International Journal of Geographical Information Science
Vol. 00, No. 00, Month 00, 1–21

RESEARCH ARTICLE

Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain).

Taras Agryzkov[a], Pablo Martí[b], Leandro Tortosa[a] and José F. Vicent[a] *
[a]*Depto. Ciencia de la Computacion e Inteligencia Artificial, University of Alicante, Campus de San Vicente, Ap. Correos 99, E–03080, Alicante, Spain*;
[b]*Departamento de Edificación y Urbanismo, University of Alicante, Spain*
(*Received 00 Month 200x; final version received 00 Month 200x*)

Among social networks, Foursquare is a useful reference for identifying recommendations about local stores, restaurants, malls, or other activities in the city. In this paper, we consider the question of whether there is a relationship between the data provided by Foursquare regarding users' tastes and preferences and fieldwork carried out in cities, especially those connected with business and leisure. Murcia was chosen for case study for two reasons: its particular characteristics and the prior knowledge resulting from the fieldwork. Since users of this network establish, what may be called, a ranking of places through their recommendations, we can plot these data with the objective of displaying the characteristics and peculiarities of the network in this city. Fieldwork from the city itself gives us a set of facilities and services observed in the city, which is a physical reality. An analysis of these data using a model based on a network centrality algorithm establishes a classification or ranking of the nodes which form the urban network. We compare the data extracted from the social network with the data collected from the fieldwork, in order to establish the appropriateness in terms of understanding the activity that takes place in this city. Moreover, this comparison allows us to draw conclusions about the degree of similarity between the preferences of Foursquare users and what was obtained through the fieldwork in the city.

Keywords: Urban analysis; social networks analysis; street networks; PageRank algorithms; data visualization.

1. Introduction

The appearance of new technologies offering information about cities and urban spaces (Lazer et al. 2009, Offenhuber et al. 2014) is continuously increasing. Among others, the data provided by sensors, public or private services, mobile devices and web services (Ciuccareli et al. 2014, Jeong et al. 2011), represent a new source of information that could be useful not only as data analysis but also in the decision-making.

*Corresponding author. Email: jvicent@ua.es

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

In general, social networks can offer different visions on diverse aspects of social, economic, and political urban live according to each network users and interests (Hong 2015). Considering the information provided by social media, several researches have used these data as a source of knowledge for urban studies, planning, and management (Jeong et al. 2009). However, very few of these experiences analyse data from Foursquare.

Noulas et al. 2013 characterizes different geographic areas in two Spanish cities, Madrid and Barcelona. The authors combine telecommunication activity and Foursquare check-ins for the purpose of identifying the prominent activities in each geographic area. In Silva et al. 2013, some Foursquare and Instagram dataset have been analysed to study user movement patterns and the activities of those who use these social network.

If we clear up the general question of whether or not we can use data from the social network Foursquare to represent the activity that takes place in a city, it is clear that the answer is that Foursquare data cannot represent reality, since we would always get a skewed and incomplete picture of veracity, due to the limitations of the social network against the global population of a city (Serrano-Estrada et al. 2014).

Regarding the role that data provided by social networks play in urban analysis (Bawa-Cavi 2011, Cerrone 2015), we point out that the information provided should be considered as partial and determined by users and providers of the service. In general, the information given by web services forms part of data owned by a private company. Thus, the opinions obtained from these web services could be filtered in a certain way that cannot be verified. Moreover, the users of social networks represent a small part of the population with special characteristics that are later described (Jeong et al. 2009).

The main objective of this paper is to check if the data generated by Foursquare users can represent or be in agreement with the commercial and social activity of the city. In particular, we want to obtain conclusions about whether the data extracted from Foursquare could be considered as a reflection of the success of the public spaces of the city. To analyse and answer these questions we organize the paper as follows: In section 2, we present some preliminaries features that serve as a starting point of the performed work. At the beginning we describe some characteristics about Foursquare users demographics. After that we introduce the basic idea of urban networks, and present the urban layout of Murcia , the city object in his stdy. Then, we briefly describe the fieldwork develop in the city, as well as the dataset extracted from Foursquare Web service. Finally the algorithm used to analyse the different dataset is presented. Section 3 shows quantitative and ranking comparisons. These comparisons allows us to determine how far the reality of the city, in terms of commercial activity, is related with the users preferences of the social network Foursquare. Finally, some conclusions are presented.

2. Preliminaries

This section contains some preliminary work, introducing the basic characteristics of the two data sources: Fousquare data and fieldwork. An algorithm model is introduced to analyse and visualize the city, from the obtained data.

2.1. *Foursquare user demographics*

The user demographics of Foursquare are found to be quite different from other social networks whose purpose is merely social interaction (Lindqvist et al. 2011). Foursquare user profile consists of young professionals with ages ranging from 25

to 35 followed by user ages ranging up to 54 years old (Cranshaw et al. 2012). See http://www.ignitesocialmedia.com/social-media-stats/2012-social-network-analysis-report/♯Foursquare for a more detailed report about the users age of Foursquare users.

Foursquare is a popular Location-based Social Networks (Noulas et al. 2011), because it mainly consists of user-generated data about georeferenced venues for business and points of interest. Each Foursquare venue has a number of associated visits and check-ins. With the main difference between that the visits are the cumulative number of people that have checked in at least once at a venue, while the number of check-ins represent the cumulative number of times that a venue has received a check-in. Foursquare dataset used in this work were obtained at the end of 2013 and constitute a historical data record. This means that the registers have been counted since the venue was discharged until the data were obtained in 2013.

We study whether the data offered by the network (in terms of user preferences) match, in a significant manner, with the activity that takes place in the city itself. Although there are more than 55 million users of Foursquare, the number of people using this social network is a small percentage of the total population of an urban core, (less than 5%). The number of check-ins counted in Spain in 2013 amounts 5393, of which 5% are made in the city of Murcia. The social media users establish a ranking of places in the city itself according to their tastes and preferences (based on visits and check-ins). We analyse and compare these data with the commercial and leisure activities that take place in the streets of the city.

2.2. *Characteristics of urban networks*

The natural complex systems frequently take form of networks where nodes and edges are embedded in space (Barthelemy 2011). In fact, the city is a complex system where a huge amount of data and activity is generated. Urban networks (Porta et al. 2006) constitute a mathematical tool that can be used to describe the topology of the city or to analyse city data. The most common approach to represent spatial urban networks is using its classical representation through the concept of primal graph (Crucitti et al. 2006). Which in turn results in to deal with a particular type of a graph, where the edges are defined by the streets and the nodes become the edge intersections. The primal graph is a powerful data structure which encapsulate the topological information and weights that we can associate either to its nodes or edges. This flexible set of data structure allows us to carry out different types of spatial analysis.

One of the fundamental characteristics of network analysis is to determine the relative importance of a particular node within a network. The idea of relative importance is related with the mathematical concept of centrality (Friedkin 1991). Over the years, network researchers have introduced a large number of centrality indices which measures the varying importance of the nodes in a network. It is worth noting such centralities as degree, closeness (Freeman 1977), betweenness (Brandes 2001, Newman 2003), eigenvector (Bonacich 1987), straightness (Shon et al. 2010) and Katz (Katz 1953) centrality that provide us a basic description of the network components.

The study of centrality measures is the most common analysis applied to urban networks because it provides the information which indicates the structural properties of the urban layout. Centrality measures are also relevant for various spatial factors affecting human life and behaviours in cities (Crucitti et al. 2006). For this reason, we use a special centrality measure as a reference for comparison with the information obtained from the analysis of the social network Foursquare.

