Dissecting the Urban Landscape for a better Management

2020

4 MAYO

Made in Game
Author: Jordi Planas

Table of Contents

INTRODUCTION

Statement of the business problem

Is there a relationship between the socioeconomic status of people living in a district and the profile of venues in it?

Target audience

- Companies wishing to stablish in Girona or any catalan city
- The local city council

Structure of the report

Data acquisition and data preparation.

- Fetch geographical information, process it and feed it to the Foursquare queries.
- Process the results of Foursquare queries
- Fetch and process statistical data

From data to information.

- Apply machine learning methodologies and techniques to extract relevant information out of data
- Data visualization will help on the understanding of the insights obtained so far.

Background

Cities, as large human settlements, flourished in parallel with the development of agriculture and have been characterized by the specialization and social division of labor. Great cities have existed in all civilizations of the human history and they have been planned and managed by bureaucracy systems which needed data science to do the job. One example that I like to use to link cities and data science is the development of the concept of the number zero. Even if in some of them the number zero was not invented, most of the ancient civilizations developed the concept of zero and used it for calculations in many areas like accounting, astronomy, geography etc.

Cities are great places to enhance human interaction both for the good and for the bad. In both cases data science is paramount in improving human life. In this project I develop a process to gather data from cities which can be used by different stakeholders. I'm mainly targeting

Companies which want to explore the socioeconomic landscape of cities to make informed decisions about their businesses

Government and governmental agencies seeking to improve resource allocation, from the deployment of police forces to budget allocation.

In all these cases, a fine granulation of the city would be of major interest as it allows us to make a zoom to small patches that can be also aggregated conveniently. Often, postal code areas are used to do this division of the city, but I think that census areas are a better choice.

Census areas in Catalonia are areas where there are between 500 and 2000 electors. One census area cannot include territories of more than one municipality, and they can be grouped administratively into census districts. Moreover, official statistical data is geographically segmented by census districts.

In this project, I use Girona as a case study that can be later extrapolated to other cities in the country. Girona will be divided in census areas and from this reference area, Foursquare will be used to locate venues in neighborhoods.

Mapping venues in towns is the first step in the process of typifying census areas with relation to the profile of facilities and commercial activities. This profiling can be complemented and/or further correlated with other statistical data like socioeconomic status of inhabitants.

For example, imagine that an area is characterized by the presence of restaurants and small households occupied mainly by young people. A real state company can focus its efforts in selling or hiring in this area to this profile of young single or couple wealthy people. If there is another area with parks and supermarkets, and households are of an average four people, this means that it may be a family neighborhood. A supermarket chain may consider it as a target neighborhood.

For governmental stakeholders, checking the reviews of businesses around an area might be a very useful source of information to plan specific support actions for this or that particular area.

DATA

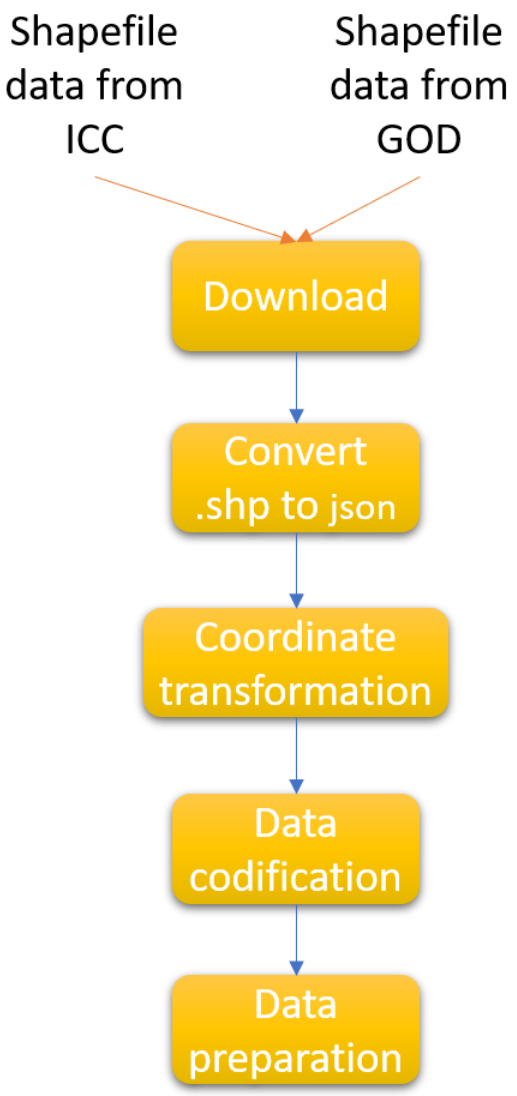
Though we live in times of data hype and terabytes of data accumulate endlessly, data is often kept at very different places, in different formats. Fortunately, there are official governmental sources of data which are reliable. These sources are not completely coherent in terms of format, but they are manageable. In this project I'm going to use three main sources of data:

