DOI 10.1590/s1982-21702019000100005

ORIGINAL ARTICLE

STRATEGY FOR EXTRACTION OF FOURSQUARE'S SOCIAL MEDIA GEOGRAPHIC INFORMATION THROUGH DATA MINING

Estratégia de Extração de Informações Geográficas de Mídia Social do Foursquare por Mineração de Dados

Paula Fernandez Costa[1] – ORCID: 0000-0002-9225-0511

Irving da Silva Badolato[1]

Rogério Luís Ribeiro Borba[3]

Julia Celia Mercedes Strauch[4, 5]

[1] Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia, Departamento de Engenharia Cartográfica, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: paulafc22@gmail.com; irvingbadolato@gmail.com

[3] Instituto Brasileiro de Geografia e Estatística, Diretoria de Geociências, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: rogerio.borba@ibge.gov.br

[4] Escola Nacional de Ciências Estatísticas, Programa de Pós-Graduação *Stricto Sensu* em População, Território e Estatísticas Públicas, Rio de Janeiro, Rio de Janeiro, Brasil

E-mail: Julia.strauch@ibge.gov.br

[5] Universidade Federal Fluminense, Instituto de Geociências, Departamento de Análise Geoambiental, Niterói, Rio de Janeiro, Brasil

Received in May 24th 2018

Accepted in September 16th 2018

Abstract:

This aim of this paper is the acquisition of geographic data from the Foursquare application, using data mining to perform exploratory and spatial analyses of the distribution of tourist attraction and their density distribution in Rio de Janeiro city. Thus, in accordance with the Extraction, Transformation, and Load methodology, three research algorithms were developed using a tree hierarchical structure to collect information for the categories of Museums, Monuments and Landmarks, Historic Sites, Scenic Lookouts, and Trails, in the foursquare database. Quantitative analysis was performed of check-ins per neighborhood of Rio de Janeiro city, and kernel density (hot spot) maps were generated The results presented in this paper show the need for the data filtering process — less than 50% of the mined data were used, and a large part of the density of the Museums, Historic Sites, and Monuments and Landmarks categories is in the center of the city; while the Scenic Lookouts and Trails categories predominate in the south zone. This kind of

How to cite this article: COSTA, P. F.; BADOLATO, I. S.; BORBA, R. L. R. and STRAUCH, J. C. M. Strategy for Extraction of Foursquare's Social Media Geographic Information Through Data Mining. *Bulletin of Geodetic Sciences*, 25(1): e2019005, 2019.

analysis was shown to be a tool to support the city's tourist management in relation to the spatial localization of these categories, the tourists' evaluations of the places, and the frequency of the target public.

Keywords: ubiquitous cartography; Social Media Geographic Information; data mining; Foursquare

Resumo:

Este trabalho tem por objetivo a aquisição de dados geográficos do aplicativo Foursquare, empregando mineração de dados para efetuar análises exploratórias e espaciais da distribuição de pontos turísticos e de sua distribuição de densidade no município do Rio de Janeiro. Desta forma, seguindo a metodologia de Extração, Transformação e Carga, foram desenvolvidos três algoritmos de pesquisa usando a estrutura hierárquica de árvore para coletar na base de dados do Foursquare informações das categorias de Museu, Monumento e Marco, Patrimônio Histórico, Vista Panorâmica e Trilha e efetuada análise quantitativa de check-ins por bairros do município do Rio de Janeiro e gerados mapas de densidade Kernel (mapas de calor). Os resultados apresentados nesse trabalho mostram a necessidade do processo de filtragem de dados, sendo que menos de 50% total dos dados minerados foram utilizados e que grande parte da densidade de Museu, Patrimônio Histórico e Monumento e Marco encontram-se no Centro da cidade e Vista Panorâmica e Trilha predominam na Zona Sul. Este tipo de análise mostrou-se uma ferramenta de apoio a gestão turística do município quanto à localização espacial dessas categorias, as avaliações do turista sobre os locais e sobre a frequência do público alvo.

Palavras-chave: Cartografia Ubíqua; Informação Geográfica por Mídia Social; Mineração de Dados; *Foursquare*.

1. Introduction

In recent years, Web 2.0 has led to the development of ubiquitous collaborative platforms available on mobile devices connected to the internet. In this new paradigm is Web Cartography (CAMPAGNA et al., 2015; GARTNER, G.; HUANG, 2016) and Ubiquitous Cartography (GOODCHILD, 2007; PETERSON, 2008; SCHUURMAN, 2009, GARTNER, G.; HUANG, 2016; USERY and VARANKA, 2018). The first is concerned with the importance of user-centered interface design, the design of dynamic map content, and mapping functions, as well as the liberation of the power to create maps for the public and amateur cartographers. Cartography, in its ubiquitous personality, concerns the viability of the interconnected user having access to maps anywhere and anytime (Gartner, 2007).

These ubiquitous collaborative platforms provide resources for people to generate and share information and multimedia online or almost in real time (Bartolomé, 2008). The information and multimedia are georeferenced and presented in a user-oriented application interface in which dynamic map content and mapping functions are made available. This offers, in real time, the growing possibility for the public and amateur cartographers to create various types of maps.

Thus, interconnected users using these platforms register their own locations of interest and produce maps that assign distinct value in spatial and temporal content. In this context, users are transformed into producers of geographic data with voluntary activities (Campagna et al., 2015). The term volunteered geographic information (VGI) was introduced by Goodchild (2007) to designate the role of society as a producer of geographic information — a function that for centuries had been performed by official government institutions.

This depicts a transition process that — according to the Cooperative Research Center for Spatial Information of Australian and New Zealand (CRCSI, 2017) — is expected to occur over the next five years, requiring new practices and innovation in government and the private sector in order to capture the power of emerging technologies and meet the future demands of users. In this transition, the focus of government should change the process of providing data to society. Data supply moves in the direction of a more collaborative and diversified information management environment in partnership with many data providers. This environment will be focused on knowledge, by increasingly providing the automated sharing of data resources, offering open and advanced analytical tools. It should be noted that in this new context, the analysis of voluntary data integrated with official data allows a better understanding of reality. This makes it possible to support design and decision making to make plans closer to the social, economic, and environmental reality (Campagna et al., 2015).

Among the forms of VGI, the data available from the Foursquare application are considered to be Social Media Geographic Information (SMGI), which is a subcategory of VGI. SMGI can be defined as the information generated by these social media platforms — multimedia content with explicit (i.e., coordinated) or implicit (i.e., place names, toponyms) information generated by the user through social media or mobile applications (Campagna et al. 2016).

Thus, based on these concepts, this work aims to mine data from the Foursquare application for the analysis of tourist attraction in the city of Rio de Janeiro, in order to identify the neighborhoods most frequented by tourists. Foursquare was chosen for this application because it is an internationally known application, and it was widely used in the international events in Rio de Janeiro in 2014 (Soccer World Cup) and 2016 (Olympic Games). Foursquare is a tourism-oriented platform that allows tourists to record, indicate, evaluate, and suggest places for excursions and dining — among other things — that are of most relevance for tourists. These locations registered in the platform are called venues, and Foursquare offers a public Application Program Interface (API) for their extraction. However, the maximum number of venues returned from a search in this API is 50; that is, when performing the search, the API returns the first 50 venues found within the limits of the specified area, and if there are more venues found in the search, the API does not list these. Thus, this work presents Foursquare's SMGI extraction strategies. The data to be mined from the application consist of locations with the categories of Museums, Historic Sites, Monuments and Landmarks, Trails, and Scenic Lookouts, in order to compose a geographic database. The acquisition of these data will enable exploratory and spatial analyses of the distribution of the main tourist attractions and their density distribution in the city of Rio de Janeiro, and the evaluation of the tourists in relation to these places that the city offers.

For a better understanding of this work, the second section describes the methodology and implementation, the third section presents the results, and the fourth section presents the final considerations.

2. Methodology and Implementation

The methodology used in this study aimed to systematize: i) the extraction of Foursquare data, ii) transformation of the data through their treatment and cleaning in order to meet demands; and (iii) the loading of the data organized into a geographical database. This methodology — known as Extract, Transform, and Load — is widely used in data warehouses (Simitsis and Vassiliadis, 2003).

Due to the restriction of the Foursquare API returning a maximum of 50 venues per search area in a spatial query, data mining was done using the concept of tree data structure. This led to the implementation of three strategies to perform the spatial division of the study region in order to scan the space and obtain all the information registered. Thus, three systems were developed in Python for the acquisition of Foursquare information: i) using a grid, ii) using Quadtree search structures, and iii) K-D tree. Based on the Quadtree algorithm, an automated model for extracting information from the application was proposed, for use in the geoprocessing environment with Python.

With the data acquired, exploratory spatial analyses were then performed in order to understand the distribution of tourists' visits in the city of Rio de Janeiro.

2.1 Foursquare Registration and Search Parameters

Initially, registration was done on the Foursquare developer website and an application project was created to obtain the accreditation keys — *ClientID* and *ClientSecret*, which are used as a requirement of the OAuth 2.0 authentication protocol. *ClientID* is considered to be public information. It is the application's identifier that is unique to the clients of the authorization server. *ClientSecret* is a confidential password used for web applications. These two parameters are required to connect to the Foursquare server.

In order to perform the search, it was necessary to use *intent:browse* as a search intent parameter — it determines the search of the venues within a defined area. This parameter requires other input parameters to define the search area by the rectangle that is delimited by northeast (NE) and southwest (SW) vertices.