Figure 1. The area of the city of Murcia studied in this research and the primal graph.

2.3. *The city of Murcia and its urban layout*

To study an urban layout, we first need to represent it by an abstract model. To represent the abstract model we use a primal graph (Crucitti et al. 2006, Porta et al. 2006), where the streets become undirected edges and nodes usually represent the intersections of the streets. Besides, we can assign nodes to some points of interest in long streets or places of high human activity. We can see the final graph of the studied area of the city of Murcia, in Figure 1 (right).

The city of Murcia is located in the south-eastern area of Spain and it is the seventh largest city in the country. In this paper we work with the historical center of the city (Figure 1, left) and its surrounding neighbourhoods. This limitation is motivated, on the one hand, because of the need of reducing the amount of data to work with, and, on the other hand, because the historical center is the most active area of the city. This selected area of study occupies 40 hectares and it is characterized by a dense concentration of commercial venues and facilities.

The next task is to identify strategic points (nodes of the graph). In this context, each identified node represents a street intersection or a place of interest in the urban space. With this in mind, 1196 nodes are identified, using a numerical identification for every node. The number of edges in the primal graph is 1869, (see Figure 1, right).

2.4. *Fieldwork data in the city of Murcia*

In this section we briefly describe the data source representing the information from a physical context of the city. The fieldwork consists of a visual inspection of services and facilities related to tertiary activity located in the area of the case study. All services located in the firs two floors of the buildings which are exposed to the public view from the street were registered. Figure 2 (left) shows the geolocated data obtained by this fieldwork. We have established three categories to organize the 2801 registered services: Shops (small stores and minor trading), Malls and department stores (big stores,

malls, department stores, and supermarkets), Food services and entertainment (food, bars, pubs, and establishment). The number of venues collected by the visual inspection in the fieldwork are:

- Shops: 2216.
- Malls and Department Stores: 33.
- Food Services and Entertainment: 552.

2.5. *Foursquare data in the city of Murcia*

Foursquare geolocated data from a specific area were obtained using an API (Application Program Interface), that retrieves the data and stores them in a database. Foursquare responses returned a collection of places, each containing fourteen elements: name, check-in count, users count , photos count, likes count, latitude, longitude, sets, twitter, rating, categories, sub-categories, sub-sub-categories, description. See http://developer.foursquare.com/start for a more detailed report about the API used in Foursquare.

Foursquare is considered to be an useful source of information for several reasons. First, because the users check-in through a mobile device and they can only do while they are physically in the venue. Check-in is considered as primary source information. Secondly, Foursquare tracks the number of users visiting a venue, thus, it is possible to know which city places attract more people. Thirdly, this information further indicate the popularity of that specific place of the city (Roick et al. 2013).

The reason to choose the city of Murcia is based on the fact that it is the fourth Spanish city in terms of amount of activity on Foursquare according to www.puromarketing.com/16/15391/comousan-espanoles-foursquare.html). This is a reliable reflection of the importance of this city in the whole country (in terms of social media users).

Therefore, we analyse the data obtained from the web service. According to the data downloaded, Foursquare categorizes each venue into five predefined categories: Outdoors & Recreation, Shops and Services, Food, Nightlife y Arts & Entertainment. In turn, each category is divided into a number of subcategories. It is worthwhile pointing out that there is a lack organization of Foursqaures's subcategories because users set up the subcategories themselves. Therefore, in order to facilitate the research we have agree on a reorganization of subcategories as follows:

- Outdoors & Recreation: 168.
- Shops: 525.
- Food: 974.
- Nightlife: 204.
- Arts & Entertainment: 84.

In Figure 2 (right) we display Foursquare data and their geolocation in the city. It must be pointed out that these data correspond to the entire city of Murcia while only a portion (historical center) concerns this study. After removing the geodata located outside of the studied area, we have obtained a total number of 1216 venues that have received at least one visit by Foursquare users. The classification is:

- Outdoors & Recreation: 89.
- Shops: 349.
- Food: 596.

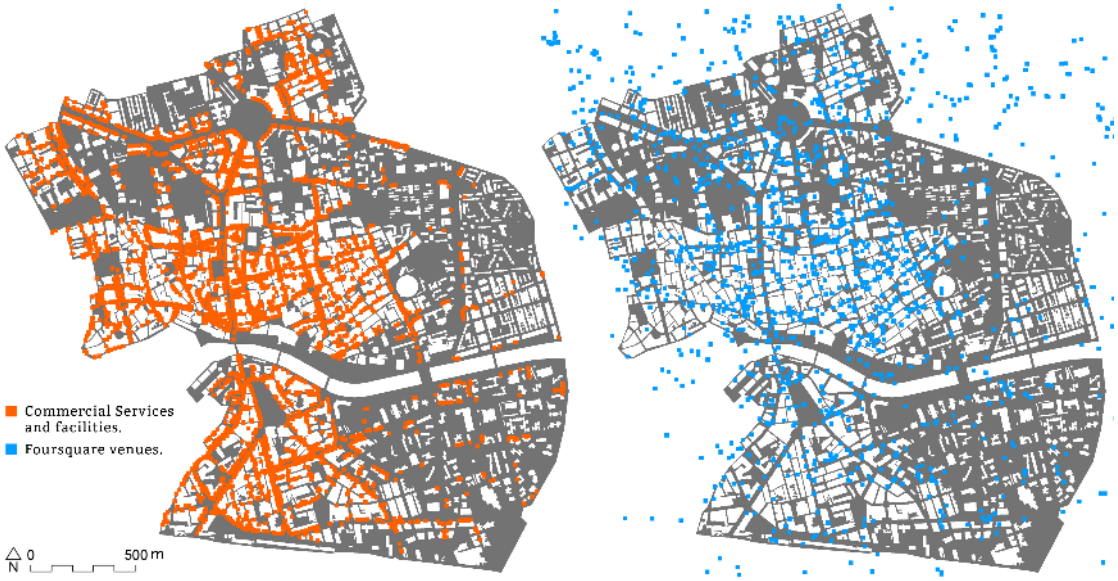

Figure 2. Geolocation of the fieldwork and Foursquare datasets.

- Night-life: 141.
- Arts & Entertainment: 41.

The 62% of the data obtained from the Web service Foursquare are located within our study area. This gives us an idea of the importance of the urban center in this city. Of the 1216 venues visited, a total of 1086 (the sum of shops, food and nightlife) correspond to the commercial sector of the city economy. This value represents 90% of the total data and it shows clearly some essential features of the city, as it is the predominance of commercial and leisure sectors.

2.6. *An algorithmic model to analyse the city*

The algorithmic model used is based on the Adapted PageRank Algorithm (APA) (Agryzkov et al. 2012) and it allows us to analyse and visualize a city, represented by a primal graph. The developed analysis is based on the information we have from the network, from both Foursquare and fieldwork data.

The fieldwork provides data (fieldwork done in 2013) which quantifies the physical services existing in the urban layout and, ultimately, can be used in the APA algorithm. This algorithm produces a classification of network nodes according to their importance or relevance within the network, taking into account the topology of the urban network and its data distribution.

The classification of the nodes allows us to determine the points or areas of greatest impact on the city, according to the information that we are evaluating. This algorithmic model uses the data collected by visual inspection, which means that it represents the reality of the establishments physically located in the city (at least, at the moment when the fieldwork was done). In this paper, the model has been used to represent the tertiary activity of the city of Murcia, taking into account the fieldwork performed (for a detailed description see Agryzkov et al. 2014).

Without going into detail, we briefly explore the features of the algorithmic model used. The APA algorithm is based on the concept of the PageRank vector that uses the

Google search engine to obtain a classification of all the Web pages (Page et al. 1999).