- Data from the Cartographic Institute of Catalonia ICC
- Data from Girona's local council that can be obtained either from Girona Open Data GOD or from l'Observatori.
- Data from Foursquare

METHODOLOGY

Prepping data

Geographic data



In the figure beside, there is a schematic representation of the transformations that have been performed in the Geografic data from ICC and GOD. In the original sources, geographical data is kept in Shapefile format which is a common format used in geographical information systems (GIS). However, the geojson format is increasingly used in data science and is the format that is used to feed map tools like Folium and Choropleth. Thus, the first step will be to transform .shp files into .json files thanks to a library ***PyShp***. I will define the function `shpToGeoJSON` were I will use the PyShp tools to read .shp files and return the information in the form of python dictionaries.

In both ICC and GOD the geographical data is kept in utm coordinates, but many current applications are better fed with WGS84 coordinates. So, I will transform the utm coordinates to WGS84 coordinates with the help of a library ***PyProj***. For this job I'll create the function `getWGS84Coordinates`.

Once the data is in json format and the coordinates are coded with the WGS84 system, I'll make some small transformations like changing some town codes by the official name of the town, this is for the sake of readability of data when observing them. I'll also create a merged code joining the district code and the section code to uniquely identify each census section in Girona.

Foursquare data

Foursquare is a powerful database which can be easily accessed with APIs provided in its developer page. However, there are some limitations that forces us to do some extra work. The main limitation I found is that the results of searches for venues are given in a circular area defined by a radius around one point. But for many applications it is important to to assign each venue to a predefined area which might have the shape of a complex polygon like the census section. In order to achieve that I will make the following steps:

1. Determine the centroid of each census are which are, for the most, quite irregular polygons. For this I use the library scipy.spatial which has a tool to calculate the convex hull of a polygon. Then the centroid of the convex hull is the mean of the x coordinates and the mean of the y coordinates of the hull. I used this approximation because it gives better results than the simpler approximation of determining the centroid as the means of the points of the raw polygon (data not shown).
2. Once the centroid is found I use the function distanceBetweenPoints to determine the longest distance between the centroid and all the vertices. This will be the distance used to assign the parameter RADIUS in the Foursquare search.
3. Next step is to retrieve Foursquare data using latitude and longitude of the centroid of each census section and the radius as the longest distance between centroid and vertices. But this will yield a lot of overlapping information as the circle defined by the radius in each census section will intersect with the circles of the neighboring sections.
4. Once venue data are obtained, venues that do not belong to the census section must be filtered out. In order to do this, I'll have to check if venue coordinates are inside the polygon which defines the census section. This can be easily done using the function polygon.contains of the Shapely library that I'll execute inside the local function ifInDistrict.
5. The final step is to convert all data in a Panda's data frame.

Statistical data

Statistical data can be easily downloaded in csv formate and then uploaded to the Jupiter notebook as a Pandas data frame.

Machine learning

In this project I have been using a clustering algorithm to group neighborhoods based on their similarity. One of the limitations of most clustering algorithms is that they have to be fed with the final number of clusters *k*. There is no proper way to optimize the number of clusters so the result is always dependent on the initial choice of the analyst. The only technique to optimize the number of clusters is the elbow technique. Plotting the within mean of squares against *k* often gives a descending curve which have a pronounced slope up to a point in which the slope decreases. The *k* value just before the breaking point is assumed to be the optimal *k* value.

