocupacionais homogêneos. De maneira geral, as categorias sócio-ocupacionais propostas, estudadas em escala intraurbana, permitiram evidenciar a estrutura social de uma sub-região do mais novo espaço metropolitano do estado de São Paulo. Embora este seja um estudo preliminar, já é capaz de identificar graus de desigualdades que segregam espacialmente de forma consistente e os grupos socioeconômicos menos privilegiados da população.
Palavras-chave: Análises Metropolitana. Microssimulação espacial. IPF. Skater.

1 INTRODUCTION

According to Quadros and Maia (2010), discussions on social policies in Brazil are largely dominated by studies that focus on identifying poverty and classifying the population according to income ranges. However, income cannot be the only factor that delimits the individual position in the social hierarchy, even though it plays an important role in individual integration in the market for goods and services. In order to be

an option for stratifying the population according to their income, the literature proposes typologies with comprehensive concepts, which approximate the society class behavior (QUADROS; MAIA, 2010).

Job occupations have come to play an essential role in shaping the modern capitalist society's structure as they are effective in providing information in greater detail about the individual's income, education, and lifestyle levels, etc. Therefore, identifying the society's socio-occupational structure enriches social analyzes related to migration, mobility, consumption, exclusion, inequality, health, among others (QUADROS; MAIA, 2010; RIBEIRO; LAGO, 2000).

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 to understand the society's relation complexity 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 socio-occupational structure distribution and composition. The analyzes are supported by a proposal to stratify Brazilian society based on the job occupations of the labor market structure used by the Brazilian Institute of Geography and Statistics - IBGE (2010).

These analyzes start from the premise that relatively homogeneous social groups can be obtained from the individual's insertion into the labor market (occupational groups) and individual income ranges (social strata). If socio-occupational stratification proposes to summarize the society patterns heterogeneity it must be able to represent relatively homogeneous population groups according to characteristics associated with this concept. It is the type of analysis that the literature calls construct validity (QUADROS; MAIA, 2010), which was analyzed in this study according to the identified socio-occupational groups composition concerning gender, color, age, education level, job occupation, income and geographical region characteristics of its members.

To fulfill the proposed objective, the next steps were followed:

a) spatial microsimulation to obtain the spatial microdata in the census tracts territorial unit, since important variables for analysis (job occupation and educational level) are only in the microdata;

b) data regionalization by the Skater method, using the variables in the census tracts territorial unit: age, gender, color, income, job occupation, and education level.

Individuals with 10 years old 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 study to combine the advantages of existing data and achieve the intended objective, both qualitatively and spatially detailed (FEITOSA; JACOVINE; ROSEMBACK, 2016).

2 MATERIALS AND METHODS

2.1 Study area

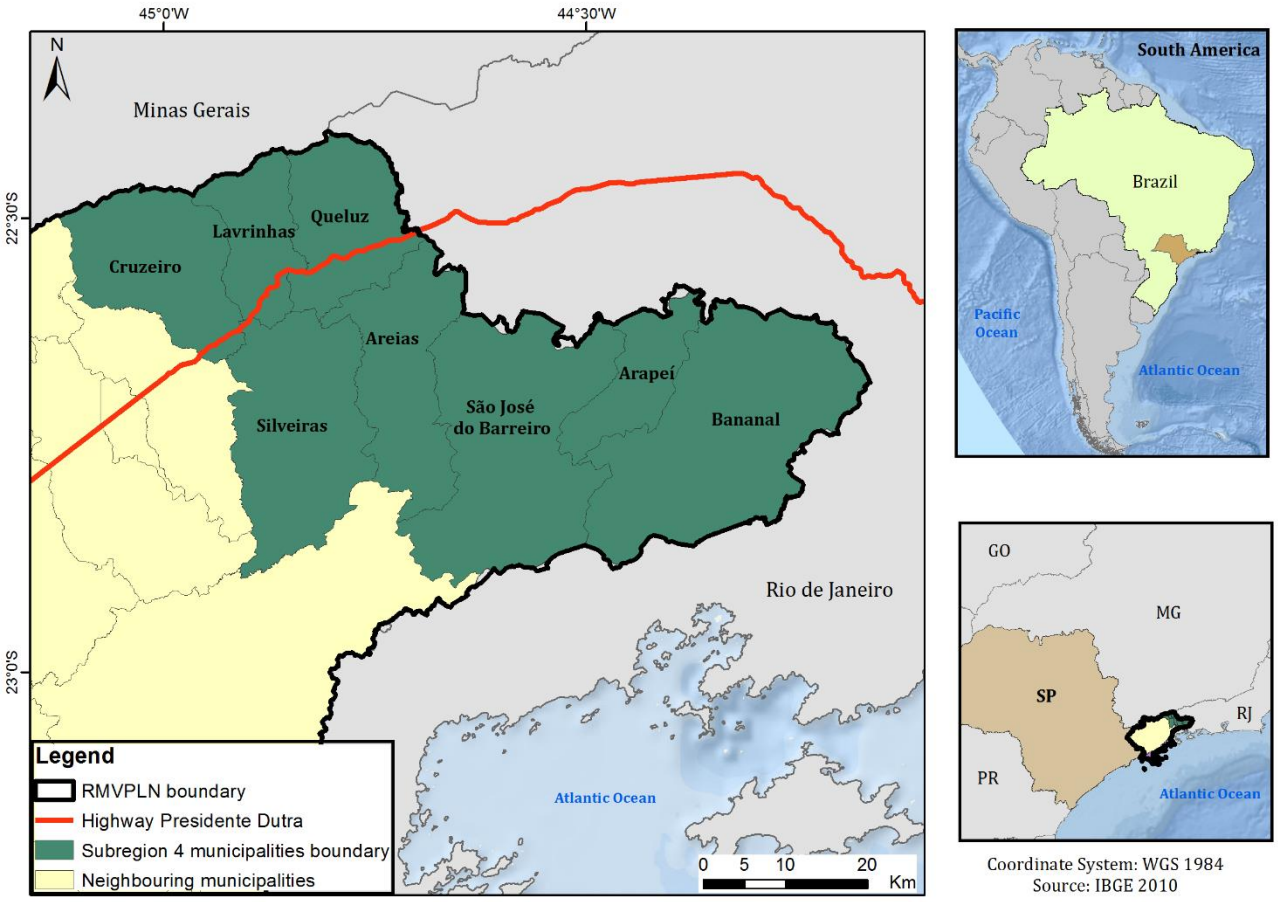
The study area is the RMVPLN, which was created by the Complementary Law n. 66 of 2011, and its effective creation in 2012 through Complementary Law no. 1166, 2012 (EMPLASA, 2012; MARIA, 2016). The RMVPLN has already born large and surrounded by conflicting interests.