However, what motivated the modification of the original PageRank model is that it computes a ranking for all the Web pages based on the Webgraph. It only takes into account the connectivity of all the pages and their location in the Webgraph. In our model we use urban networks and therefore the PageRank is not appropriate. We need to consider some other information in addition to the location of the nodes in the graph. So, the original PageRank algorithm has been adapted to consider the influence of any dataset that provide information on the urban network.

Figure 3. A summarized scheme of the algorithmic model.

Figure 3 shows an scheme of the key points of the algorithmic model, in which we highlight three main elements to form a matrix, whose dominant eigenvector provides a ranking of the nodes according to their importance in the network.

The central point behind the APA algorithm for ranking the nodes is the construction of a data matrix D, which allows us to represent numerically the information of the network that we are going to analyse and measure.

The first of these elements is a stochastic matrix A, which is not exactly the adjacency matrix of the graph which represents the network, but is related to the degree of the nodes.

Let us assume that we have a set of n nodes $p_1, p_2, ..., p_n$ and we denote by c_i, for $i = 1, 2, ..., n$, the number of outgoing connections from the node p_i to other nodes.

This matrix is $A = (a_{ij})_{i,j=1}^{n} \in \mathrm{R}^{n \times n}$ where

$$a_{ij} = \begin{cases} \frac{1}{c_j} & \text{if there is a link from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The matrix A is built in such a way that its columns provide us with valuable information about the outgoing connections of the nodes. It is evident that, among other well-known properties, A is non-negative, stochastic by columns, irreducible and primitive.

To take into account the different importance of each node in the network, it is necessary to define a new matrix called data matrix, and denoted by D. The rows of this matrix D, of size $n \times k$, represent the n nodes of the urban network, and its k columns represent the characteristics associated with the nodes. These characteristics may be different, according to the problem or network studied. Specifically, an element $d_{ij} \in D$ is the value we attach to the property k_j at node p_i.

However, if we establish k characteristics associated to the nodes of the network, not all of them have the same relevance or influence in the solution of the problem. Therefore, we construct a vector $\vec{v_0}$, which constitutes a multiplicative factor for the different items in the data matrix D. For instance, if we are studying three sectors, Food, Shops and Malls, the vector $\vec{v_0}$ will have three components that assign a specific relevance to each

one. If we only want to study the first sector (Food service and entertainment), we take $\vec{v_0} = (1, 0, 0)$. Likewise, if we want to study the second sector (Shops), then $\vec{v_0} = (0, 1, 0)$ while if we want to emphasize only the third sector, (Malls and department stores), $\vec{v_0} = (0, 0, 1)$. If we highlight all equally important sectors, then $\vec{v_0} = (1, 1, 1)$.

For instance, if we consider in the data matrix D different types of commercial activity, perhaps each activity will have more or less importance according to the study we are performing. This way, the vector $\vec{v_0}$ allows us to previously determine the importance we want to assign to any of the characteristics measured in D. Therefore, we multiply $D \cdot \vec{v_0} = \vec{v}$ in order to assign to each node the importance we want to give to each type of activity or information represented in D. Note that $\vec{v} \in \mathrm{R}^{n\times 1}$.

After obtaining $\vec{v}$ we normalize this vector following the classical method of dividing each component by its modulus and denote it as $\vec{v^*}$. Then, we construct the matrix $V \in \mathrm{R}^{n\times n}$ as a matrix in which all of its components in the i-th row are equal to $\vec{v^*}$. In practice, we can say that we copy the vector $\vec{v^*}$ in every column of the matrix V (n times). After that, we are ready to define the matrix M' as

$$M' = (1 - \alpha)A + \alpha V. \tag{2}$$

The parameter α was originally introduced is the PageRank algorithm used by Google search engine to classify Web pages. After numerical experiments they established that $\alpha = 0.15$ (Bianchini et al. 2005, Page et al. 1999). We also use $\alpha = 0.15$, following PageRank algorithm suggestions. Once M' is constructed, the eigenvector associated to the eigenvalue 1 for the matrix M' is our ranking vector, and it provides us with the ranking values for the nodes of the network.

For a detailed description of the model and the mathematical background in which is based, see Agryzkov et al. 2012, Berkin 2005, Page et al. 1999.

3. Measuring some urban activities from Foursquare data

The rationale for conducting comparative analysis between Foursquare data and field-work data, in the city of Murcia, is twofold. On the one hand, we face two ratings of public space. The data that social network users generate with their activity suggest a ranking of sites or venues within the city itself. This means that by counting the numbers of visits to venues it is possible to establish a rating following the tastes and preferences of the users of the social network. Moreover, we have an algorithmic model that allows us to classify the nodes of the primal graph of the city taking into account their connectivity and the data present in their surroundings. Therefore, the results of the algorithmic model also constitute a ranking of geographic locations. Since we have two rankings obtained from different sources, but in the same geographical space, it is important to carry out a comparison of both classifications to ascertain any differences or similarities between both data sources.

Moreover, the study produce an accurate quantification of the data used. Since the Foursquare data in the city of Murcia are related to the users tastes, we can conduct a comparative graphical analysis of the dataset obtained from the social network with field-work data (which represent the physical reality of the city with regards to the commercial sector).

Consequently, in this section we perform two comparisons.

- **Quantitative comparison**. Foursquare dataset in the city of Murcia was compared

with the fieldwork data obtained in the city.

- **Rankings comparison**. The ranking of the venues established by Foursquare users was compared with the ranking of nodes given by the algorithmic model.

3.1. *Quantitative comparison*

Initially, we perform a quantitative comparison between the geolocated Foursquare data and data collected from the streets' network by visual inspection of tertiary activity in the city (fieldwork).

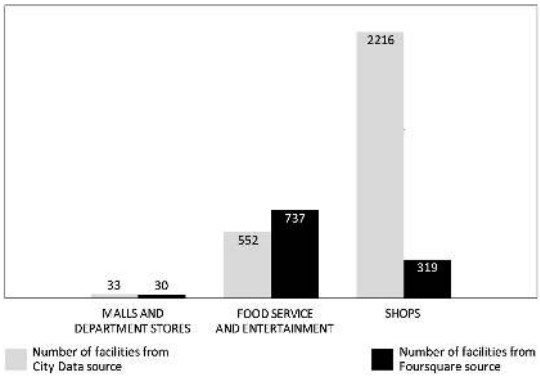

Figure 4. Quantitative comparison of fieldwork and Foursquare datasets.

In the fieldwork 2801 venues were registered relating to Shops, Malls and department stores, and Food service and entertainment. Based on the Foursquare data, 1086 places have been visited at least once by users of this social network. By computing all three categories we can appreciate a variation of 1715 facilities. The bar chart of Figure 4 illustrates the statistical information of both data sources: Foursquare and fieldwork. Malls and department stores sector presents a nearly identical record, with an average deviation of 1%. In the case of the Shops sector, the results show a remarkable divergence of 86%. In the fieldwork, a total of 2216 facilities were collected, but only 319 of these venues record at least one visit from Foursquare users. In the case of Food service and entertainment sector, the results show a variation of 34%, significantly lower than the Shops sector. Here, a total of 737 venues received at least a visit.