In this project I have processed our data with three different algorithms from the sklearn library:

- K-Means
- Spectral Clustering
- Agglomerative Clustering

I have been using different *k* values ranging from 2 to 7 because we only have 9 samples corresponding to the 9 neighborhoods in Girona. Thus, I have performed 18 clustering analysis for each data set. As my data sets contain a limited number of samples, I explored an alternative method to the elbow method for parameter optimization.

After performing the 18 clustering analysis, I processed the results to obtain the optimum k with the following criteria:

1. Select the combinations of algorithm + *k* value that yield the maximum number of clusters with more than one element in it.
2. If there are more than one such combinations, then we choose the minimum *k* value greater than 2.

With this rather simple criterion I can ensure that samples are distributed as evenly as possible among clusters and that the number of clusters with only one element is minimal.

RESULTS

Parameter optimization

In the following table we can see the 18 results of applying the different clustering conditions to Foursquare data on venues in Girona neighborhoods. The results are expressed in lists that contain the labels of each of the 9 samples. Thus, if position 1 of the list is "1" and postion 7 is also "1" it means that sample 1 and sample 7 belong to the same cluster. The results are grouped in columns by clustering method and rows by *k* value.

	K-Means	Spectral	Agglomerative
N clusters	kluster labels	kluster labels	kluster labels
2	[1 1 1 1 1 0 1 1 1 1]	[0 0 1 0 1 0 0 0 0]	[0 0 0 0 1 0 0 0 0]
3	[1 1 1 1 1 0 2 1 1 1]	[0 0 0 0 2 1 0 0 0]	[0 0 0 0 1 2 0 0 0]
4	[1 1 1 1 0 3 2 1 1 1]	[0 0 0 3 2 1 0 0 0]	[0 0 0 3 1 2 0 0 0]
5	[2 2 4 0 1 3 2 2 2]	[1 1 0 4 2 3 1 1 0]	[0 0 4 3 1 2 0 0 0]
6	[2 2 4 0 1 3 2 5 2]	[2 2 0 5 3 1 2 4 0]	[0 0 4 3 1 5 0 2 0]
7	[4 4 5 0 3 2 4 6 1]	[1 1 4 3 0 5 1 6 2]	[0 0 4 3 1 5 0 6 2]

Here we can see two things

1. K-Means and Agglomerative clustering methods yield exactly the same results. Spectral clustering yields the same results when *k* is 3, 4 and 7.
2. There are a few differences in the Spectral clustering results for k 1, 5 and 6 with respect to the other two methods. In all three cases the samples are more evenly distributed among clusters than in the other two methods.

Regarding optimization of *k* values, according to the proposed optimization method, the best clustering conditions are with Spectral clustering and *k* 5. This is because it is the method that clusters samples with two clusters with more than one sample in it and it is the lowest k value different than 2.

Now, if we look at the other two data sets, the ones with socioeconomic data, we see the following clustering.

Education Level

N clusters	K-Means kluster labels	Spectral kluster labels	Agglomerative kluster labels
2	[1 1 0 0 1 0 0 0 1]	[0 0 1 1 0 1 1 1 0]	[1 1 0 0 1 0 0 0 1]
3	[1 1 2 0 1 2 2 2 1]	[0 0 1 1 0 2 2 2 0]	[1 1 0 2 1 0 0 0 1]
4	[1 1 3 0 1 2 2 2 1]	[3 3 2 1 3 0 0 0 3]	[0 0 3 2 0 1 1 1 0]
5	[1 1 4 0 3 2 2 2 1]	[0 0 2 1 4 3 3 3 0]	[1 1 3 2 4 0 0 0 1]
6	[1 1 3 0 4 2 2 5 1]	[5 0 1 2 4 3 3 3 5]	[0 0 3 5 4 1 1 2 0]
7	[4 1 3 0 5 2 2 6 1]	[5 1 0 2 4 3 3 3 6]	[6 0 3 5 4 1 1 2 0]