The *intent:browse* parameter differs from other parameters such as *intent:global*, which only returns venues considered to be relevant independent of the location or the *near* parameter, which returns venues via the name of the location informed; however, the return criterion is unknown.

The other search parameters used were the category identification code (i.e., *categoryID*) and the return limit number for the venues. The latter is due to Foursquare's API limiting the number of venues it returns from a search. If this parameter is not specified, the number of venues returned is not necessarily the maximum (50 venues).

2.2 Extraction Algorithms for Foursquare's API

Initially, a code was implemented based on the grid search structure, in an attempt to obtain all the venues of Rio de Janeiro in the categories of interest for tourism management in Rio de Janeiro city: Museums, Historic Sites, Monuments and Landmarks, Trails, and Scenic Lookouts. This code is in accordance with the principle of spatial division into squares of dimensions of 0.1° of the total area delimited, in order to extract the information from the 4 x 7 subareas, as shown in Figure 1. The algorithm follows the programming logic presented by the pseudo code located in Figure 2.

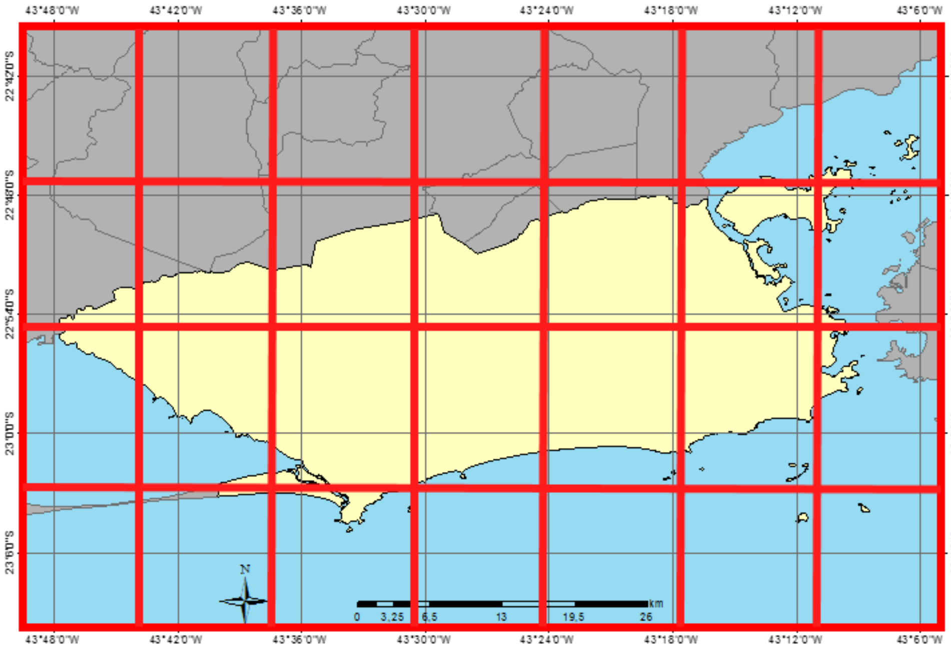


Figure 1: Spatial division of Rio de Janeiro city into grids

The inefficiency of the data mining done using the grid search structure was evident when considering the high number of venues lost due to the API returning only a maximum of 50 venues. Since the points are not uniformly distributed, the uniform grid will be very small in some areas, where there are no points, or too large, when more than 50 results return. Thus, codes were developed with search based on Quadtree and K-D Tree structures, which enabled subdivision of the area when reaching the 50 venue limit.

The Quadtree algorithm follows the principle of performing four spatial subdivisions of the area that delimits the city when the return from the search is equal to or greater than the established node of the tree. The polygon used to mark the total area is the same as the grid algorithm. The algorithm considers the restriction of the Foursquare API in returning a maximum of 50 venues in the search area, and when it reaches this, the algorithm introduces a new spatial division node. Likewise, the algorithm limits the depth of the search tree up to ten spatial subdivisions. This depth variable was implemented in order to eliminate the risk of the system entering an infinite cycle of area divisions in the case of a region small enough that the system could not divide it with 50 or more venues, in which case there would be no stop conditions for the algorithm. This algorithm follows the programming logic presented by the pseudo code of Figure 3.

During the code execution process, it could be seen that the API behaves differently (i.e, the total number of venues returned is different) according to the variation in the node established in the code. The variation in the result for the different nodes established was due to the Foursquare API. Due to not having access to the API code, it was not possible to determine the cause of this problem, only to avoid it with the spatial division methods. Therefore, the data mining process used considered the values of 30, 20, 10 and 1 for a maximum number of nodes (*NodeDepth*) in order to study the most efficient and appropriate method for acquiring all the information of interest. For example, in the case of a node value of 30, the area is subdivided until 30 or fewer venues are returned; whereas, in the case of node value of 1, the area is subdivided until only 1 venue is returned. The node value determines whether or not the search tree will perform a spatial division. Thus, for the Quadtree and K-D Tree algorithms implemented, the tree will continue to perform spatial divisions until the information of the obtained node is equal to or less than the value of the pre-established node.

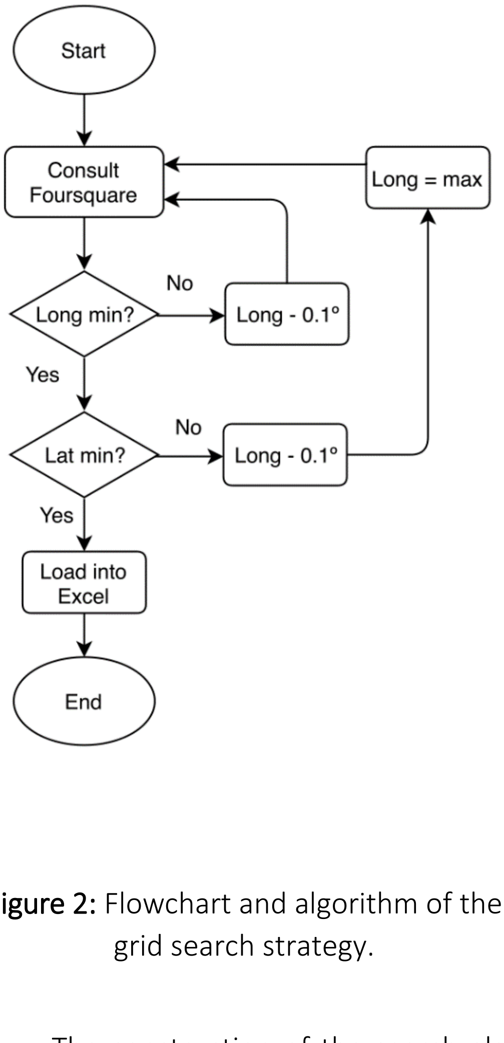


Figure 2: Flowchart and algorithm of the grid search strategy.

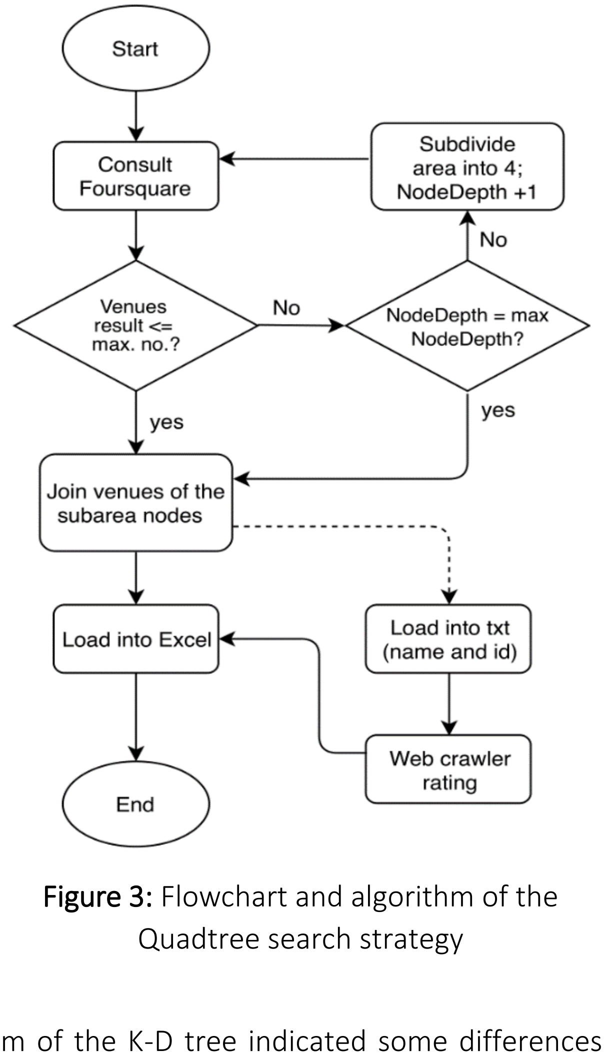

Start
Consult Foursquare
Subdivide area into 4; NodeDepth +1
No
Venues result <= max. no.?
No
NodeDepth = max NodeDepth?
yes
yes
Join venues of the subarea nodes
Load into Excel
Load into txt (name and id)
Web crawler rating
End

Figure 3: Flowchart and algorithm of the Quadtree search strategy

The construction of the search algorithm of the K-D tree indicated some differences in relation to the Quadtree. The K-D tree method involves two subdivisions; that is, dividing the area in half. However, this is done by alternating the subdivision by axis (latitude and longitude) for each level of the tree, following the same node and depth criteria of the Quadtree. The K-D tree algorithm and the pseudo code are illustrated in Figure 4.