The Vale do Paraíba region has historically been of significant importance, showing broad economic growth in the coffee-growing period in Brazil. The region was marked in the 17th century by the so-called "Gold Cycle", which promoted the interior of the country job occupation due to the change in population flow. The southeast region promoted food production for the new settlements with the highest exploitation rate (NASCIMENTO; RICCI; RODRIGUES, 2014). In this context, the Paraíba River valley was colonized by tropeiros and pioneers who founded the first villages in the region, called Jacareí, Guaratinguetá, and Taubaté among others. Since then, the region has become an important circulation axis, with roads linking the Rio de Janeiro to Minas Gerais state. Several indigenous trails led to the towns and cities connection and the formation of others that currently make up the Historic Valley, such as Cunha, São Luis do Paraitinga, and Paraibuna (MÜLLER, 1969).

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, 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, 2019). 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 national coffee period and currently invisible to the current RMVPLN planning. The subregion comprises eight municipalities which are: Cruzeiro, Lavrinhas, Queluz, Silveiras, Areias, Sao Jose do Barreiro, Arapeí, and Bananal showed in Figure 1.

Figure 1 – RMVPLN Subregion 4 location.



Source: The authors (2020).

2.2 Database

The data used in this study come from the 2010 Brazilian Demographic Census, conducted by IBGE. The Census is the most comprehensive statistical survey conducted in Brazil, collecting data about the composition and characteristics of the population, families, households, and their surroundings and is available to all municipalities in the country (IBGE, 2010a; 2011).

The IBGE applies two questionnaires types called the basic and the sample. The Basic Questionnaire (BQ), which has 37 items, and is applied to all households, except those selected for the sample, and contains the general household and resident's characteristics questions. The Sample Questionnaire (SQ), which is more extensive has 108 items and is applied to only 11% of households selected from the sample. Also, the BQ covers other household characteristics and has important social, economic, and demographic questions (IBGE, 2010a, 2011). Table 1 summarizes the data used in this study.

Universe data from the BQ are available in tables and a file aggregated by Census tracts. Census tracts according to IBGE are: "[...] the smallest territorial unit, formed by continuous area, entirely contained in an urban or rural area, with adequate size for the research operation and whose set exhausts the entire National Territory, which ensures full country coverage" (IBGE, 2011, p.4). The variables used in the demographic census data were: gender, color/race, age, and job occupation's yield.

The microdata from the SQ is available in tables and in a territorial unit called Weighting Area, which is according to IBGE: "[...] 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). The variables used in the demographic census microdata sample were: gender, color/race, age, job occupation's yield, and educational level.

Table 1 – Data used. Legend: MUN.: Municipal; CT: Census Tracts; WA: Weighing Area.

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
Demographic Census Microdata	Table	WA	IBGE	2010

Source: The authors (2020).

2.3 Spatial microsimulation and the IPF method

According to Lovelace and Dumont (2016), to understand the spatial microsimulation concept 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 microsimulation field. The second part, micro, shows the information and degree detail level that can be achieved. The third part, simulation, as in all modeling analysis, brings the idea of producing data estimations.

Thus, spatial microsimulation is understood in this study as "the creation, analysis, and modeling of data at the individual level allocated to geographical zones" (LOVELACE; DUMONT, 2016, p.7). It is important to highlight that, strictly speaking, new individuals and information are not being created with spatial microsimulation. During spatial microsimulation, what happens is the repetition, given their individual representativeness (in this study case for each census tracts) in the microdata (spatial level of weighting areas), although in a different order and in different combinations. Thus, spatial microsimulation does not increase the dataset diversity it simply alters its spatial aggregation (LOVELACE; DUMONT, 2016).

Essentially, spatial microsimulation calculates the representativeness of each individual in the sample for each census tract. The more similar the general census tracts attributes are to the analyzed individual characteristics, the greater is its representativeness. This results in a synthetic population, that is, it estimates in which census tracts each individual who answered the sample questionnaire may be (MIRANTI et al., 2016).

Further to explaining the spatial microsimulation process, it is necessary to understand that initially there are two data types. The first is aggregate in a certain spatial unit, in this case, aggregated by census tracts, and the other is disaggregated, called microdata that is in the spatial weighting area scale. In order to analyze socio-occupational groups, the aim is to have available information present to set 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 the data must meet a requirement set, something that varies between the several existing spatial microsimulation techniques. However, all methods have in common the requirement that both aggregate and microdata data must have the same variables, called constraint variables. In addition, databases must be organized and systematized in specific ways. After these requirements are met, the microsimulation result is the individual allocations, known as new microdata, to the census tracts, thus bringing information that was present only in the coarser-resolution spatial to finer spatial resolution, thus opening up many

opportunities for territorial analyze and interpretation (JACOVINE, 2017).