We observe a 40% average value of deviation in the amounts recorded between Foursquare and fieldwork dataset. These discrepancies are completely logical and expected, considering the characteristics of both data sources. Visual inspection shows all existing trade in the city, regardless of other factors. However, users of Foursquare set their preferences on venues and places in different parts of the city. Users with their visits are, somehow, making recommendations to other users of what they consider interesting, either in terms of appearance or environment. It was noted that many of the small businesses in the city do not generate a great deal of interest from Foursquare users. Therefore, the largest deviation between the data appears when we compare Shops sector of both data sources. However, when we analyse the data related to the Food service and entertainment sector, the discrepancy is negligible. Foursquare users are prone to evaluate in a positive way coffee places, bars, restaurants, etc.

It is clear from the datasets presented, that there are more places related to the Food service and entertainment sector registered by Foursquare web service, than those col-

lected in the fieldwork. Taking into account that there are facilities related to the Food service sector that are not deleted from the social network data base when a business is sold or those is a change of ownership. Additionally, in the city of Murcia, as in most large cities, there has been a recent proliferation of businesses related to the food and recreation services.

The analysis of quantitative comparison between these two data sources may have its limitations, since it is clear that it present a partial reality based on the sample size that Foursqaure offers. However, we consider these data to be worthy of analysis because this contrast in data volumes in different categories or sectors provides us some insights into the characteristics and peculiarities of the social network users in this city.

With this comparative analysis, we can conclude that Foursquare in the city of Murcia is mainly focused on the Food sector and entertainment venues. In addition, it follows that the user prototype of the network in this city likes to enjoy public places in the city, with friends, especially those related to the food activity, cafes, bars and pubs, which makes this city a lively place and with an intense life in their public places.

So, we can say that in the case of the city of Murcia, Foursquare virtual databases contain enough information to be compared to a fieldwork in the case of Food service and entertainment sector, at least in a quantitative sense.

This quantitative comparison also serves as a starting point for the subsequent comparative study, which we discuss in the next section.

3.2. *Rankings comparison*

In this section, a comparison of *popularities* is performed. We compare two ratings arising from the two data sources discussed in Section 3.1. On the one hand, we have a classification of the nodes of the urban network given by the mathematical model presented in Section 2.6. On the other hand, a popularity ranking offered by Foursquare, through the user visit records. From these data we can estimate that places with the higher number of visits are the most popular.

It is interesting to compare both rankings to establish possible coincidences, as well as to determine whether there is a correlation between the results provided by the algorithmic model and the ranking of popularity provided by the social network. Note that the ranking of nodes provided by the APA algorithm takes into account information collected objectively by visual inspection, as well as the network topology itself.

To perform the comparison, we follow the scheme summarized in Figure 5. On the left branch we have the data physically extracted from the city (visual inspection), while on the right one we have the data from the social network (Foursquare).

Regarding the physical data source, the work has involved the collection and registration of the facilities studied (Shops, Malls and department stores and Food service and entertainment). Once this information has been registered, it is associated to the nodes of the graph. Then, the algorithmic model is applied and a ranking of the nodes is obtained. This classification is the *algorithmic ranking* in the diagram in Figure 5 and is denoted in Table 1 by r_1.

Concerning process related to Foursquare data source, we first obtained the geolocated data of the social network. Then, this information was registered and allocated to the nodes of the graph. Subsequently, a ranking of visits per node was obtained. This classification is the *Foursquare ranking* in the scheme in Figure 5 and is denoted by r_2 (see Table 1).

In both analysis performed in this work (Fieldwork and Foursquare), the same pro-

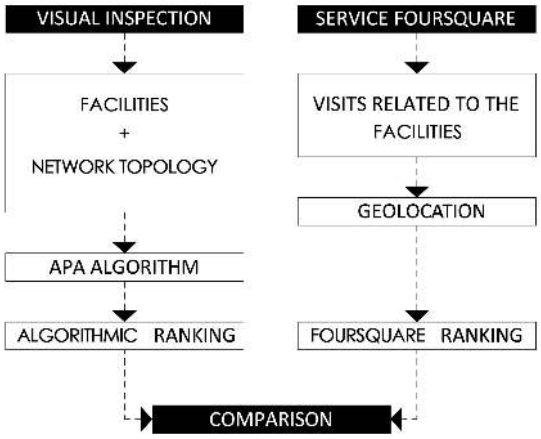

Figure 5. Workflow summary in the popularity comparison of Foursquare service and fieldwork.

cedure to assign data to the nodes of the network has been used. The data we need to assign are located in a block of buildings, so we have taken the criterion that these services influence into the vertices of the polygon that covers the block of buildings. Obviously, there are other criteria such as proximity or assign to the edge, but we have opted for one that allows us to automate the allocation process efficiently. Besides, this criterion does not imply that it is better or worse, it's just another one.

In order to clarify this process we show an example that describes the steps involved in the process. Selecting a small piece of the urban network of Murcia (Figure 6), along with those of the geolocated venues. In the example of figure 6, we have three venues or facilities e_1, e_2, e_3 and several nodes. The first step of the allocation process is the transformation of the graph into a set of adjacent polygons, where its nodes define the vertices of the polygons. If we consider a particular venue, we must determine the polygon, in the graph, in which this venue is located. Then, this venue is associated with the vertices of the polygon which contains the venue. This process must be performed for all the commercial venues studied. In the case of Foursquare, the nodes store the sum of the visits that each of the venues assigned to that node has received.

Consequently, the task of mapping the information of the nodes of the network has two parts: the location of the polygons that contains the commercial facility and the allocation to its vertices (nodes of the graph). After performing the allocation process, we have a quantification of the commercial activity of the city because each node of the urban network has a number of commercial services that allows us to establish classifications of the nodes (*algorithmic model ranking* and *Foursquare ranking*).

We briefly expose the model used to perform the visualization of the rankings.

With the two rankings analysed, we obtain vectors representing a classification of network nodes ($\vec{x} = (x_1, x_2, x_3, ..., x_n)$, with $min \leq x_i \leq \max$). This means that the value $x_i \in \vec{x}$ represents the ranking of the node v_i in the network. Since this vector does not provide any visual information, we need to transform the values of the vector $\vec{x}$, in a chromatic scale that easily allows us to visualize the importance of each node within the network. For this purpose, the *Hot-Cold Scale*, based on the RGB model (Red, Green, Blue) is used. Consequently, we have two distinct scales: the scale of values $x_i \in \vec{x}$ and, the Hot-Cold Scale. The first step is the transformation of a particular numerical value x_i into the corresponding value of the color scale; as shown in Figure 7 (b), each node has an associated color. The second step for the final display consists of performing a

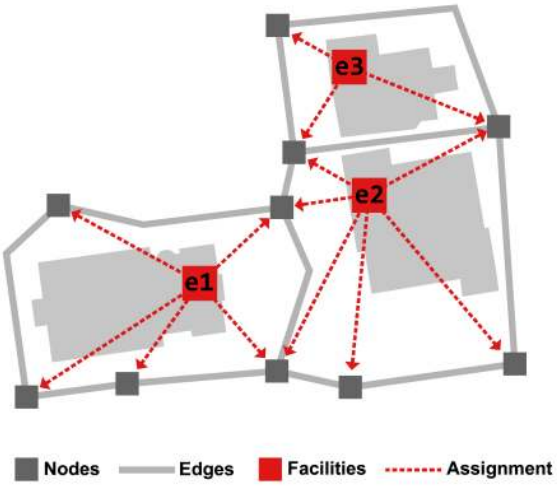

Figure 6. Assigning data to the nodes of the network.

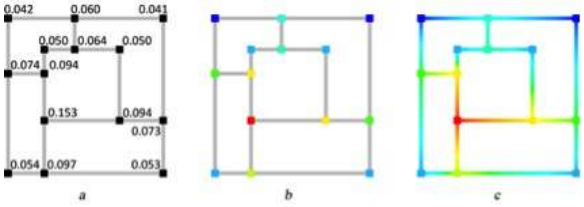

Figure 7. Transformation of numerical values to chromatic scale.

standard linear interpolation with the aim to obtain the graphical representation of the edges, see Figure 7 (c).