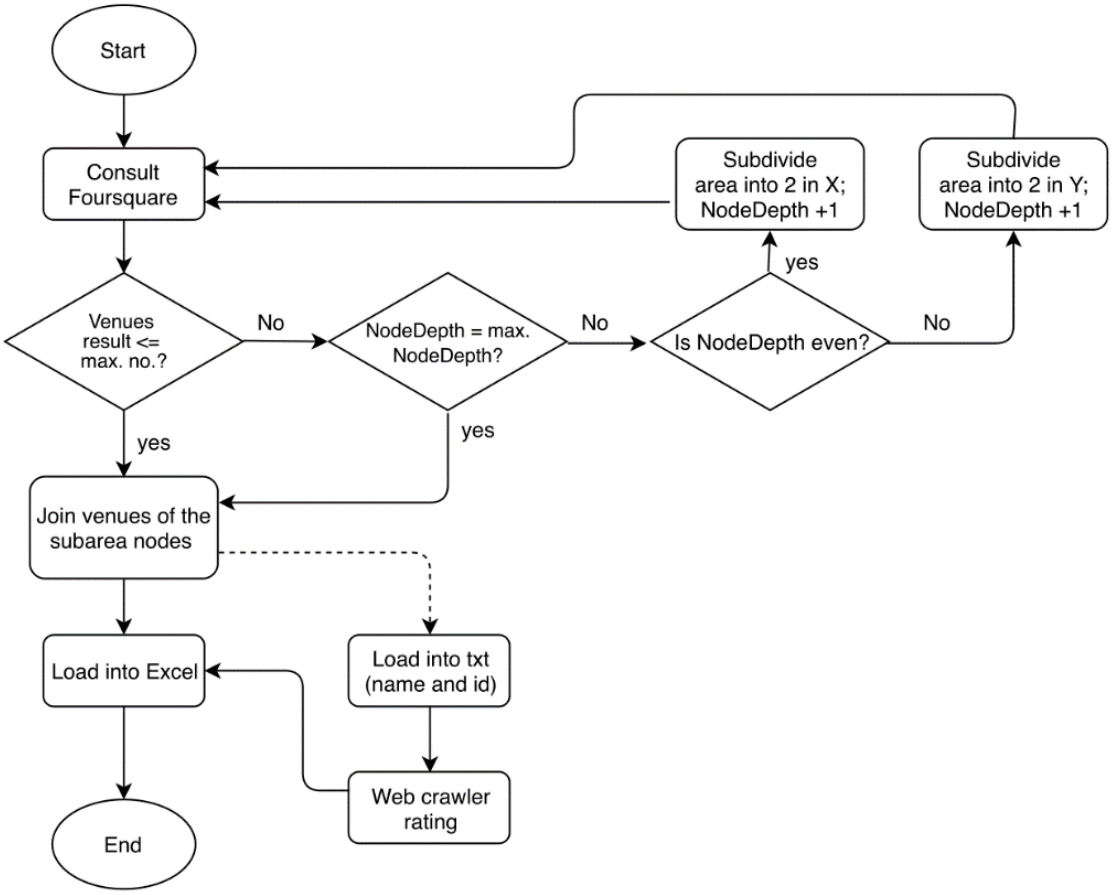
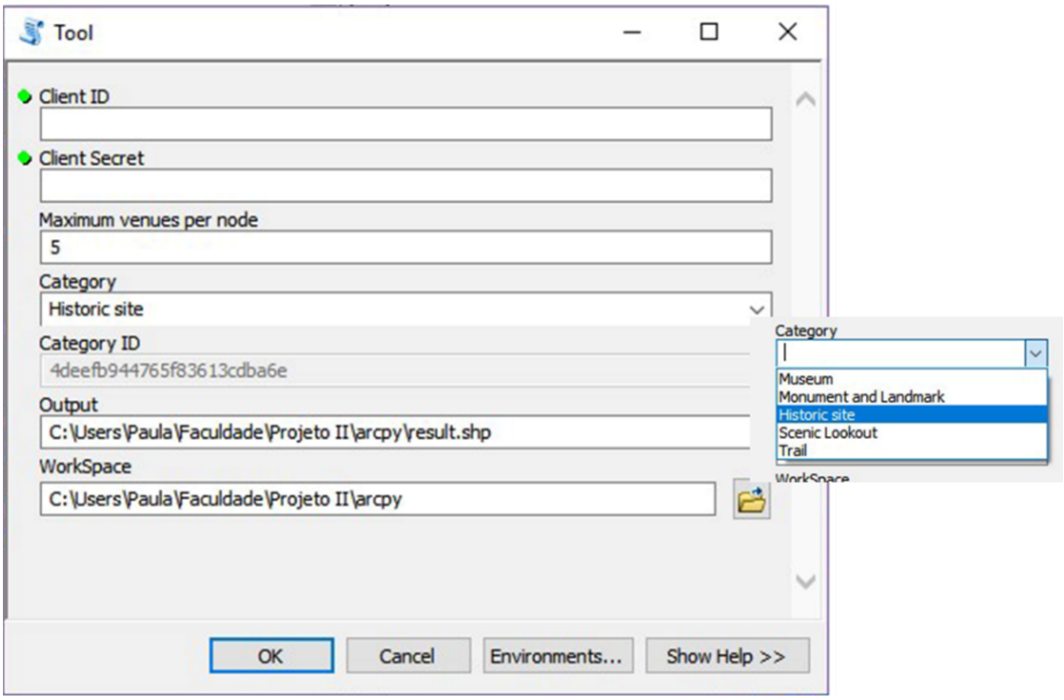

Start
Consult Foursquare
Subdivide area into 2 in X; NodeDepth +1
Subdivide area into 2 in Y; NodeDepth +1
yes
Venues result <= max. no.?
No
NodeDepth = max. NodeDepth?
No
Is NodeDepth even?
No
yes
yes
Join venues of the subarea nodes
Load into Excel
Load into txt (name and id)
Web crawler rating
End

Figure 4: Flowchart of the K-D tree algorithm.

For all three algorithms, the data were exported to an Excel worksheet. However, in order to use them in geoprocessing software, they must first be transformed into points in a coordinate system in shapefile format, through latitude and longitude attributes. In order to automate this process, a tool was implemented for automating the extraction from the Foursquare API, using programming in Python. The script programmed has a Quadtree search algorithm for Foursquare's API, with input parameters as shown in Figure 5: *ClientID* and *ClientSecret*; node value; and a list of categories to be searched in a drop-down list (Museums, Monuments and Landmarks, Historic Sites, Scenic Lookouts, and Trails).

As shown in Figure 5, upon selecting the *Category* parameter, the category ID is automatically informed as an uneditable parameter. In this figure, the *Output* parameter is the location and name of the output shapefile, and *WorkSpace* parameter is the address of the folder where the shapefile is located.

Thus, when implemented as a toolbox in a geographic information system, the script automatically generates the shapefile points in the SIRGAS 2000 geographic coordinate system, including the data attributes table.



Client ID

Client Secret

Maximum venues per node: 5

Category: Historic site

Category ID: 4deefb944765f83613cdba6e

Output: C:\Users\Paula\Faculdade\Projeto II\arcpy\result.shp

WorkSpace: C:\Users\Paula\Faculdade\Projeto II\arcpy

Figure 5: Input parameters of the script used in Foursquare's API, as well as details of the Category menu in a drop-down list.

2.3 Database

To enable the analysis of the data, a geographic database was created from the model with the classes described using the Object Modeling Technique for Geographic Applications (OMT-G) — see Figure 6. The venue categories used in the search were as follows: Museums, Historic Sites, Monuments and Landmarks, Scenic Lookouts, and Trails. Venues have several attributes that can be accessed in data mining; however, the algorithm constructed for this study was limited to obtaining the location attributes (latitude and longitude), name, venue ID, check-in number, comment number (tip), category name that it belongs to, and category ID.