Because it has many applications in different contexts, there are numerous spatial microsimulation techniques available in the literature. In this study the Iterative Proportional Fitting (IPF) reweighting technique was chosen because of the existence of publicly accessible 2010 IBGE Census microdata. In addition, several studies show that the reweighting technique is the most efficient method (HERMES; POULSEN, 2012); and because it is more commonly used, is a simple technique, easy to understand and replicable (HERMES; POULSEN, 2012; LOVELACE et al., 2015). When used in spatial microsimulation, IPF can help overcome the limitations of extensive and geographically aggregated data sources (LOVELACE et al., 2015).

The IPF method is known and used historically in statistics, to adjust known margins of restriction variables. Like any other spatial microsimulation method, it consists of estimating and allocating microdata in spatial scales or geographic clippings of interest, such as census tracts, neighborhoods, etc. For this, the method confronts different databases, such as microdata and aggregated data, but with variables, seeking to calculate the individual representativeness in each interested area. The more representative an individual's characteristics are for a given area, the greater the weight attributed to it. On the other hand, the rarer the individual characteristics, the lower their weight (JACOVINE, 2017; LOVELACE; DUMONT, 2016; WHITWORTH et al., 2013).

The IPF requires two aggregate data types, which are the spatial information data that presents the total count number for each of its composing variables; and data at the individual microdata level, which presents a greater richness of variables, besides allowing one or more characteristics to be associated to the same individual. Regarding the variables used, it is subdivided into two groups by the IPF, based on the function they fulfill: restriction variables and target variables. Responsible for allowing the method to function properly, the restriction variables presence on both bases is vital. This is because they enable the connection between these two universes, allowing estimates for the target variables to be generated. The target variables are those you want to know better but do not present data or information at a given scalar level. The suitability of the variables and the constraint's number to be used in the process depends on the ultimate estimation purpose (LOVELACE; DUMONT, 2016; WHITWORTH et al., 2013).

Regarding the variables selected to obtain socio-occupational groups from RMVPLN Subregion 4 the restriction variables case, characteristics as race/color, gender, age, and the job occupation's yield were used. These variables have important characteristics that interfere with what is expected to be estimated, therefore justifying their choice. Regarding the target variables, the chosen ones were job occupation and educational level, both only present in microdata. This happens because, with these two variables, it is possible to obtain the main factors for the socio-occupational group's creation. 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 individual 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 E DUMONT, 2016).

Once the initial weight is set, the IPF can then be executed. For this, from Eq. (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 tracts, 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, so the algorithm will move to the next constraint variable and the same path will be taken. It is noteworthy that, 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 used all the constraint variables. What is verified, therefore, is that the procedure is made by each one the restriction variable, so that, at the end of the process, all individuals and their characteristics will have their weights computed for each census tract (JACOVINE, 2017; LOVELACE E DUMONT, 2016).

$$Pn_i = \frac{P_i * Agreg_{var}}{Micro_{var}} \quad (1)$$

where,

Pn_i: New weight;

P_i: Initial or previous iteration weight;

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

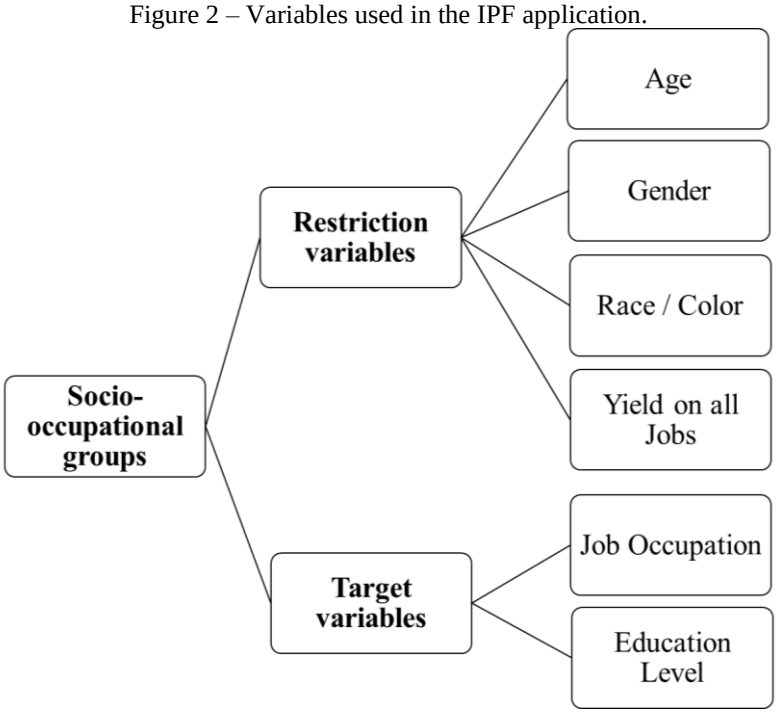
$Micro_{var}$: microdata for the same variable as the aggregate data.

With the weights generated and expressed in integers, the next step performed is the data expansion. This consists of creating tables with individual records associated with certain territory portions. Thus, there is spatial microdata (JACOVINE, 2017; LOVELACE; DUMONT, 2016).

2.4 Variables description

As explained in item 2.3, the IPF uses two variables sets to execute the method: restriction variables and variables of interest. The variables chosen for this study are summarized in Figure 2.

Figure 2 – Variables used in the IPF application.



Source: The authors (2020).

As already mentioned, all variables come from the 2010 Census data, and IBGE provides, together with the data, the description of each one of them and what was considered. This description is important for future data analysis.

The next variables definitions presented in Chart 1 come from the “Brazilian Demographic Census Dictionary of Variables Description” (2010):

Chart 1 – Variables description from the Brazilian Demographic Census Dictionary.