It is worth pointing out that the three categories established in this work besides being differentiated by their particular commercial characteristic are also differentiated by the amount of people the place is able to accommodate within. For example, it is obvious that venues of the category Malls and Department Stores always receive more visits than services of Shops category due to its larger size. In order to avoid the distortion produced in the scales of the different categories, each analysis is applied for each category separately.

Likewise, to homogenize the results of each category, we proceed to scale the dataset.

In Figures 8-10 we display on the left the ranking of nodes using the algorithm model and, on the right, Foursquare ranking. The red tones represent, in the case of the theoretical model, the highest values of the ranking and, in the case of Foursquare, the largest number of visits.

Figure 8 compares the rankings of the Shops sector, where a significant difference between the results obtained from the two methods of analysis is observed. This difference is justified by the absence of records of many establishments related to this commercial sector in databases of Foursquare. We must also take into account two factors: firstly, that only a small part of the population of Murcia uses Foursquare and secondly related to the type of facilities or shops located in the studied area, where traditional shops are a majority. Usually, these shops are not of interest to Foursquare users. However, it is noteworthy that the most valued venues by the users of this social media match with those values that achieve greater importance according to the algorithmic model.

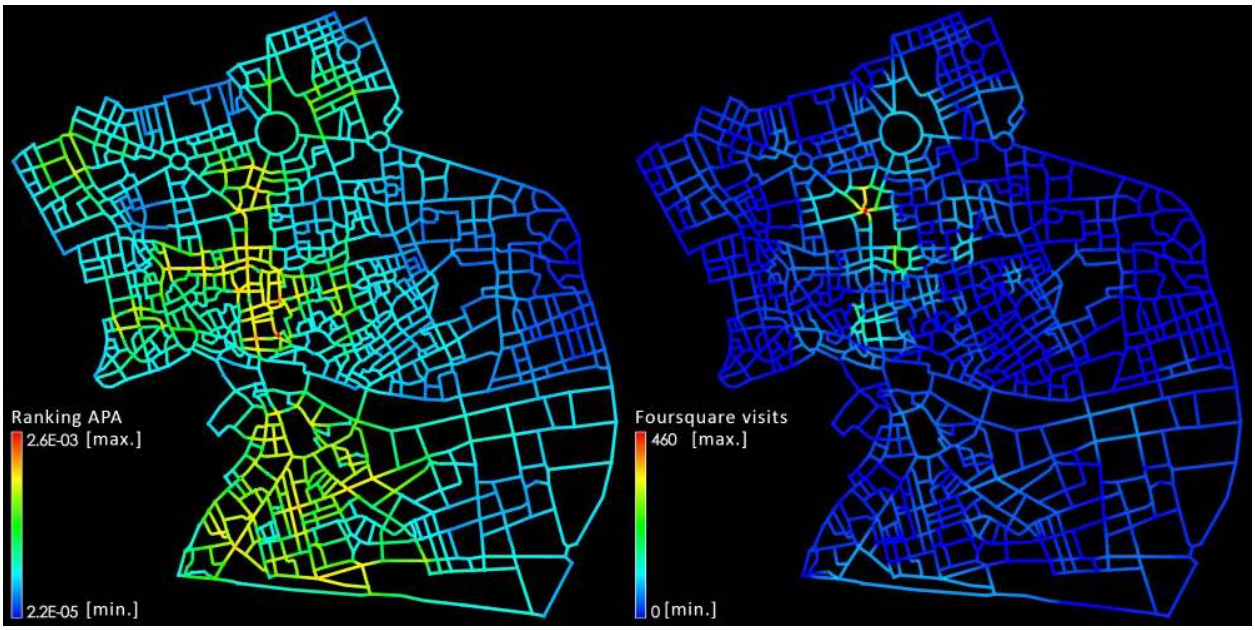

Figure 8. Shops sector. Rankings comparison between the algorithmic model (left image) and Foursquare (right image).

Figure 9. Food service and entertainment sector. Rankings comparison between the algorithmic model (left image) and Foursquare (right image).

Figure 9 provides a comparison of the rankings of the Food service and entertainment sector which stresses the similarity of places with high levels of popularity located in the historical part of the town. Moreover, both places with maximum recorded values in the ranking are located around a particular location (Santo Domingo square).

Figure 10 compares the rankings of the Malls and department stores sector, where similarity is evident in the distribution of the venues identified with high values of pop-

Figure 10. Malls and department stores. Rankings comparison between the algorithmic model (left image) and Foursquare (right image).

ularity. In both cases, a lack of high values can be observed in the area comprising the historic centre of the city.

Ideally, we would like to establish a comparison that allows us to check whether the venues that occupy the top positions in the number of visits by Foursquare users and those which are located in high positions of the algorithmic ranking are the same, or if there exist significant differences. This question constitutes a comparison of the two rankings.

The ranking provided by the algorithmic model does not consider the preferences of people and other subjective factors, since it only considers the influences from the network topology and the information contained in the data matrix. However, the data provided by the Foursquare network are associated to the tastes and preferences of users.

We want to analyse similarities and differences between both rankings for each node. Due to the characteristics of the two data source, we focus on the nodes that are at the top of the popularity ranking of Foursquare in order to determine whether these nodes are also relevant in the algorithmic model ranking.

The procedure is as follows: we obtained for each nodes, which occupied a higher value in the Foursquare ranking, its position in the algorithmic ranking. Then, we subtracted their positions in both rankings (in absolute value). This difference is denoted as dif. Finally, we establish a percentage deviation using this difference. A concrete example helps to clarify this process. We focus on the Food service and entertainment sector of activity because it is the sector where we have a more balanced quantity of data. The node that occupies the first position in the Foursquare ranking is identified with the number 857 in the graph. This node holds the position 51 in the algorithmic ranking. The difference between the two positions in the respective rankings, denoted by *dif*, is $dif = |1 - 51| = 50$.

As we have 1196 nodes classified in a vector representing the urban streets of the city

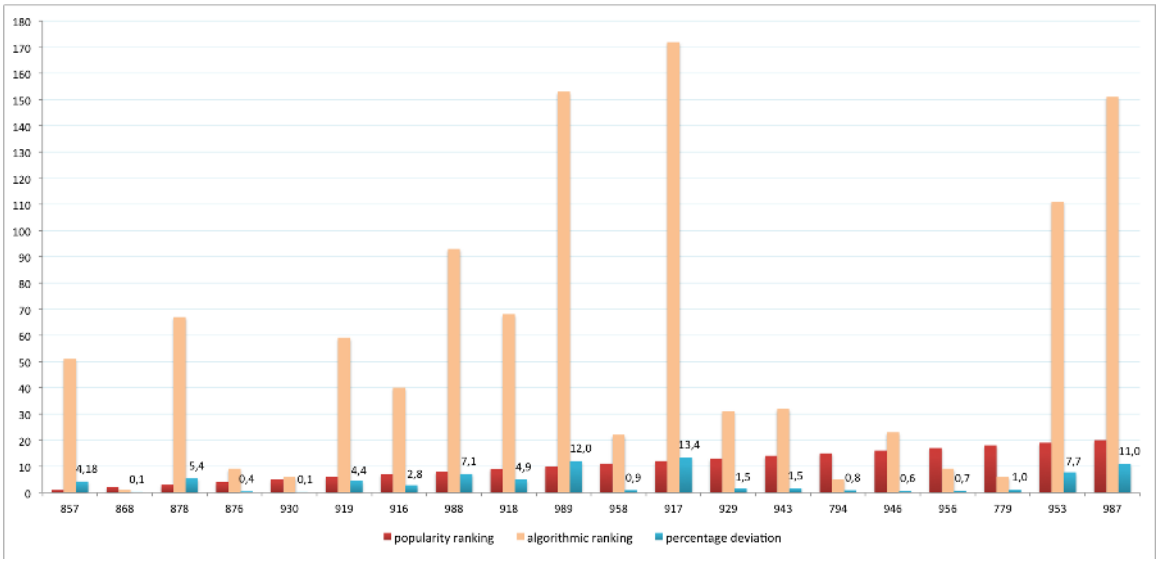

Figure 11. Percentage deviation of the 20 first nodes between algorithmic and Foursquare ranking related to the Food service sector.