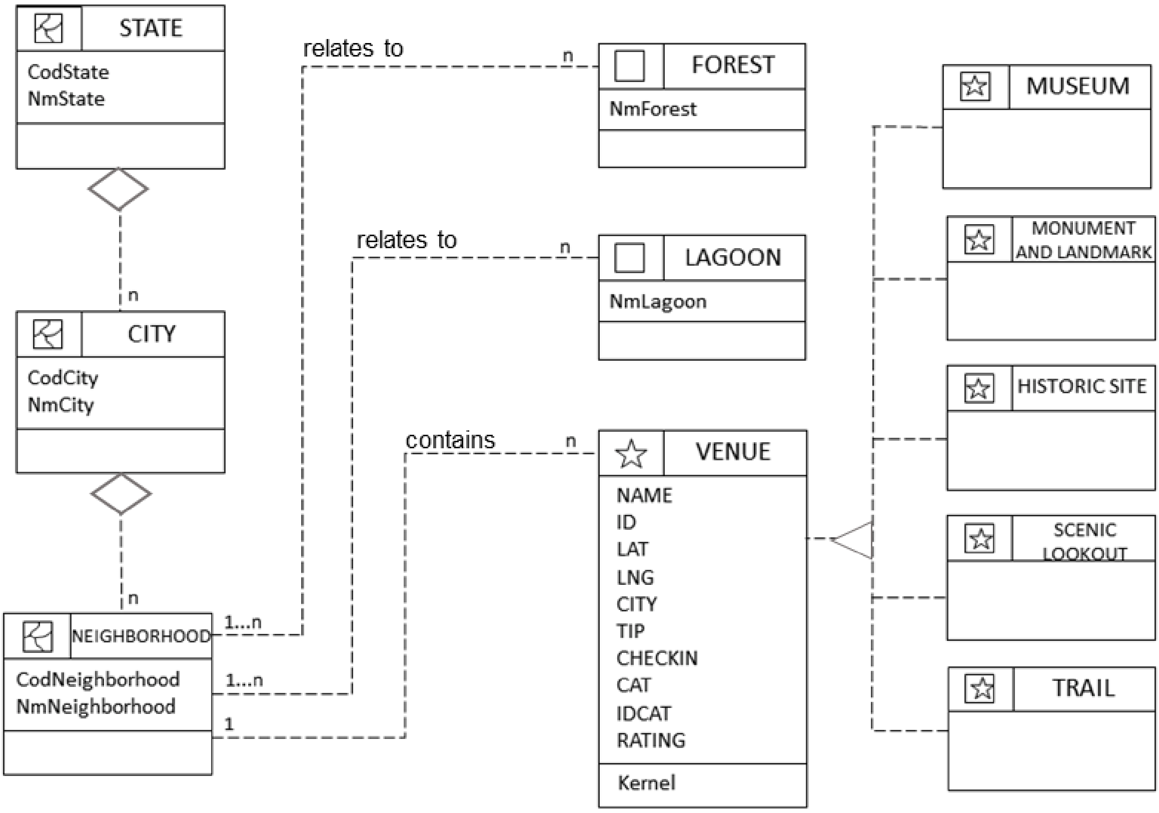


Figure 6: Diagram of OMT-G class

2.4 Acquisition of the tourists' evaluations

Foursquare uses an algorithm to rank the venues based on users evaluations of the places, based on a variety of signals derived from users' interactions with venues. These signals may be explicit or implicit. In the first case, for example, the tourist leaves a positive or negative comment explaining the like or dislike rating. In the second case, the fact that a location has many loyal customers with experiences there, increases its credibility. However, it must be noted that not all sites are ranked, which may be due to the type of place or because Foursquare cannot gather enough information to generate the evaluation.

Although the user ratings of locations are an attribute of the venue in the API, it turns out that this value is not returnable in data mining. However, these ratings are viewed on the webpages of venues. Thus, it was necessary to implement a code to track the network and get the users' ratings of the locations. For this task, an algorithm named GenerateFile () was implemented to create two text files: one containing the list of names, and the other with the ID list of each venue. It was also necessary to implement a parsing routine to analyze the list of location names and modify them to the structure used by Foursquare to describe Web addresses. For example, Museu do Amanhã (Museum of Tomorrow) would become “museu-do-amanha” in the parsing process. These two text files form the elements that make up the URL of the web page for a venue, and they are read by the program that tracks the network in order to find user ratings for that location. For this code, we used Beautiful Soup, which is a Python library that parses HTML and

XML documents. Subsequently, the evaluations (ratings) obtained via this process were added to the attribute tables of venues contained in the geographic database.

2.5 Data refinement and form of results representation

A filtering process was required to perform the analysis of Rio de Janeiro, because much of the information obtained in the categories was not in accordance with the proposal of this work — it could be seen that some venues are in neighboring municipalities and that some places have few check-ins or have no evaluations. Thus, a subset was created to be analyzed via the following criteria (shown in Figure 7): i) venues that are in the city of Rio de Janeiro, and ii) venues that had more than 50 check-ins or a rating higher than zero. For the first criterion, the clip operation was used; while for the second criterion, a selection of attributes was made for venues with more than 50 check-ins or that had a non-zero evaluation.

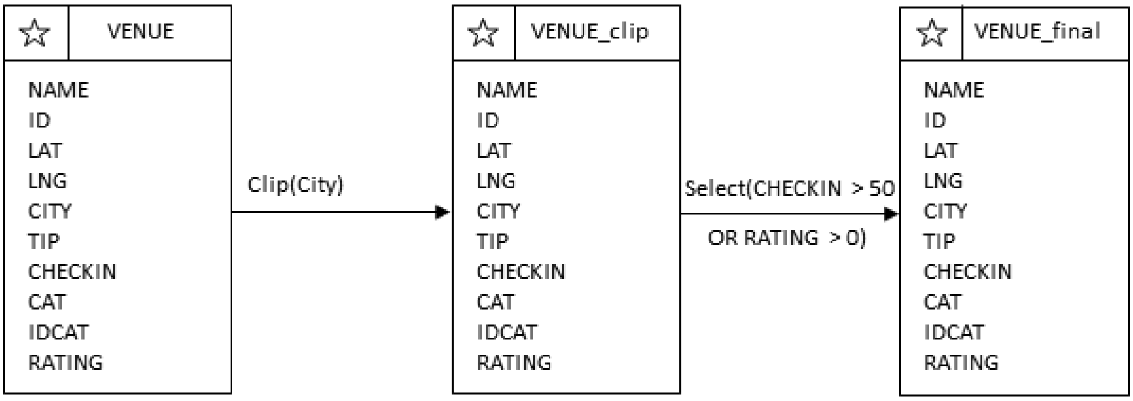


Figure 7: Diagram of OMT-G transformation

In order to perform the exploratory and spatial analyses of the distribution of the main tourist attraction and their density distribution in the city of Rio de Janeiro, density estimation maps (using the Kernel method) and quantitative thematic maps were prepared for each category for check-in numbers by neighborhood. The density maps aimed for a holistic observation of the distribution of venues in the city by category; while the quantitative thematic maps provided an objective view in relation to the neighborhoods most visited by tourists. The representations of the thematic classes of these maps are shown in Figures 8 and 9.

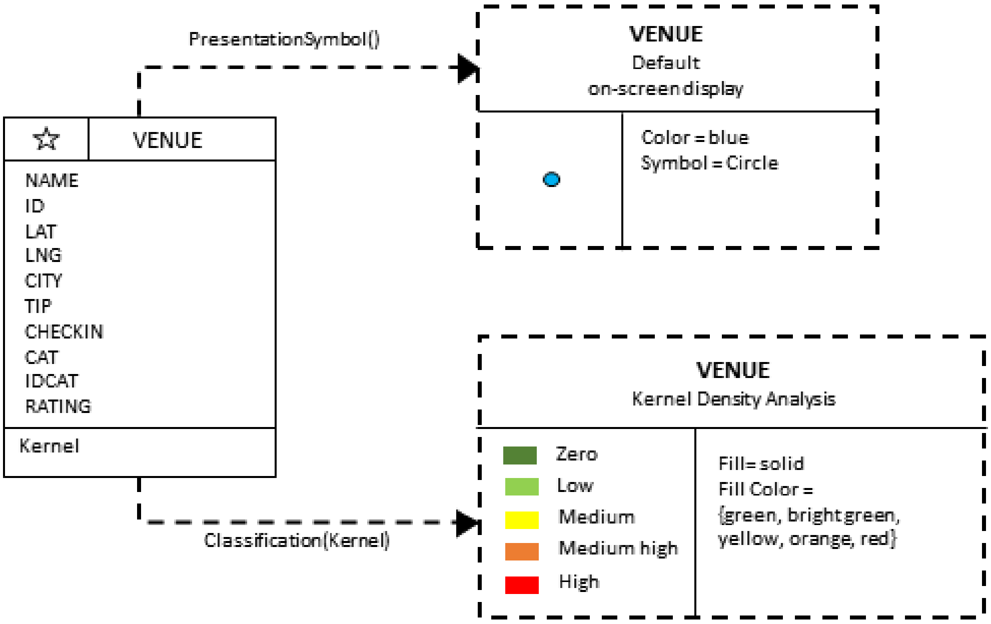


Figure 8: Diagram of OMT-G presentation for the venue density map.

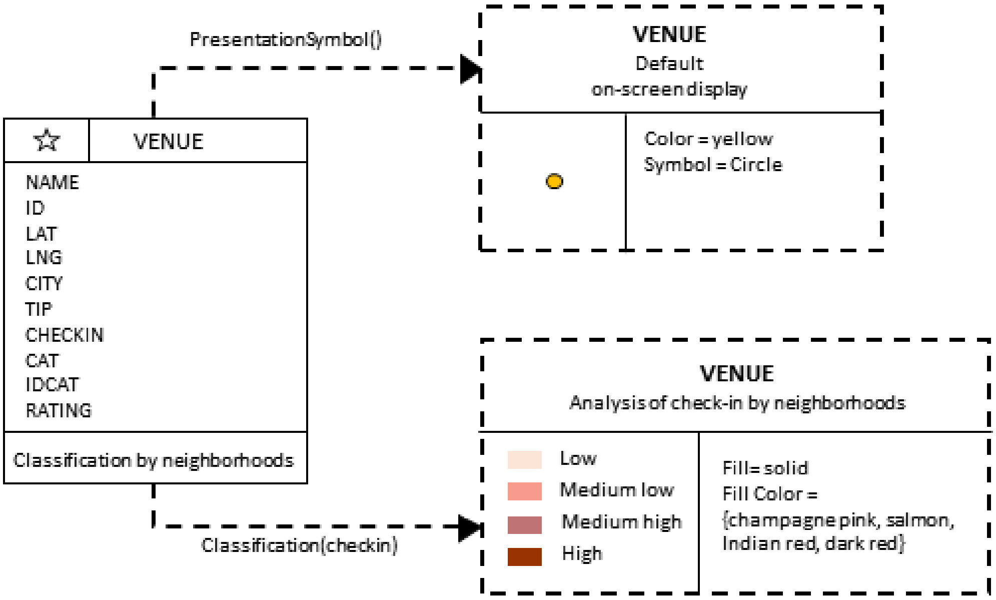


Figure 9: Diagram of OMT-G presentation for the classification of check-in by neighborhood.