Variable	Description
Age	Age of the person in full years on the search reference date.
Gender	Gender of the person enrolled. Classified in: 1 – Male; 2 – Female.
Color or race	Color or race as declared by the registered person. Classified in: 1 - White: for the person who declared himself white. 2 - Black: for the person who declared himself black. 3 - Yellow: for the person who declared himself to be yellow (of oriental origin: Japanese, Chinese, Korean; etc.). 4 - Parda: for the person who declared himself brown. 5 - Indigenous: for the person who declared to be indigenous or Indian. This classification applies to both indigenous people who lived on indigenous lands and those who lived outside them. 9 – Ignored.
Yield on all jobs	Gross income from all jobs in minimum Brazilian wages. Without information: for whom, in the 25 to 31 July 2010 week: - was younger than 10 years old; or - did not work earning money, products, goods or benefits; and - did not have any paid work from which he was temporarily removed; and - did not help without any payment in paid work as a household resident; and - worked or not in planting, raising animals, or fishing, just to feed the residents.
Job occupation	Occupation in the job you had. This question investigated the occupation that the person had in the only job or in the main job that he had in the reference week. Without information: for whom, in the 25 to 31 July 2010 week: - was younger than 10 years old; or - did not work earning money, products, goods or benefits; and - did not have any paid work from which he was temporarily removed; and - did not help without any payment in paid work as a household resident; and - did not work on planting, raising animals, or fishing, just to feed the household residents. Some job occupation categories examples are: - Directors and managers; - Sciences and intellectuals professionals; - Mid-level technicians professionals; - Administrative support workers; - Service workers, merchants, and markets vendors; - Skilled workers in agriculture, forestry, hunting, and fishing; - Skilled workers, construction workers and craftsmen, mechanical arts and other crafts; - Plant and machine operators and assemblers; - Members of the armed forces, police and military firemen; - Elementary occupations; - No occupation (or that did not fit the requirements described above).
Education level	Education level of the person enrolled. Classified in: 1 - Without instruction and incomplete elementary school; 2 - Complete elementary school and incomplete high school; 3 - Complete high school and incomplete higher education; 4 - Complete higher education.

Source: IBGE (2010).

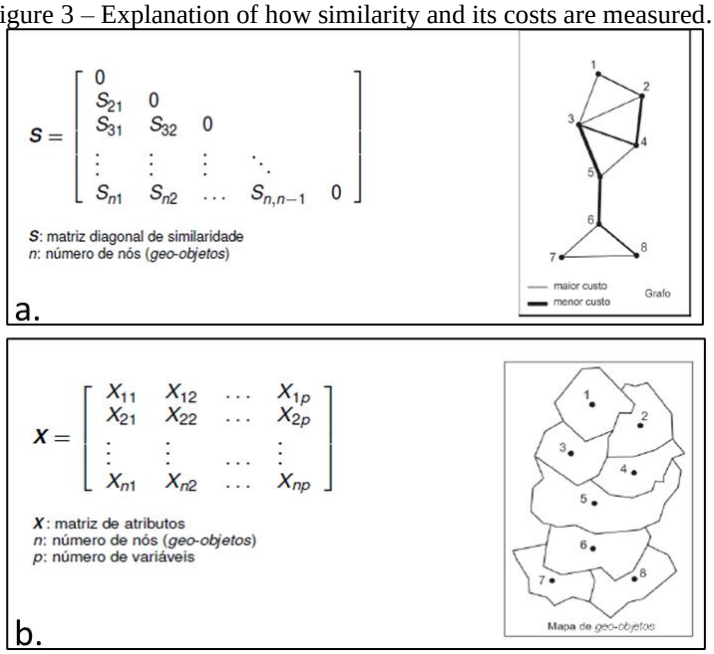
2.5 Skater regionalization

Regionalization can be seen as a classification procedure applied to geo-objects with polygonal representation. It requires contiguity between the same class geo-objects, 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 geo-objects (centroids) location 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 MST construction is based on similarity measures between geo-objects, analyzing the graph edges costs. Initially, costs are calculated using a metric that assesses the similarity only 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 an Snxn matrix (ASSUNÇÃO et al., 2006), presented in Figure 3.a. Similarly, the *p* attributes or variables associated with each of the *n* geo objects can also be represented by an X_{nxp} matrix (Figure 3.b.).

Figure 3 – Explanation of how similarity and its costs are measured.



Source: Camargo and Monteiro (2010).

The similarity coefficient is measured by the Minkowski metric, represented by Eq. (2).

$$S_{ij}^{(\lambda)} = \left[\sum_{l=1}^{p} |X_{il} - X_{jl}|^{\lambda}\right]^{1/\lambda} \qquad \lambda > 0 \tag{2}$$

where:

i e j: geo-object indexers;

I: variable indexer (attribute);

X_{iI} *e* X_{jI}: the value of the ith variable associated with 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 study was performed.

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

3 RESULTS AND DISCUSSION

Table 2 shows the socioeconomic synthesis of the municipalities in RMVPLN Subregion 4. It can be noticed that the Cruzeiro municipality has the largest population and Gross domestic product (GDP) compared

to the rest of the municipalities. The information about the average income, average education, and the busiest occupation were all the same for the municipalities. Elementary occupations are professionals who work as kitchen helpers, cleaning staff, street vendors, garbage collectors, among others (IBGE, 2010b). This occupation type is the one that contains the largest number of people and it is probably due to the low average wages and schooling found in the Subregion 4 municipalities history.