of Murcia, we can compute the percentage deviation as

$$\frac{50}{1196} \cdot 100 = 4.18\%.$$

The node in the second position in the Foursquare ranking (node 868) is placed in the first position in the algorithmic ranking. So the difference is only one position, and the percentage deviation is 0.08%.

In Figure 11 we have selected the top 20 ranking nodes that we obtained from the Foursquare ranking. In the abscissa axis the identifier of the node is shown. The vertical axis represents the ranking position of the node within the network. Each node has three bars, the first one indicates the position in the ranking given by Foursquare (note that it grows in a unit since we have taken the 20 nodes with a higher popularity value). The second bar indicates its position in the algorithmic ranking and third one shows the percentage deviation between both values. The percentage value of deviation appears explicitly in the corresponding bar.

Analysing in detail Figure 11, we realize that the similarities in the values of the different rankings studied for the first 20 nodes with higher number of visits, are remarkable. From the set of these first 20 nodes represented in the graph, eight of them have a percentage deviation of 1% or less. Moreover, only three of them present a percentage deviation greater than 10%, with the maximum value reached of 13.4%.

We highlight, for instance, the node 868 of the graph (Figure 11). This node ranks second in the Foursquare ranking, whereas it is the node with higher centrality value when we apply the algorithmic model. The node 930 of the graph also shows a difference of only one place in both rankings. Therefore, we conclude that nodes which have preferred and recommended venues for Foursquare users, are relevant and have an important centrality within the urban network (in terms of Food services sector).

Table 1 shows the properties of the first 50 nodes in the ranking of visits provided by Foursquare. The first column represents the numerical identification of each node, the second column shows the position of the nodes in the ranking obtained by running

Food service entertainment					Shops					Malls and department stores				
Node Id.	r_1	r_2	dif	%	Node Id.	r_1	r_2	dif	%	Node Id.	r_1	r_2	dif	%
857	51	1	50	4.2	832	12	1	11	0.9	830	72	1	71	5.9
868	1	2	1	0.1	482	163	2	161	13.5	832	1	2	1	0.1
878	67	3	64	5.4	816	106	3	103	8.6	814	158	3	155	13.0
876	9	4	5	0.4	869	151	4	147	12.3	816	160	4	156	13.0
930	6	5	1	0.1	1169	100	5	95	7.9	817	91	5	86	7.2
919	59	6	53	4.4	487	85	6	79	6.6	833	110	6	104	8.7
916	40	7	33	2.8	865	377	7	370	30.9	834	159	7	152	12.7
988	93	8	85	7.1	870	23	8	15	1.3	835	78	8	70	5.9
918	68	9	59	4.9	488	7	9	2	0.2	836	7	9	2	0.2
989	153	10	143	12.0	779	453	10	443	37.0	865	170	10	160	13.4
958	22	11	11	0.9	793	771	11	760	63.5	1169	5	11	6	0.5
917	172	12	160	13.4	794	430	12	418	34.9	779	288	12	276	23.1
929	31	13	18	1.5	478	350	13	337	28.2	793	311	13	298	24.9
943	32	14	18	1.5	860	800	14	786	65.7	794	105	14	91	7.6
794	5	15	10	0.8	861	753	15	738	61.7	671	26	15	11	0.9
946	23	16	7	0.6	864	581	16	565	47.2	672	37	16	21	1.8
956	9	17	8	0.7	771	3	17	14	1.2	681	48	17	31	2.6
779	6	18	12	1.0	769	34	18	16	1.3	1055	23	18	5	0.4
953	111	19	92	7.7	480	248	19	229	19.1	1060	20	19	1	0.1
987	151	20	131	11.0	770	45	20	25	2.1	1065	33	20	13	1.1
984	146	21	125	10.5	829	541	21	520	43.5	1111	53	21	32	2.7
986	147	22	125	10.5	788	418	22	396	33.1	1113	22	22	0	0.0
877	176	23	153	12.8	789	396	23	373	31.2	1114	43	23	20	1.7
792	39	24	15	1.3	792	519	24	495	41.4	1138	50	24	26	2.2
951	116	25	91	7.6	756	328	25	303	25.3	676	521	25	496	41.5
793	30	26	4	0.3	778	247	26	221	18.5	677	520	26	494	41.3
949	20	27	7	0.6	767	9	27	18	1.5	660	267	27	240	20.1
955	16	28	12	1.0	477	524	28	496	41.5	668	501	28	473	39.5
789	33	29	4	0.3	339	655	29	626	52.3	669	603	29	574	48.0
875	14	30	16	1.3	830	204	30	174	14.5	670	279	30	249	20.8
756	37	31	6	0.5	764	10	31	21	1.8	681	48	31	17	1.4
879	21	32	11	0.9	338	639	32	607	50.8	691	642	32	610	51.0
881	2	33	31	2.6	814	43	33	10	0.8	698	707	33	674	56.4
855	57	34	23	1.9	815	65	34	31	2.6	1144	410	34	376	31.4
873	53	35	18	1.5	859	376	35	341	28.5	756	179	35	144	12.0
952	88	36	52	4.3	879	323	36	287	24.0	769	494	36	458	38.3
858	119	37	82	6.9	1114	342	37	305	25.5	770	446	37	409	34.2
683	442	38	404	33.8	1138	397	38	359	30.0	771	193	38	155	13.0
684	435	39	396	33.1	337	943	39	904	75.6	778	199	39	160	13.4
980	86	40	46	3.8	855	313	40	273	22.8	788	490	40	450	37.6
849	312	41	271	22.7	858	559	41	518	43.3	789	395	41	354	29.6
716	50	42	8	0.7	878	464	42	422	35.3	792	444	42	402	33.6
880	13	43	30	2.5	1136	409	43	366	30.6	53	24	43	19	1.6
856	237	44	193	16.1	817	15	44	29	2.4	66	66	44	22	1.8
914	24	45	21	1.8	684	303	45	258	21.6	72	92	45	47	3.9
926	10	46	36	3.0	1137	441	46	395	33.0	85	10	46	36	3.0
976	72	47	25	2.1	1060	161	47	114	9.5	86	15	47	32	2.7
1033	35	48	13	1.1	1008	1	48	47	3.9	87	2	48	46	3.8
967	25	49	24	2.0	1065	284	49	235	19.6	88	56	49	7	0.6
945	76	50	26	2.2	693	935	50	885	74.0	90	6	50	44	3.7

Table 1. The first 50 nodes in the Foursquare ranking (number of visits) and differences with the algorithmic ranking.

the APA algorithm (r_1). In the third column we can see the position of the Foursquare ranking (r_2). The fourth column shows the difference between the values of both rankings, in absolute value (dif). Finally, the last column represents the percentage deviation. The results, for the three categories, have been displayed: Food service and entertainment, Shops, and Malls and department stores.