3. Results and Discussion

On May 30, 2017, the Grid, Quadtree, and K-D tree search algorithms were applied to Foursquare's API. As described in Table 1, a different numbers of venues was returned for each algorithm, in particular for the Quadtree and K-D tree search algorithms that implemented nodes between 30, 20, 10, and 1.

Table 1: Number of venues returned, per category, for the Grid, Quadtree, and K-D tree algorithms.

Category	Grid	Quadtree venues				K-D tree venues			
	venues	ND 30	ND 20	ND 10	ND 1	ND 30	ND 20	ND 10	ND 1
Museums	119	86	135	166	182	81	130	166	182
Monuments	72	28	54	72	72	28	34	52	72
Historic Sites	239	158	158	301	344	140	140	301	344
Scenic Lookouts	157	238	238	238	238	124	181	238	238
Trails	137	24	138	138	138	24	45	138	138

Analyzing the results shown in Table 1, it could be seen that at its initial nodes depth (ND) of 30 and 20, the Grid algorithm had results better than or similar to the Quadtree and K-D tree algorithms. However, as the value of the node depth of the two other algorithms decreases, the inefficiency of the Grid algorithm can be seen for the search of venues in the city. Similarly, when comparing the results obtained by the Quadtree and KD tree algorithms described in Table 1, in their variations for the nodes depth between 30 and 10, it could be seen that the return for the Quadtree algorithm is slightly superior to that of the KD tree algorithm, since it returns more venues. However, for node depth 1, both attained the same number of venues for all the categories. The analysis by category (Figure 10) shows that the largest number of check-ins is in the Historic Sites category, followed by Scenic Lookouts, Museums, Monuments and Landmarks, and Trails.

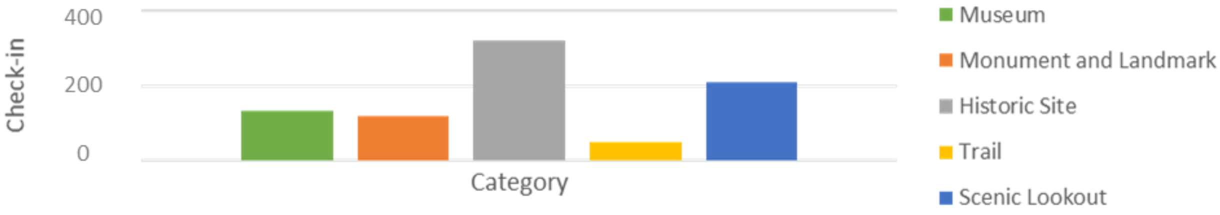


Figure 10: Foursquare's check-ins, by tourism category, in Rio de Janeiro city.

The data returned from the Quadtree algorithm with one node were filtered considering the following criteria: i) the venues within the limits of the Rio de Janeiro city, and ii) those that had more than 50 check-ins or a rating higher than zero. Thus, in the analysis, we considered a subset of the venues found by the Quadtree algorithms for node one, according to the quantitative described in Table 2 for each category. It could be seen that, on average, 85% of the data were in Rio de Janeiro city and that only 31% were considered by the second criterion, and in the filtering, more than 50% of the data acquired for each category were not used in the spatial analysis process, due to being outside the study area and they are not meaningful in this analysis.

Table 2: Number of venues selected for spatial analysis.

	Quadtree venues — Node 1				
Category	Total	In Rio	%	Check-in > 50 or rating > 0	%
Museums	182	157	86.26	85	46.70
Monuments	72	58	80.56	25	34.72
Historic sites	344	276	80.23	78	22.67
Scenic lookouts	238	205	86.13	77	32.35
Trails	138	125	90.58	29	21.01

Table 3 was prepared by analyzing the list of the 10 venues with the highest rating, and using the highest number of check-ins as the tie-breaking criterion. It should be noted that among the venues in the Museums category, many are repeated in the Monuments and landmarks, and Historic Sites categories, and vice versa; for example, the Paço Imperial (Imperial Palace) is in the Museums category (it is currently a Cultural Center, in which various types of exhibition are held), but it is also in the Monuments and Landmarks category, due to being constructed in 1873 as the home of the Viceroys of Brazil. Similarly, the Palácio Tiradentes is in both the Museums and Historic Sites categories.

Table 3: Order of venues with highest ratings

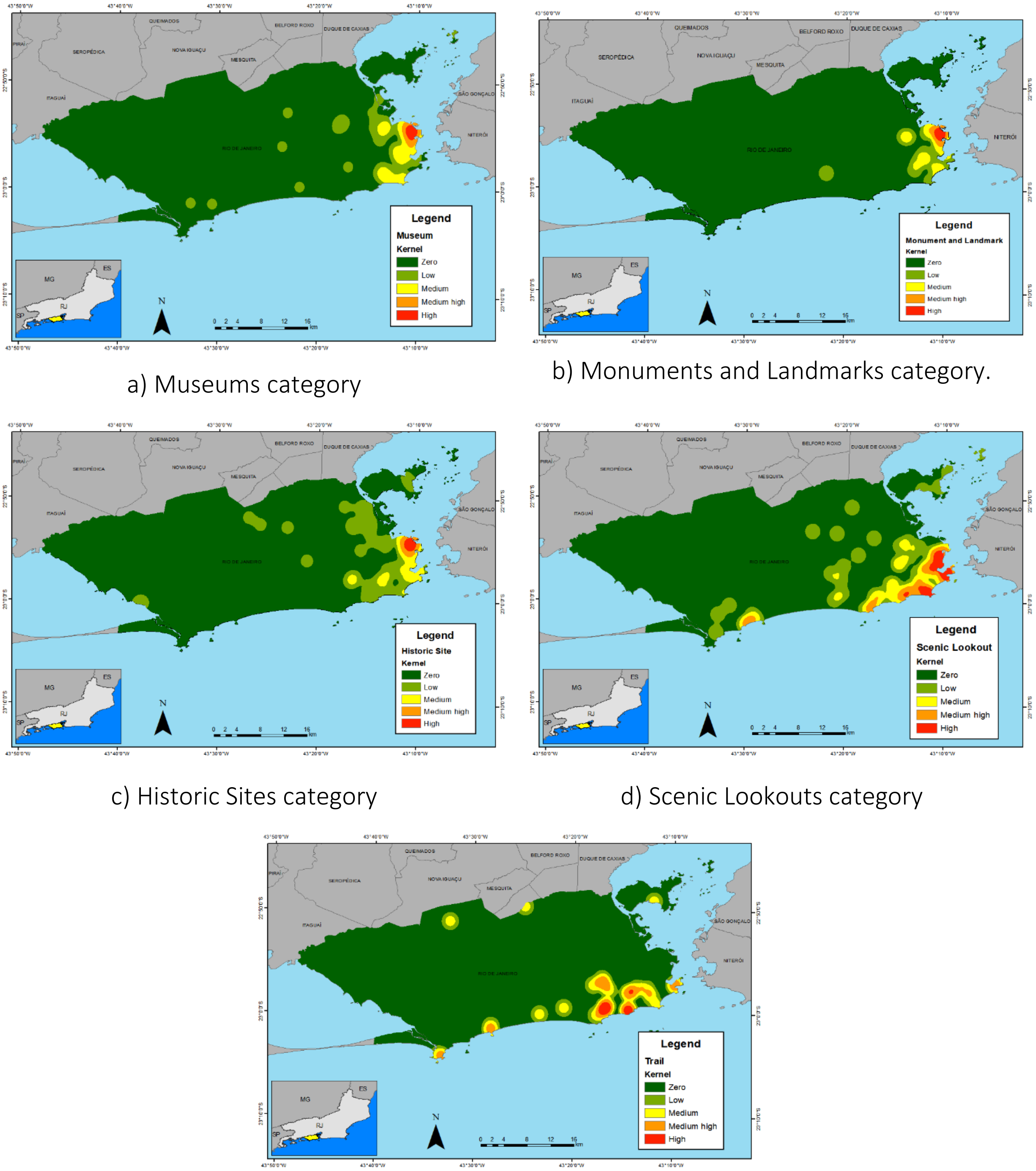
Order	Name	Category	Rating	Check-ins
1	Pedra do Arpoador	Scenic lookout	9.70	25,534
2	Morro da Urca	Mountain	9.70	13,117
3	Morro do Pão de Açúcar	Mountain	9.60	35,854
4	Igreja e Mosteiro de São Bento	Church	9.60	2,949
5	Mirante do Leblon	Scenic lookout	9.50	10,812
6	Vista Chinesa	Scenic lookout	9.50	6,614
7	Cristo Redentor	Monuments and landmarks	9.40	41,181
8	Forte de Copacabana	Military base	9.40	22,355
9	Mureta da Urca	Scenic lookout	9.40	10,335
10	Parque das Ruínas	Historic Site	9.40	6,619

In the ordered list of the tourist places with the highest ratings from visitors, it is interesting to see — in accordance with the tie-breaking criterion of number of check-ins (Table 3) — that among the ten places shown, four are in the Scenic Lookouts category (Pedra do Arpoador, Mirante do Leblon, Vista Chinesa, and Mureta da Urca), and that Morro da Urca and Morro do Pão de Açúcar (Sugarloaf Mountain) also have the subcategory of Scenic Lookout. In other words, six out of the ten highest rated places are in the Scenic Lookouts category, which confirms that the city of Rio de Janeiro has much natural beauty as tourist highlights.