This Subregion, also known as the Historic Valley, had its heyday in the coffee period, however, afterward with the decline in wealth generated by the coffee cycle, the region was called by Monteiro Lobato “dead cities” (PHILIPPINI, 2019). This expression was given to the region since there was a period of economic difficulty after the coffee era (PHILIPPINI, 2019). Currently, the region's economy is geared towards small businesses, historic and ecological tourism (NASCIMENTO, 2015), which explains that elementary occupations are the occupation of most people who move to work in the region. The Cruzeiro municipality stands out in the region because s it has a national reference industry, called “Iochpe Maxion”, which is also the largest contractor in the region, but even with this scenario, it is still far from being comparable to other RMVPLN locations (NASCIMENTO; RICCI; RODRIGUES, 2014).

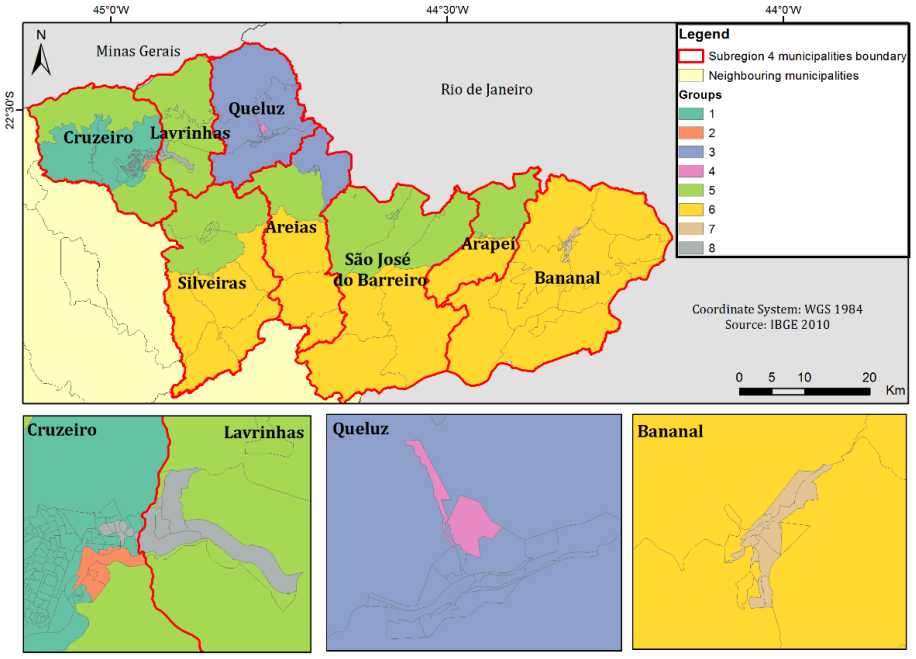
Table 2 – Socioeconomic synthesis of the municipalities in RMVPLN Subregion 4.

Municipalities	Population	GDP	Average Income (minimum wages)	Average Schooling	Job Occupation
Arapeí	2,493	22,642.91	more 1/2 to 2	Without instruction and incomplete elementary school	Elementary occupations
Areias	3,696	23,168.75			
Bananal	10,223	119,562.11			
Cruzeiro	77,039	1,450,013.25			
Lavrinhas	6,590	76,796.32			
Queluz	11,309	109,056.55			
São José do Barreiro	4,077	28,250.40			
Silveiras	5,792	37,239.46			

Source: IBGE (2011).

Figure 4 shows the result after the regionalization by the Skater method of the census tracts present in Sub-region 4. Eight homogeneous socio-occupational groups were formed. The simulation expanded and allocated the original microdata to sectors and allowed for a much more detailed spatial distribution of job occupation, the main variable for the analysis of socio-occupational groups.

Figure 4– Regionalization result using the Skater method.



Source: The authors (2020).

The historical problem of Sub-region 4 explained above, is also reflected in the groups formed. Cruzeiro is the city with the largest census tract number (Table 3) and presents the largest population and GDP. Therefore, it is possible to observe that within the municipality boundaries there are a greater number of groups compared to others in the region.

For example, Group 1 is located only in the Cruzeiro municipality, because the socio-occupational characteristic of this group exists only within that census tract. It contains 57,516 people living in it, of which 59% are women and 41% men. About, 59% of the individuals declared themselves to be white, 21% brown, and 5% black. The group 1 population is very well divided between the age groups, however, the age groups with the highest people numbers are 10 to 20 years old (21%), 21 to 30 years old (18%), and 31 to 40 years old (16%), respectively. The average income is up to two minimum wages (36%), but almost 45% of the individuals do not have income, which is explained by the high individual's concentration with incomplete elementary education (38%). The job occupations with the highest percentages are unoccupied (60%), elementary occupations (8%) and service workers, salespeople in shops and markets (7%).

Otherwise, Groups 2 and 3 have very similar characteristics and are not part of the same group due to the contiguity criterion. These groups together contain 15,524 people, with an average of 56% women and 44% men. About 58% of the individuals declared themselves to be white, 35% brown, and 7% black. The age groups with the highest people are 10 to 20 years old (46%), 31 to 40 years old (14%), and 61 years old or more (10%), respectively. The average income is up to two minimum wages (37%), but 51% of individuals do not have income, as explained by the high presence of individuals with incomplete elementary education (53%). The job occupations with the highest percentages are unoccupied (71%), elementary occupations (11%), workers in the services that sell trades, and markets (5%).

In addition, Groups 5, 6, and 7 also have very similar characteristics in common and were the groups with more census tracts in different municipalities and portray the sub-region reality as a whole. These groups contain 22,312 people, on average 45% of them women and 55% men. About 50% of individuals declare themselves to be white, 38% brown, and 12% black. The age groups with the highest people are 10 to 20 years old (24%), 31 to 40 years old (16%), and 61 years old or more (17%), respectively. However, the other age groups contain percentages close to these, showing that this group is balanced between the studied ages. The average income is up to two minimum wages (55%), but 43% of individuals do not have income, explained by the high individuals with incomplete elementary education (62%). The job occupations with the highest percentages are unoccupied (55%), elementary occupations (16%), service workers selling trades, and markets (6%).