If we calculate the average values in the percentage deviations of the rankings for these 50 nodes, we obtain the following results:

- Food service and entertainment: 5.40%.
- Shops: 25.76%.
- Malls and department stores: 14.70%.

As we have concluded in the quantitative analysis developed in section 3.1, the most comparable data within the commercial sectors is the Food service and entertainment, where most of the visits are focused. However, regarding the Shops sector, the data are not overly comparable due to the significant differences of the quantity of data collected.

It is significant that the average of the deviations of the 50 first nodes in the Food service and entertainment sector only reaches 5.4%. This means a significant overlap between the rankings offered by the algorithmic model and the preferences and recommendations given by the social network users.

If we focus on the Shops sector, the average deviation of the values in the rankings is about 25%. This result was expected given the differences observed in the volume of data. Taking the first 10 nodes in Table 1, the percentage deviation is only 11%. This reduction is significant and suggests that the most visited venues by Foursquare users in the Shops sector are also important locations in the network, after running the APA algorithm.

Considering the Malls and department stores sector, the average deviation of the values in the rankings is about 14%; an intermediate value between the other sectors studied. We can conclude that there is not the degree of similarity that occurs in the Food service sector, but neither the high dispersion that occurs in the Shops sector. Taking the first 10 nodes in Table 1, the percentage deviation is only 8%. The first two nodes that appear in the ranking of Foursquare users, correspond to the geolocation of the principal Mall in the city. Obviously, there is a correlation between the size of the Mall and the position in the rankings and it is an attractor which receives more than 450 visits in the social network. This fact is reflected in the algorithmic analysis conducted. Specifically, the node 832 (the closest to the Mall) appears in the second position of the Foursquare ranking, while it appears in the first position of the algorithmic ranking.

The bar chart in Figure 12 shows the differences between the arithmetical average in the three sectors analysed. It can be seen that the average of the percentage deviation in Food and Malls sectors is much lower than in the other. On the other hand, as might be expected looking at Figure 13, the standard deviations of the three sectors are high, giving an idea of the dispersion of the data.

When we make an estimate, an error is made but its magnitude is unknown. To avoid this, we must turn to the estimation by interval, which provides a range of values into which the estimated values are, with a certain probability called confidence level. Thus, considering the level of confidence and the interval width, the magnitude of the error can be assessed. That is, with the confidence interval we can make assumptions about what we expect for the values. Obviously, a greater confidence level implies a higher amplitude of the interval, and therefore less accuracy in the estimation. Usually, confidence levels are in the range of 90% to 99%. Table 2 summarizes the intervals in which the arithmetic average of the deviation between the ranking of Foursquare and the algorithmic model has a confidence level of 99%. That is, any estimates for the percentage deviation lies within the interval with a probability of 99%. After studying these ranges, we conclude that the amplitude of the intervals in the Food Services sector is much smaller than the rest of the sectors. With this, it can be estimated with a very high confidence (99%) that

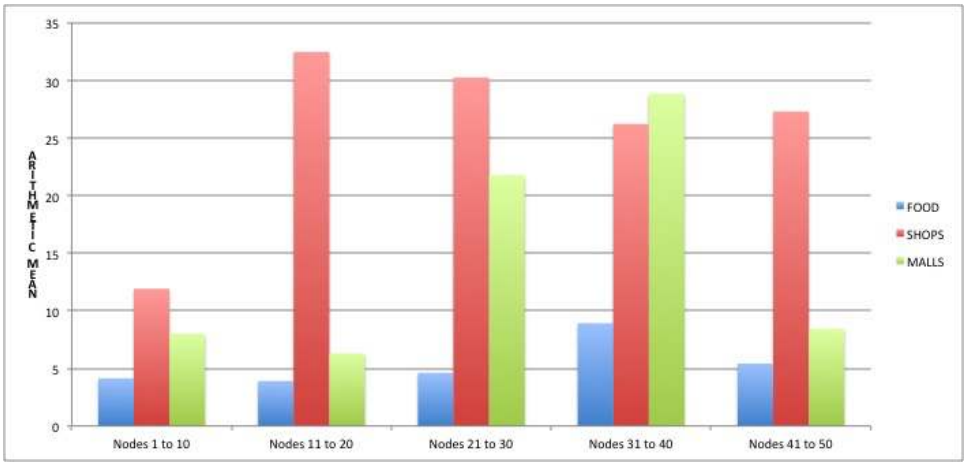

Figure 12. Arithmetic mean of the percentage deviation in the three sectors studied.

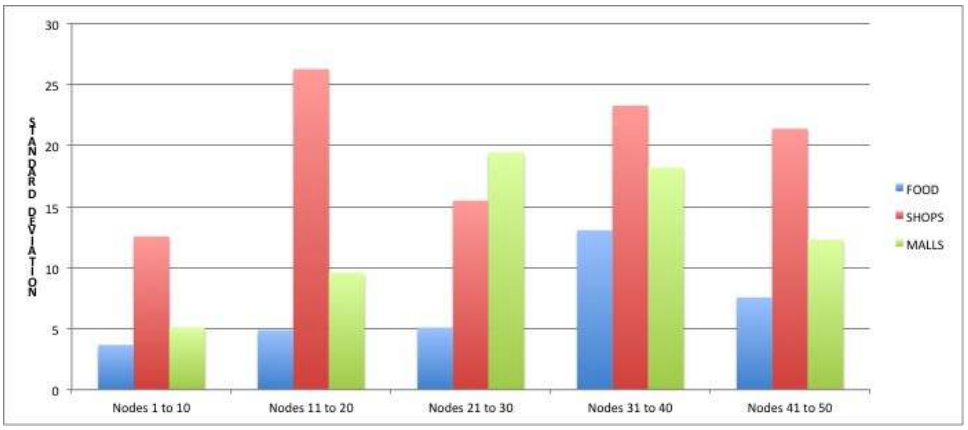

Figure 13. Standard deviation of the percentage deviation in the three sectors studied.

	Food service	Shops	Malls and department stores
Nodes 1 to 10	$[1.15, 7.13]$	$[1.68, 22.16]$	$[3.85, 12.17]$
Nodes 11 to 20	$[0, 7.9]$	$[11.05, 53.93]$	$[0, 14.12]$
Nodes 21 to 30	$[0.46, 8.78]$	$[17.64, 42.92]$	$[5.92, 27.64]$
Nodes 31 to 40	$[0, 9.58]$	$[7.24, 45.24]$	$[14.03, 33.71]$
Nodes 41 to 50	$[0, 11.59]$	$[9.86, 44.78]$	$[0, 18.46]$

Table 2. Intervals in which the average of the percentage deviation has a confidence level of 99%

the difference of the percentages between the Foursquare ranking and the algorithmic model ranking does not exceed of the range of the interval.

4. Conclusion

The initial question was whether there is a relationship between data generated by Foursquare users and the activity that actually takes place in the city, specifically those related to the commercial and leisure sectors. We studied the case of Murcia, located on the Mediterranean coast of Spain, where the good weather encourage to use of public

spaces in the city. The results of the study confirm that in fact there is a relationship between both data, especially when we study the Food service sector. Moreover, we can say that most venues that receive a large amount of visits by Foursquare users are public places that have a great importance in the topology of the urban network, especially when the city is a mathematical structure (primal graph).

Two comparative analysis have been performed. The first one is a quantitative analysis comparing the volume of data extracted from Foursquare Web service and the volume of data collected by fieldwork. From this quantitative comparison analysis, we conclude that Foursquare data in Murcia are focused on the Food service and leisure sector rather than other sectors such as shopping or business.