To analyze the intensity of venues at a point in the city of Rio de Janeiro a density map was created using the quadratic kernel function. This function allows view spatially a density pattern of venues converting the points into a continuous surface through a degree of smoothing given by bandwidth. In the case of venues, an optimal scanning radius of 10 m and a resolution of 1 m were used in order to minimize the Mean Square Error.

When analyzing the density estimation maps of venues in Figure 11, it can be seen that the Museums category (Figure11a) is dense in the city's south zone, which indicates a possible cultural investment in this region, represented by the Gávea Planetarium and Eva Klabin Foundation, as well as new museums such as the Museu do Meio Ambiente (Environment Museum) inaugurated in 2008 in the Jardim Botânico neighborhood, and the Centro Cultural Oi Futuro Flamengo (Oi Futuro Flamengo Cultural Centre) inaugurated in 2005 in the Flamengo neighborhood. The Monuments and Landmarks category has a high density in the Centro region of Rio de Janeiro (Figure 11b), represented by famous landmarks such as the bustling Praça XV de Novembro, Itamaraty Palace, and the Monumento Nacional aos Mortos da II Guerra Mundial (Monument to the Dead of World War II), among others. Similar to the density estimation maps for the "Museums" and "Monuments and Landmarks" categories, the density estimation map for the "Historic Sites" category (Figure 11c) shows a high density in the Centro neighborhood, with emphasis on tourist attraction such as the Igreja Matriz Nossa Senhora da Candelária (Nossa Senhora da Candelária Church), Academia Brasileira de Letras (Brazilian Academy of Letters), Escadaria de Selarón (Selaron's Steps), and Duque de Caxias Palace.

Unlike the categories previously presented (Museums, Monuments and Landmarks, and Historic Sites), the density estimation map for the Scenic Lookouts category (Figure 11d) indicates spots along the seashore and in the hilly areas of the city of Rio de Janeiro, when exiting the concentration of the Centro area of the city shown in the other categories under study. This category has a high concentration of spots in the south zone of Rio de Janeiro, where one can find Morro Pão de Açúcar (Sugarloaf Mountain), Mirante Dona Marta, Pedra do Arpoador, and Mureta da Urca. In this category, the medium-high density areas are located at the beginning of the Barra da Tijuca neighborhood and in Recreio dos Bandeirantes, with some scenic lookouts such as Mirante do Roncador, Mirante do Pontal, and Pedra da Macumba. When analyzing the density map of the Trails category (Figure 11e), high density can be seen in the south zone, particularly near the neighborhoods of: São Conrado, with the Pedra Bonita trail; Vidigal, with the Dois Irmãos; and Alto da Boa Vista, with the highlights being Cachoeira do Chuveiro and Cachoeira dos Primatas. The medium-high density of this category is concentrated in the neighborhoods of: Urca, with the Trilha Morro da Urca trail; Recreio dos Bandeirantes, with Trilha da Prainha and Pedra Pontal; and Grumari, with the Pedra do Telégrafo and Pedra da Tartaruga trails.



a) Museums category

b) Monuments and Landmarks category.

c) Historic Sites category

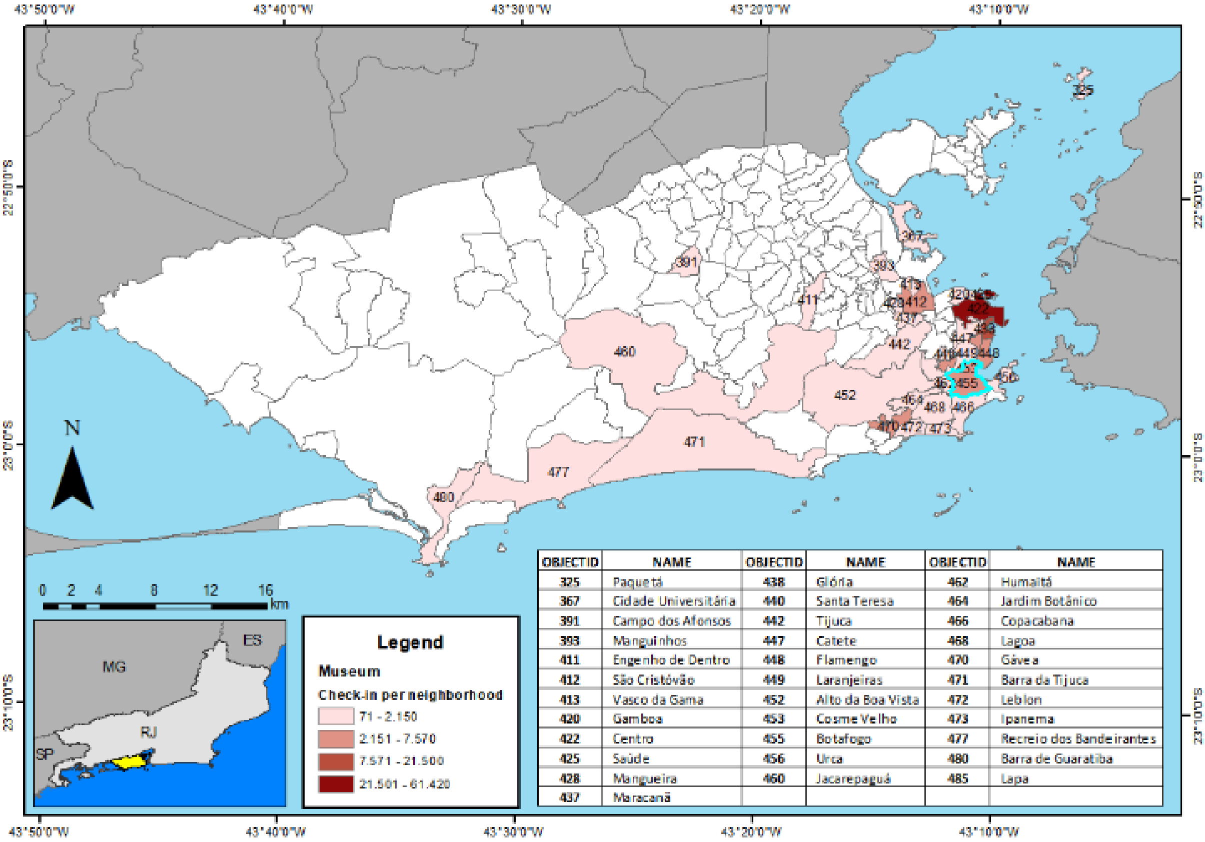
d) Scenic Lookouts category

e) Trails category

Figure 11: Estimation of venue densities per category

In Figure 12, the check-in map for the Museums category shows agreement with the density of points, in which the Centro neighborhood — which had the highest number of check-ins — is the one with the highest density of Museums. Gávea had a high number of visits — due solely to the Planetarium, which increases its importance and relevance for the neighborhood. The São Cristóvão neighborhood is also notable for the number of check-ins — the National

Museum and the Conde de Linhares Military Museum are located here. Despite neither neighborhood being in a high density location, they both have great cultural value.



OBJECTID	NAME	OBJECTID	NAME	OBJECTID	NAME
325	Paquetá	438	Glória	462	Humaitá
367	Cidade Universitária	440	Santa Teresa	464	Jardim Botânico
391	Campo dos Afonsos	442	Tijuca	466	Copacabana
393	Manguinhos	447	Catete	468	Lagoa
411	Engenho de Dentro	448	Flamengo	470	Gávea
412	São Cristóvão	449	Laranjeiras	471	Barra da Tijuca
413	Vasco da Gama	452	Alto da Boa Vista	472	Leblon
420	Gamboa	453	Cosme Velho	473	Ipanema
422	Centro	455	Botafogo	477	Recreio dos Bandeirantes
425	Saúde	456	Urca	480	Barra de Guaratiba
428	Mangueira	460	Jacarepaguá	485	Lapa
437	Maracanã				

Figure 12: Number of check-ins in the Museums category.

Analyzing the map for number of check-ins in the "Monuments and Landmarks" category (Figure 13), other relevant spots of significant tourist interest — but which appear outside the high-density zone — are revealed in this category. Among these, in the Santa Tereza neighborhood — one of the city's postcard locations — is Cristo Redentor (Christ the Redeemer), with about 40,000 check-ins, and in Copacabana there is the Carlos Drummon de Andrade statue and Princess Isabel monument.

Ao analisar o mapa por *check-in* (Figura13) é revelado outros pontos da categoria Monumento e Marco com relevância que aparecem fora da zona de alta densidade, mas que possuem grande interesse turístico. Dentre estes, no bairro de Santa Tereza, um dos cartões postais da cidade, o Cristo Redentor com cerca de 40 mil *check-ins* e em Copacabana tem-se a Estátua de Carlos Drummond de Andrade e Monumento à Princesa Isabel.