Then, Group 8 is found largely in the Lavrinhas municipality. It contains 8,645 people, of which 55% are women and 45% men. About 40% of individuals declare themselves to be white, 28% brown, and 5% black. The age groups with the highest people are 10 to 20 years old (40%), 21 to 30 years old (25%), and 51 to 60 years old (12%), respectively. The average income is up to two minimum wages (25%), but with 67% of individuals without income, explained by the high number of individuals with incomplete elementary education (47%). The job occupations with the highest percentages are unoccupied (71%), elementary occupations (9%) and service workers, salespeople in shops and markets (5%).

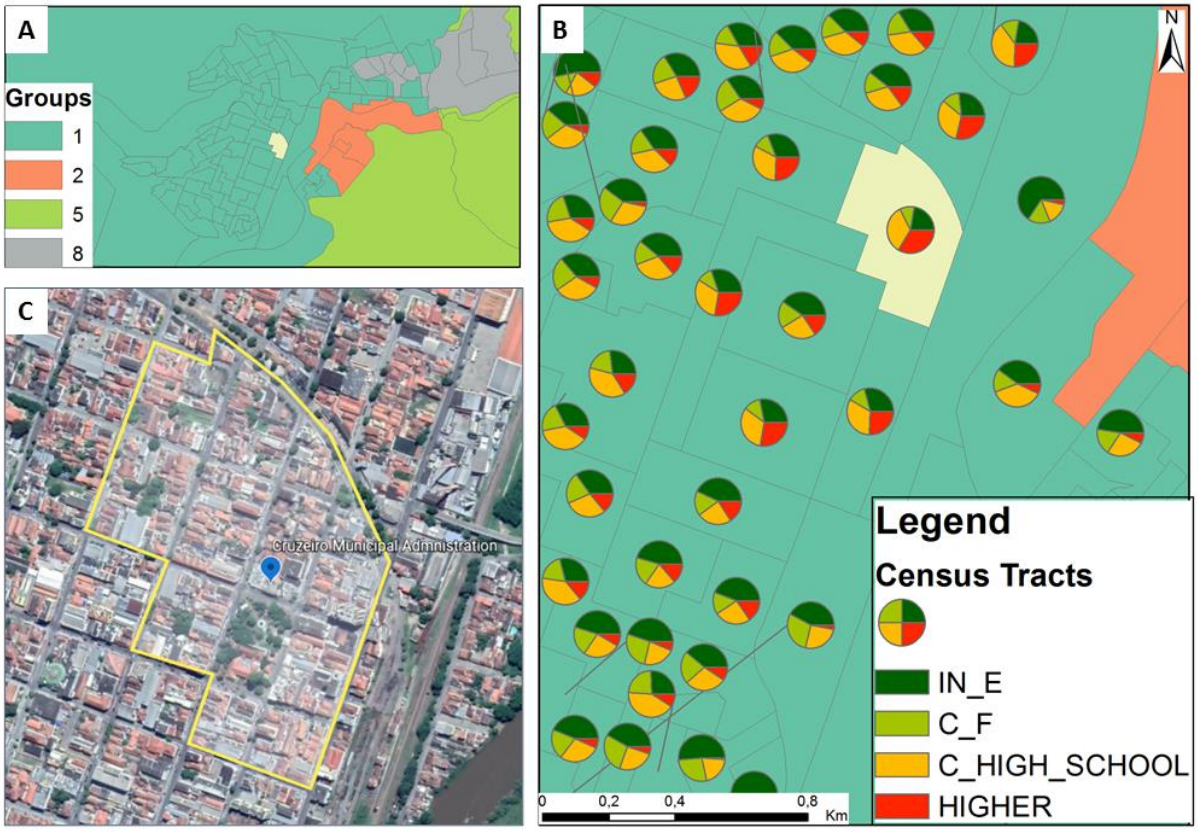
Finally, Group 4 is formed by two census tracts that do not contain information about the population, that is, they are blank and without data for confidentiality and secrecy reasons, if the sector has less than five records (households), this information is not included in the Demographic Census tables.

It is important to note that the men and women percentages do not remain across groups, as gender is one of the variables that were used as a criterion for grouping. Therefore, the method grouped census tracts with men and women percentages alike, with groups that together have more women and others that together have more men. The percentage of men is 46% and women is 54%, in sub-region 4. Another important point is that all groups had high percentages of people without an income job occupation. This is explained by the fact that the IBGE only considers job occupation and income variables in all jobs as people who work in formal labor, excluding those who are retired, pensioners, have informal work, etc.

Figure 5.A shows the most densely populated region in Group 1, where is found in the census sector with the highest education level in RMVPLN Subregion 4. In detail, Figure 5.B shows the education level

distribution by census sector in this region of Group 1, where the census sector "351340505000005" stands out, which showed the highest level of education. This sector presented 36% of residents with a university degree and 35% with complete high school and incomplete higher education, this reality differs from the rest of the Subregion under analysis where most people have an average education level without instruction and incomplete elementary school. Finally, Figure 5.C shows that the sector is located in the central region of the municipality of Cruzeiro, close to the city hall, which is considered a middle-class neighborhood.