The quantity of data collected from the two sources allows us to perform a second analysis, which is a comparison between the ranking established by Foursquare users and the ranking given by the APA algorithmic model. Foursquare data provides a ranking of venues based on the number of visitors according to users tastes. This allows us to develop a visual map of urban activities in the city, based on the social network users criteria (less than 5% of the total population of the city). The data collected from the city (fieldwork) about its commercial activity and the application of the APA algorithm provides a new ranking of nodes according to their importance in the network.

To carry out the comparison, we divide the commercial activities into three sectors: Food service and entertainment, Shops, and Malls and department stores. Having performed this comparative analysis, we have detected a notable concordance among the most popular places, according to users of the social network, and nodes with a higher rate in the ranking of importance by the algorithmic model, especially in the Food service and entertainment sector.

According to the analysis the percentage deviation of each node of the urban network, per sectors, we conclude that in the commercial sector of Food service and entertainment is where the greatest concordance can be seen in the data, with an average, for the 50 first positions of the visits of venues ranking, of 5.4%. When we consider the 20 first positions, only a percent deviation of 4% occurs.

Another issue raised in the paper, is whether virtual databases such as Foursquare, can replace the data obtained from the fieldwork. It is clear that data from social networks give us a partial representation of reality that, in general terms, do not replace those obtained by a visual inspection of the physical reality. However, there are certain sectors of commercial activity in the fieldwork that are notably reflected by the dataset extracted from the social network Foursquare, as is the case with Food service sector.

Another conclusion is that the success of the public places depends on the large number of tertiary services located in their environment. It also depends on the connectivity of these places with other places supported by large number of tertiary services.

Summarizing, this work offers three main contributions. First, we use an algorithm based on APA algorithm for discovering the most relevant geographic areas of the city, taking into account the network topology and data distribution. Then, we introduce a methodology based on geolocated data provided by Foursquare for exploring the user tastes of this social network. Finally, we provide a comparison of both rankings using a network visualization model.

Acknowledgement

This work was partially supported by Spanish Govern, Ministerio de Economía y Competividad, which reference number is TIN2014-53855-P.

References

Agryzkov, T., Oliver, J.L., Tortosa, L. and Vicent, J., 2012. An algorithm for ranking the nodes of an urban network based on concept of PageRank vector. *Applied Mathematics and Computation*, 219, 2186–2193. doi:10.1016/j.amc.2012.08.064

Agryzkov, T., Oliver, J.L., Tortosa, L. and Vicent, J., 2014. Analyzing the commercial activities of a street network by ranking their nodes - a case study Murcia. *International Journal of Geographical Information Science*, 28 (3), 479–495. doi:10.1080/13658816.2013.854370

Barthelemy, M., 2001. Spatial networks. *Physics Reports*, 499 (1), 1–101.

Bawa-Cavia, A., 2011. Sensing the Urban: Using Location-Based Social Network Data in Urban Analysis. *First Workshop on Pervasive Urban Applications (PURBA).*

Berkhin, P., 2005. A survey on PageRank computing. *Internet Mathematics*, 2 (1), 73–120.

Bianchini, M., Gori, M., Scarselli, F.,2005. Inside PageRank. *ACM Transactions on Internet Technology*, 5 (1), 92–128.

Bonacich, P., 1987. Power and centrality: a family of measures. *American Journal of Sociology* 92, 1170–1182.

Brandes, U., 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163–177.

Cerrone, D., 2015. A Sense of Place. *Turku Urban Research Programmes*. Research Report 1/2015. Retrieved from http://beta.turku.fi/sites/default/files/atoms/files/

Ciuccarelli, P., Lupi, G., Simeone, L., 2014. Visualizing the Data City. *Social Media as a Source of Knowledge for Urban Planning and Management*, Springer, Heidelberg.

Cranshaw, J., Schwartz, R., Hong, J. I., Sadeh, N., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *Proceedings of the 6th International AAAI Conference on Webblogs and Social Media*, Trinity College, Dublin, Ireland.

Crucitti P., Latora V. and Porta S., 2006. The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and its Applications*, 369 (2), 853–866. doi:10.1016/j.physa.2005.12.063

Crucitti, P., Latora, V., Porta, S., 2006. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3), 036125.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. textitSociometry, 40(1), 35–41.

Friedkin, N.E., 1991. Theoretical foundations for centrality measures. *American Journal of Sociology*, 96, 1478–1504.

Hong, I., 2015 Spatial Analysis of Location-Based Social Networks in Seoul, Korea. *Journal of Geographic Information System*, 7, 259–265. doi:10.4236/jgis.2015.73020

Jeong, S.K., Ban, Y.U., 2011. Computational algorithms to evaluate design solutions using Space Syntax. *Computer-Aided Design*, 43(6), 664–676. doi:10.1016/j.cad.2011.02.011

Jiang, B., 2009. Ranking spaces for predicting human movement in an urban environ-

ment. *International Journal of Geographical Information Science*, 23 (7), 823–837. doi:10.1080/13658810802022822

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., et al., 2009. Computational Social Science. *Science*, 323 (5915), 721–723.

Katz, L., 1953. A new index for sociometric data analysis. *Psychometrica*, 18, 39–43.

Lindqvist, J., Cranshaw, J., Weise, J., Hong, J., Zimmermann, J., 2011. Im the Mayor of My House: Examining Why People Use foursquare a Social-Driven Location Sharing Application. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems HCI'11*, New York, USA, pp. 2409–2418.

Newman, M.E.J., 2003. A measure of betweenness centrality based on random walks. *Research Report*. Retrieved from http://arXiv:cond-mat/0309045.

Noulas, A., Scellato, S., Mascolo, C., Pontil, M., 2011. An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 570-573.

Noulas, A., Mascolo, C., Frias-Martinez, E., 2013. Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. *Proceedings of the IEEE 14th International Conference on Mobile Data Management*, Milan, Italy, 167-176.

Offenhuber, D., Ratti, C., 2014. Decoding the City. *Urbanism in the Age of Big Data*, Birkhuser, Basel.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The pagerank citation ranking: Bringing order to the web. *Technical report 1999-66*, Stanford InfoLab, Stanford, CA, USA.

Porta, S., Crucitti, P., Latora, V., 2006. The network analysis of urban streets: a primal approach. *Environment and Planning B: Planning and Design*, 33, 705–725.

Roick, O., Heuser, S., 2013. Location based social networks - definition, current state of the art and research agenda. *Transactions in GIS*, 17, 763–784. 10.1111/tgis.12032

Serrano-Estrada, L., Serrano-Salazar, S., Alvarez F., 2014. Las redes sociales y los SIG como herramientas para conocer las preferencias sociales en las ciudades turísticas: el caso de Benidorm, *XVI Congreso Nacional de Tecnologías de la Información Geográfica, Tecnologías de la información para nuevas formas de ver el territorio (in Spanish)*, Alicante, Spain, 1005–1012.

Sohn, K., Kim, D., 2010. Zonal centrality measures and the neighbourhood effect. *Transportation Research Part A: Policy and Practice*, 44 (9), 733–743.

Silva, T.H., Vaz de Melo, P.O., Almeida, J.M., Salles, J., Loureiro, A.A., 2013. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, ACM*, New York, USA.

Silva, T.H., Vaz de Melo, P.O., Almeida, J.M., Salles, J., Loureiro, A.A., 2014. You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. *Proceedings of 8th AAAI International Conference on Weblogs and Social Media, AAAI*, Michigan, USA.