People in household

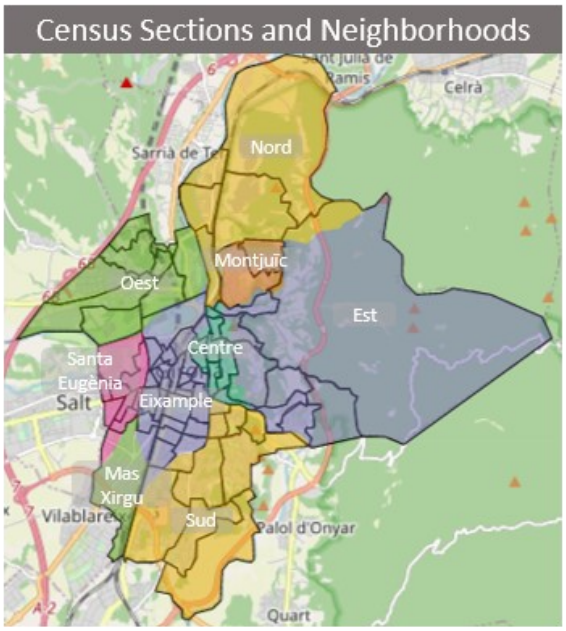
N clusters	K-Means kluster labels	Spectral kluster labels	Agglomerative kluster labels
2	[1 0 0 1 0 0 0 0 0]	[0 0 0 1 0 0 0 0 0]	[0 0 0 1 0 0 0 0 0]
3	[2 2 1 0 1 2 1 1 1]	[0 0 0 1 2 0 0 0 0]	[2 2 0 1 0 2 0 0 0]
4	[0 0 3 1 2 0 3 3 3]	[3 0 0 1 2 0 0 0 0]	[0 0 2 3 1 0 2 2 2]
5	[2 3 1 0 4 3 1 3 1]	[0 4 2 1 3 2 4 2 4]	[4 2 0 3 1 2 0 0 0]
6	[2 3 5 0 4 3 1 5 1]	[4 0 5 1 3 0 2 0 2]	[4 2 0 3 1 2 5 0 5]
7	[2 3 5 0 4 3 1 6 1]	[4 5 6 0 2 1 3 1 3]	[4 0 6 3 1 0 2 5 2]

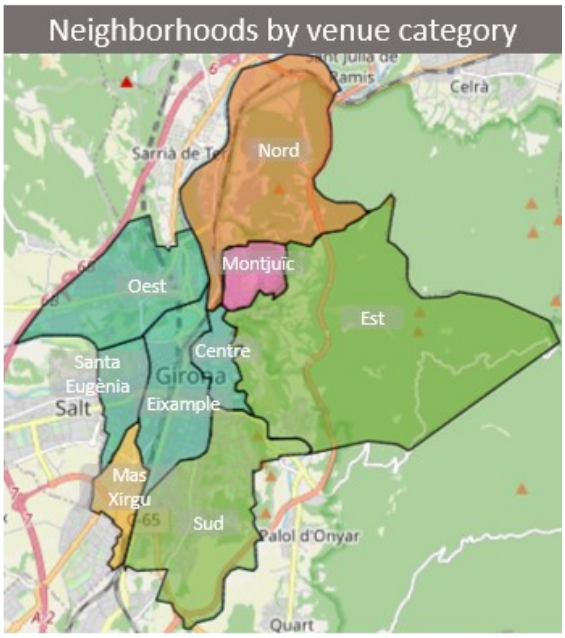
As happened with the Foursquare Venues data set, there are many identical results across the three methods, but there are also som differences. In summary, according to my optimization criteria, the best clustering conditions for the data set of population education evel asre with Spectral clustering and *k* 3. Instead the optimal clustering conditions are for the data set containing the number of people living in each household are obteined with K-Means method and *k* 6.