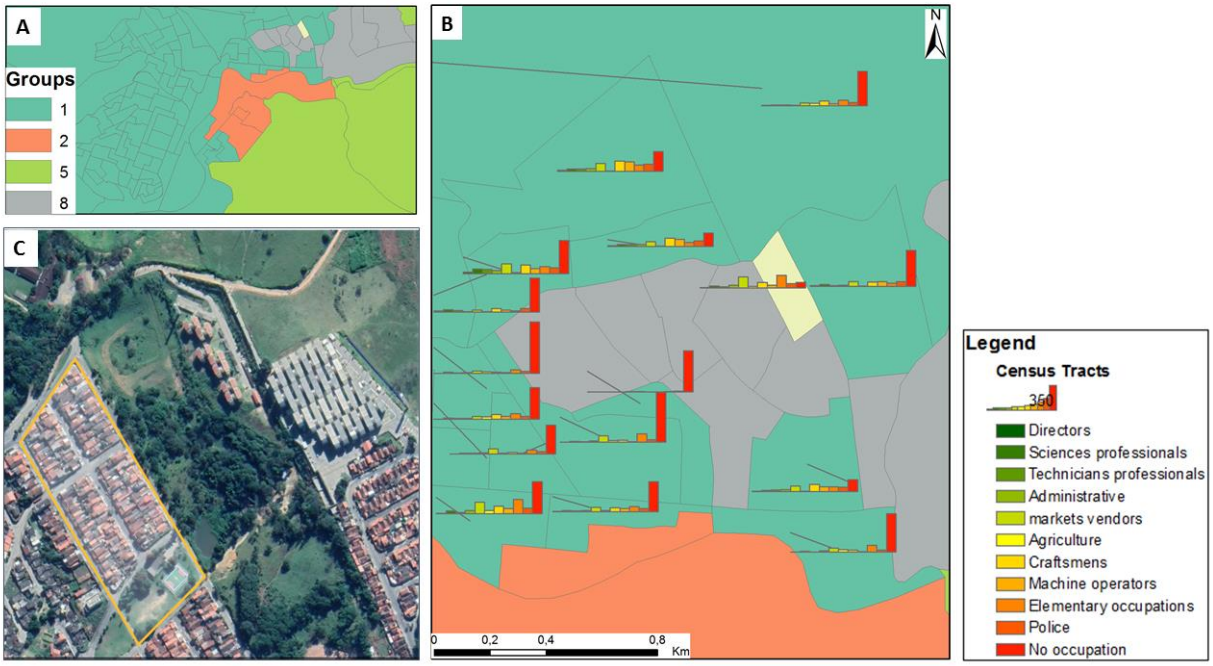
Figure 5 – Education level for Group 1. Legend: IN_E: Without instruction and incomplete elementary school; C_F: Complete elementary school and incomplete high school; C_HIGH_SCHOOL: Complete high school and incomplete higher education; HIGHER: Complete higher education.



Source: A) and B) The authors; C) Image obtained from Maxar Technologies CNES/Airbus (Google Earth).

Figure 6.A shows a Cruzeiro region located in Group 1, which is also densely located, however, it is located near the border with the Lavrinhas municipality. In this region, the census sector presents an interesting occupation distribution. Figure 6.B shows the job occupation distribution by census sector in this region, where the census sector "351340505000071" stands out, which presented an interesting distribution among occupations. This sector presented 24% of the residents with the occupation service workers, merchants, and market vendors and 26% with elementary occupations, however, when observing the occupations graph, a balance between them is perceived, different from that shown by the others census sectors. Finally, Figure 6.C shows that this sector location is very close to a popular set of buildings constructed by the government for the low-income population.

Figure 6– Job occupations for Group 1. Legend: Directors: Directors and managers; Sciences professionals: Sciences and intellectuals professionals; Technicians professionals: Mid-level technicians professionals; Administrative: Administrative support workers; market vendors: Service workers, merchants, and markets vendors; Agriculture: Skilled workers in agriculture, forestry, hunting, and fishing; Craftsmen's: Skilled workers, construction workers and craftsmen, mechanical arts and other crafts; Machine operators: Plant and machine operators and assemblers; Police: Members of the armed forces, police and military firemen.



Source: A) and B) The authors; C) Image obtained from Maxar Technologies CNES/Airbus (Google Earth).

The analyzes by census tracts presented in this study were only possible through the spatial micro-simulation technique since only the Census data SQ contains this information in the scale of weighting areas. With this information and presented in more detail in Figures 5 and 6, it is possible to plan a comprehensive analysis of the regions with job occupation lack, or low education levels, to understand what is happening in the area and thinking in some solutions to solve these problems. Preliminary studies like these reinforce the importance of the technique and how the spatial microdata enables analyzes that can have direct impacts on territorial planning at various scales at the RMVPLN Subregion 4.

4 CONCLUSIONS

This study showed how spatial microsimulation techniques introduce new socio-occupational groups possibilities in more detailed spatial units, allowing the phenomenon intra-urban analyzes. Although the data aggregated by the census tracts show good spatial resolution, there is no possibility of having variables such as "job occupation" at this level. 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 join into 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 elevated inequality degree within the subregion and consistently discriminating important population socioeconomic groups as presented in the more detailed analysis of group 1.

Additional testing should be performed to ensure that the spatial microdata resulting is as representative as possible within the data limitation. This requires exploring different constraint variables and validating the resulting estimates. It is also important to test and compare different spatial microsimulation methods, exploring their main characteristics, variability, and validity against the resulting external data sets, to arrive at a better estimate. In addition, modifying the neighborhood criteria for the neighborhood matrix

considered in this article to apply the Skater method may also offer improvements to the proposed socio-occupational structure.

This article is an extended and improved version of Oliveira, Anazawa, and Monteiro (2019), presented in XVII Brazilian Symposium on GeoInformatics (GEOINFO 2019).

Acknowledgement

The authors thank the financial support from Brazil's Coordination for the Improvement of Higher Education Personnel (CAPES) process 88882.330694/2019-1 and the São Paulo Research Foundation (FAPESP) process 2018/25525-2.

Authors contributions

Antonio Monteiro contributed to the conceptualization and together with Tathiane Anazawa managed and supervised the project. Gabriela Oliveira contributed to data curation, formal analysis, execution of the proposed methodology, validation, visualization and writing. Tathiane Anazawa ended the writing with revisions and editions.