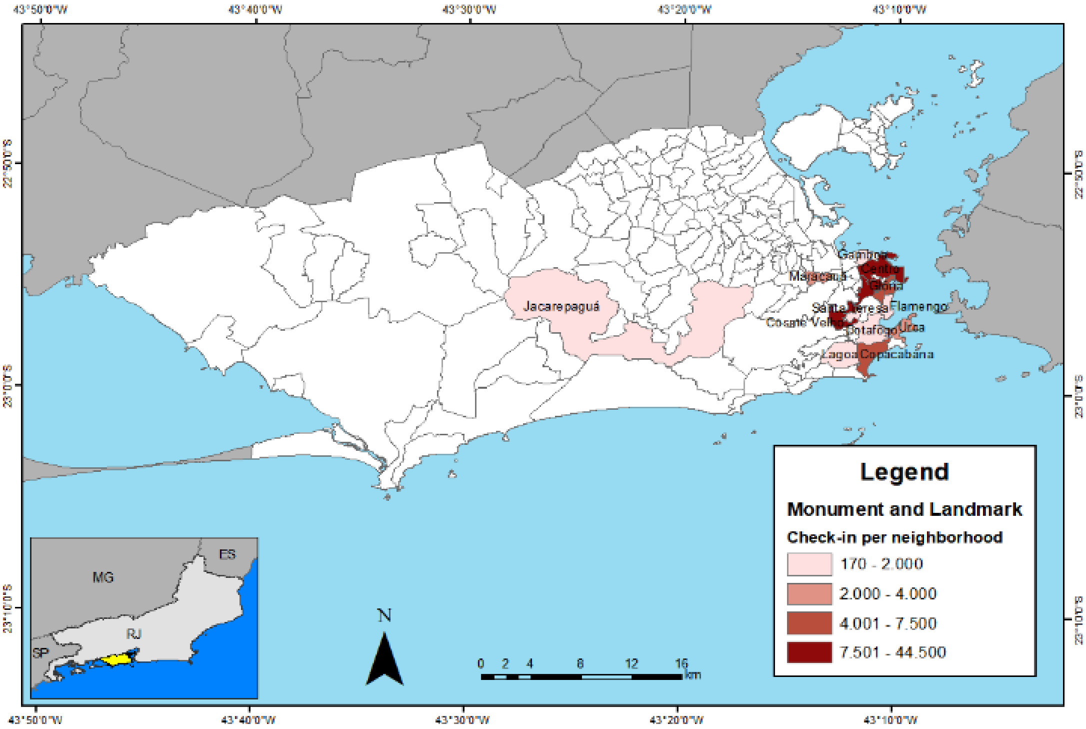
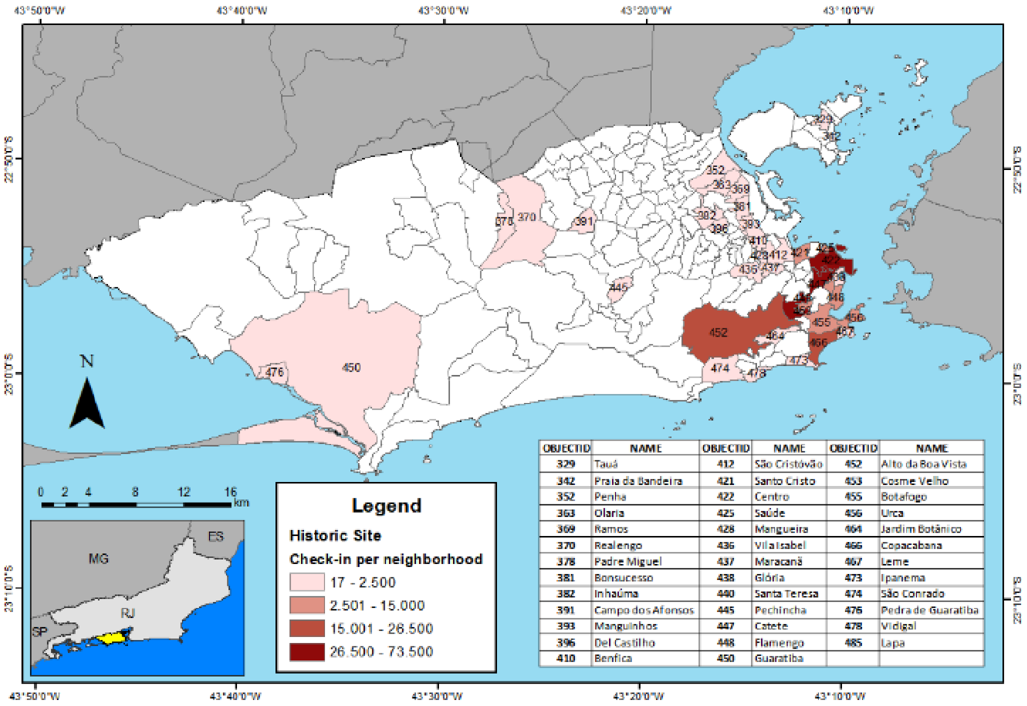


Figure 13: Number of check-ins in the Monuments and Landmarks category.

The most visited Historic Sites (Figure 14) are in the south zone, around the neighborhoods of Flamengo, Botafogo, Urca, and Leme, which have medium density. The Historic Sites category was the one that was most closely distributed among neighborhoods away from the seaside (e.g., Alto da Boa Vista, Cosme Velho, Guaratiba, and Realengo).



OBJECTID	NAME	OBJECTID	NAME	OBJECTID	NAME
329	Tauá	412	São Cristóvão	452	Alto da Boa Vista
342	Praia da Bandeira	421	Santo Cristo	453	Cosme Velho
352	Penha	422	Centro	455	Botafogo
363	Olaria	425	Saúde	456	Urca
369	Ramos	428	Mangueira	464	Jardim Botânico
370	Realengo	436	Vila Isabel	466	Copacabana
378	Padre Miguel	437	Maracanã	467	Leme
381	Bonsucesso	438	Glória	473	Ipanema
382	Inhaúma	440	Santa Teresa	474	São Conrado
391	Campo dos Afonsos	445	Pechincha	476	Pedra de Guaratiba
393	Manguinhos	447	Catete	478	Vidigal
396	Del Castilho	448	Flamengo	485	Lapa
410	Benfica	450	Guaratiba		

Figure 14: Number of check-ins in the Historic Sites category.

The most visited scenic lookouts (Figure 15) are in the neighborhoods of: Alto da Boa Vista, where the Mesa do Imperador and the Vista Chinesa stand out; Ipanema, with Pedra do Arpoador; Copacabana, with the Copacabana Fort; and Urca, with the highest number of check-ins in Sugarloaf Mountain, Morro da Urca, and Mureta da Urca.

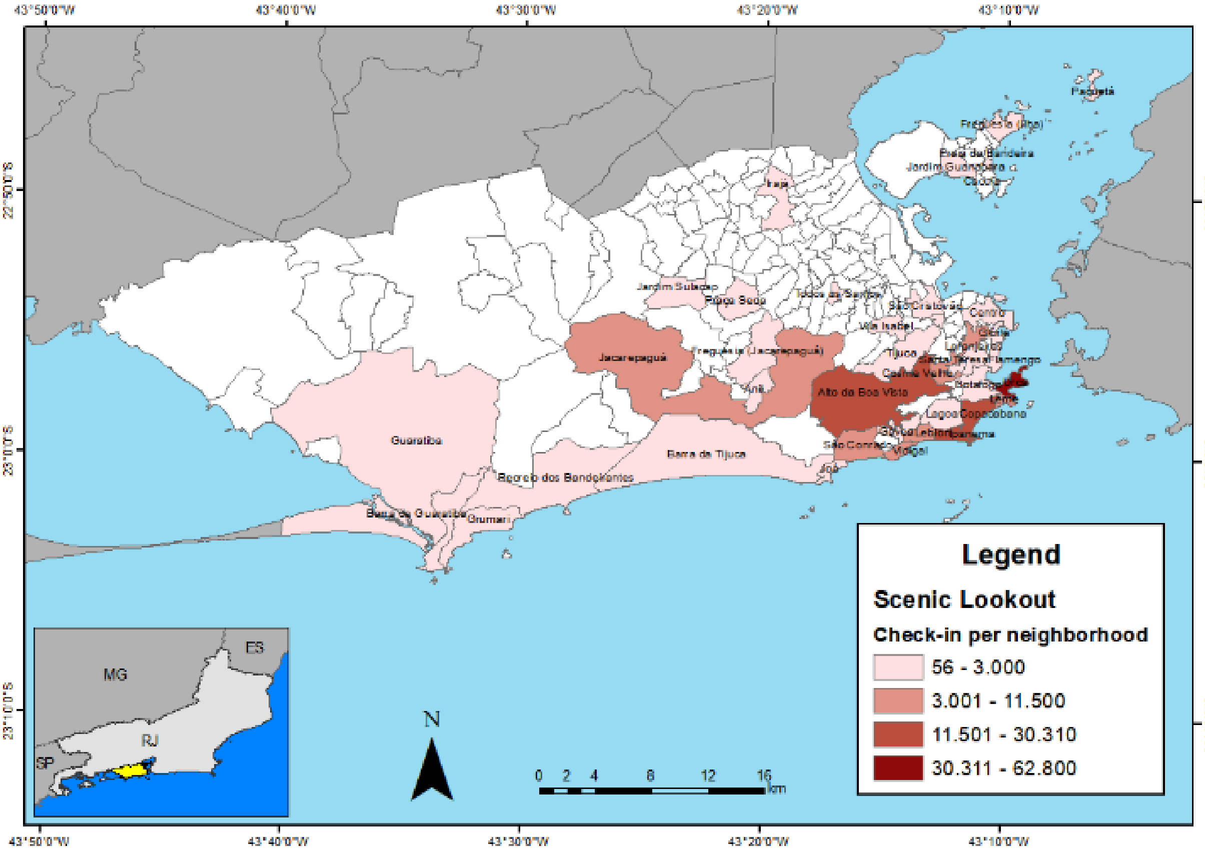


Figure 15: Number of check-ins for the Scenic Lookouts category

When analyzing the check-in map for the Trails category (Figure 16), the following neighborhoods stood out: Ipanema, with Pedra do Arpoador; Urca, with Sugarloaf Mountain; and São Conrado with Pedra da Gávea and Trilha da Pedra Bonita.