Conflicts of interest

The authors declare no conflict of interest.

References

ASSUNÇÃO, R. M.; NEVES, M. C.; CÂMARA, G.; FREITAS, C. C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. **International Journal of Geographical Information Science**, p. 1–29, 2006. DOI. 10.1080/13658810600665111.

CAMARGO, E. C. G.; MONTEIRO, A. M. V. **Regionalização via Skater**: curso Análise Espacial de Dados Geográficos, nov. de 2010. 53 p. Disponível em: http://www.dpi.inpe.br/cursos/ser301/SlidesAulas/Aula_Edu_Skater_2010.pdf . Acesso em: 11 set. 2019.

EMPRESA PAULISTA DE PLANEJAMENTO METROPOLITANO (EMPLASA). Região metropolitana do Vale do Paraíba e Litoral Norte. Empresa Paulista de Planejamento Metropolitano S.A, **Imprensa Oficial do Governo do Estado de São Paulo**, São Paulo, 2012. Disponível em: <https://bibliotecavirtual.emplasa.sp.gov.br/ExibirDetalhes.aspx?funcao=kcDocumentos&id=2715&lingua=PT>.

EMPRESA PAULISTA DE PLANEJAMENTO METROPOLITANO (EMPLASA). Sobre a região metropolitana Vale do Paraíba e Litoral Norte. 2019. Disponível em: <https://emplasa.sp.gov.br/RMVPLN>. Acesso em: 5 jul. 2020.

FEITOSA, F.; JACOVINE, T. C.; ROSEMBACK, R. G. Small area housing deficit estimation: a spatial microsimulation approach. **Brazilian Journal of Cartography**, v. 68, n. 6, p. 1157–1169, 2016.

HERMES, K.; POULSEN, M. 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, 2012. DOI. 10.1016/j.compenvurbsys.2012.03.005.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Descrição das variáveis da Amostra do Censo Demográfico 2010.** Rio de Janeiro: IBGE, 2010a.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Classificação de Ocupações para Pesquisas Domiciliares - COD**. Rio de Janeiro: IBGE, 2010b.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Base de informações do Censo Demográfico 2010 :** Resultados do Universo por setor censitário. Rio de Janeiro: IBGE, 2011.

JACOVINE, T. C. **Estimativas de deficit habitacional para pequenas áreas**: uma proposta de abordagem baseada em microssimulação espacial. 195p. Dissertação (Mestrado em Planejamento e Gestão do Território) - Universidade Federal do ABC, São Bernardo do Campo, 2017.

LOVELACE, R. et al. Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. **Journal of Artificial Societies and Social Simulation**, v. 18, n. 2, p. 1–15, 2015. DOI. 10.18564/jasss.2768.

LOVELACE, R.; DUMONT, M. **Spatial Microsimulation with R**. Chapman & Hall/CRC The R Series, 2016.

MARIA, J. M. **Região e regionalização:** estudo da região metropolitana do Vale do Paraíba e Litoral Norte. 43p. Monografia (Bacharelado em Geografia) - Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista Júlio de Mesquita Filho, Rio Claro, 2016.

MIRANTI, R. et al. Measuring small area inequality using spatial microsimulation: Lessons learned from Australia. **International Journal of Microsimulation**, v. 8, n. 2, p. 152–175, 2016. DOI. 10.34196/ijm.00118

MÜLLER, N. L. **O Fato Urbano na Babia do Rio Paraíba, Estado de São Paulo**. Rio de Janeiro: Fundação IBGE, 1969.

NASCIMENTO, R. P. **Características regionais e oportunidades locais na formação de mão de obra:** análise comparativa de duas sub-regiões do vale do Paraíba Paulista. 2015. 164p. Dissertação (Mestrado em Planejamento e Desenvolvimento Regional do Programa) - Universidade de Taubaté, Taubaté, 2015.

NASCIMENTO, R. P.; RICCI, F.; RODRIGUES, M. D. S. Desenvolvimento endógeno da região metropolitana do Vale do Paraíba e Litoral Norte: uma análise do quociente locacional. In: CONGRESSO INTERNACIONAL DE CIÊNCIA, TECNOLOGIA E DESENVOLVIMENTO, 3., 2014. **Anais...** 2014. p.19

PHILIPPINI, R. A. S. **Fazenda de café do Vale Histórico:** perspectiva de práticas educativas de história e cultura afro- brasileiras em espaços não formais de educação. 2019. 154p. Dissertação (Mestrado em Educação e Desenvolvimento Humano) - Universidade de Taubaté, Taubaté, 2019.

QUADROS, W. J. DE; MAIA, A. G. Estrutura sócio-ocupacional no Brasil. **R. Econ. contemp.**, v.14, p. 443–468, 2010.

RIBEIRO, L. C. DE Q.; LAGO, L. C. DO. O espaço social das grandes metrópoles brasileiras: São Paulo, Rio de Janeiro e Belo Horizonte. **Revista Brasileira de Estudos Urbanos e Regionais**, n. 3, p. 111, 2000. DOI. 10.22296/2317-1529.2000n3p111.

WHITWORTH, A. (org). **Evaluations and improvements in small area estimation methodologies**. Sheffield: University of Sheffield, 2013.

Main author biography

Gabriela Carvalho de Oliveira was born in Cachoeira Paulista - SP in 1995. She holds a bachelor's degree in Environmental Engineering from São Paulo State University (Unesp), Institute of Science and Technology, São José dos Campos and a master's degree in Remote Sensing from the Brazilian Institute for Space Research (INPE). Previously, she worked as a research assistant and GIS consultant, currently she works as an environmental consultant. Her research interests include the use of GIS and new techniques of spatial analysis, such as microsimulation, applied to terrotorial planning..