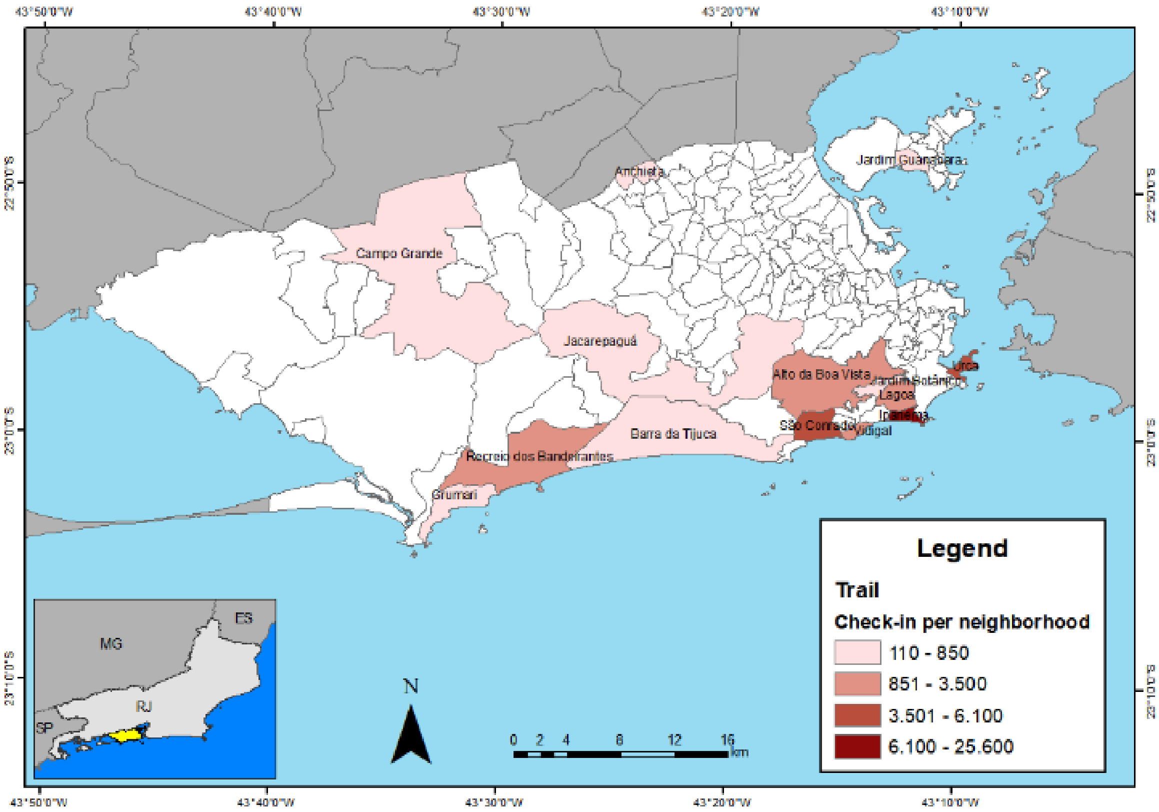


Figure 16: Number of check-ins for the Trails category.

In general, it can be concluded that for the Museums, Historic Sites, and Monuments and Landmarks categories, the high density is concentrated in the Centro area of Rio de Janeiro; however, there are places outside this high-density zone with great touristic value and a high number of visits. For example, the "Trails" and "Scenic Lookouts" categories are located and most visited predominantly in the south zone of the city, with some isolated parts in the west zone in the neighborhoods of Recreio dos Bandeirantes and Grumari. It is also worth mentioning — as shown in the graph of Figure 10 — that the places most sought for visitation are the historic sites of Rio de Janeiro, followed by the scenic lookouts. In the latter, it is worth emphasizing the benefit of the geographic locus of the city of Rio de Janeiro for tourism.

4. Final Considerations

Among the algorithms implemented to perform the search in Foursquare's API, the Grid algorithm was the least efficient. In fact, it is linked to the total number of results allowed per query. Since the points are not uniformly distributed, the uniform grid will be very small in some areas, where there are no points, or too large, when more than 50 results return.

In the development of this work, in the data mining of the Foursquare application, many difficulties were encountered that had to be overcome; for example, the constraint of only 50 venues per search area, the limits on number of searches per hour, and inconsistent behavior when the Quadtree and K-D tree search algorithms were applied based on the variation in decision making for spatial subdivisions. The lack of information regarding the evaluation of venues proved to be very problematic as well, due to its importance in quantitatively demonstrating the preferences of the population; hence the need to develop another system in order to perform searches separated by the web pages and to obtain the evaluations.

Nevertheless, for this analysis, as a tourism application, Foursquare proved to be a very versatile and coherent database with respect to presenting the values and attributes of places, performing the categorization of the data and facilitating the search for what is of interest to the user.

It is also worth noting the importance of performing a filtering of data in this project in order to them to be used, given that less than half of the information obtained was used via the selection process. However, filtering of data should occur according to the need of each project.

Thus, investment in areas that have a desirable flow of tourist visits — those quantitatively indicated as having more than 50 check-ins or that already had an evaluation via the Foursquare application — was considered for the tourism management. In other words, the relevance of the venue given via Foursquare is considered, even if it has not reached the limited visitation mark — stipulated in the filtering — in the application. This kind of analysis is important for mapping priority areas for investment in tourism.

For future work, an attempt will be made to optimize the search algorithm so that it operates in a hybrid manner regarding the search of venues; that is, the hybrid form would involve joining the grid spatial division — presented as the first search algorithm of this project — with the search of Quadtree spatial subdivisions controlled by decision making regarding the information returned and the depth of subdivisions to be performed.

This type of analysis offers a critical view of the municipality's tourism management regarding: the spatial location of these tourism categories, the tourists' evaluations of the places, and the frequency of the target public. In general, it can be concluded that the data behaved coherently regarding the reality of Rio de Janeiro, thus making Foursquare a highly recommendable application for conducting studies and making decisions for tourism management.

This work showed how SMGI can be used for spatial analyses together with traditional databases. Unlike authoritative geographic information used in planning, user-generated content provides data that enriches the knowledge of the area through information that is not normally available from official bodies, thus supporting a more comprehensive view of the system in relation to questions of geographic, social, and cultural analysis. In general, SMGI applications process data in real time, which can lead to an analysis of predictive models and identification of trends and phenomena that are affecting the study area.

REFERENCES

BARTOLOMÉ, A. Web 2.0 and new learning paradigms. ELearning papers, v. 8, p. 1-10, 2008.

CAMPAGNA, M., FLORIS, R., MASSA, P., GIRSHEVA, A., IVANOV, K. The role of social media geographic information (SMGI) in spatial planning. In: Planning support systems and smart cities. Springer International Publishing, 2015. p. 41-60.

CAMPAGNA, M., MASSA, P., FLORIS, R. The role of social media geographic information (SMGI) in geodesign. Jornal Digital Landscape Architecture, 2016.

CRCSI - AUSTRALIA AND NEW ZEALAND COOPERATIVE RESEARCH CENTRE FOR SPATIAL INFORMATION. Towards a Spatial Knowledge Infrastructure. White Paper March, 2017

GARTNER, G., BENNETT, D. A., MORITA, T. Towards ubiquitous cartography. Cartography and Geographic Information Science, v. 34, n. 4, p. 247-257, 2007.

GARTNER, G.; HUANG, H. Recent research developments in modern cartography in Europe, International Journal of Cartography, 2:1, 1-5, 2016.

GOODCHILD, M. F. Citizens as sensors: the world of volunteered geography. GeoJournal, v. 69, n. 4, p. 211-221, 2007.

MENEGUETTE, A. A. C. Cartografia no século 21: revisitando conceitos e definições. Revista Geografia e Pesquisa, Ourinhos, v.6, n.1, jan./jun. 2012.

PETERSON, M. P. Trends in Internet and Ubiquitous Cartography. Third International Joint Workshop of Ubiquitous, Pervasive, and Internet Mapping - UBIMap 2008. Shepherdstown, WV, Sept. 10-11, 2008.

SCHUURMAN, N. The new Brave NewWorld: geography,GIS,and the emergence of ubiquitous mapping and data. Environment and Planning D: Society and Space, volume 27, pages 571-580, 2009

SIMITSIS, A., VASSILIADIS, P. A Methodology for the Conceptual Modeling of ETL Processes. In: CAiSE workshops. 2003.

USERY, E. Lynn; VARANKA, Dalia. The Evolution of Cartography in the Digital Age: From Digitizing Vertices to Intelligent Maps. In: Conference proceedings. AutoCarto/UCGIS 2018 and The 22nd International Reserach Symposium on Computer´-based cartography and GIScience - Frontiers of GeospatialData Science, Madison, Wisconsin, USA, May 22-24, 2018