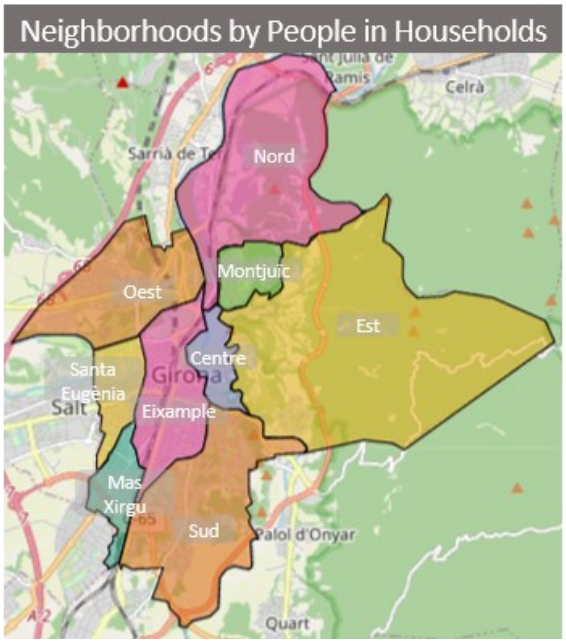
Girona neigborhoods

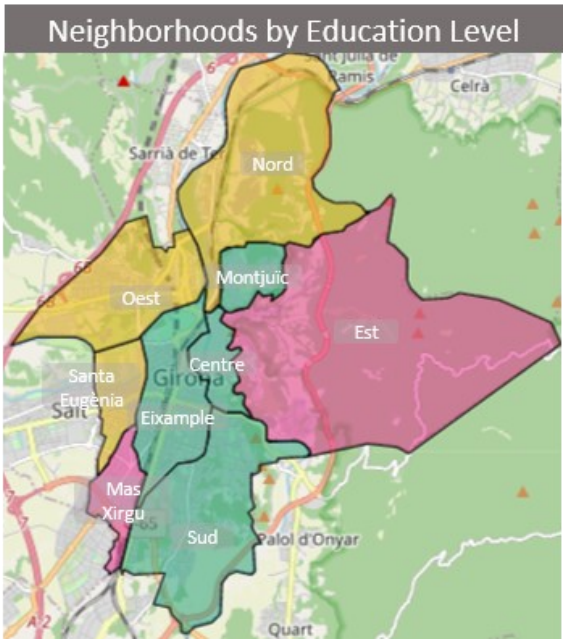
After a marathon of data processing and parameter optimization, we are at the final 100 m of the finish line, were we are going to uncover the results of clustering neighborhoods.

In the following figure, there are 4 maps of Girona with different information. The map at the top left side is the basic map where I show the overlapping between census sections and neighborhoods. A part from some small non-overlapping, the neighborhoods are clusters of census sections equivalent to census districts.









In the top right map, you can see the result of clustering neighborhoods by the profile of venue categories. Here you find a main cluster composed of four neighborhoods, then there is a cluster with two big neighborhoods and finally there are three neighborhoods each making a separate cluster.

When looking at education level there is also one big cluster containing 4 neighborhoods, but its composition is not the same than in the big cluster of venues category. Then, there is a cluster with three neighborhoods and finally, a cluster with two neighborhoods.

If we look at the number of people in households, we see a quite diverse map with three clusters with two neighborhoods and then three neighborhoods that cannot be clustered.

DISCUSSION

Neighborhoods by venues

The total amount of venues recorded in the Foursquare database from Girona is less than 300, a number which is far below the actual number of venues. Girona is a very dynamic town and the commercial center of a rather big geographical area. It accumulates many governmental offices and it hosts a university which has had a great impact on the city over the last 30 years. However underrepresented, the Foursquare venues are able to capture the special profile of the different neighborhoods. There is a great cluster containing four neighborhoods, Centre, Eixample, Santa Eugènia and Oest. They are the four neighborhoods that constitute what we can call the center of the city. Although the socioeconomic profile of these four neighborhoods is not necessarily similar in the streets there is a complex mixture of restaurants, supermarkets, groceries, hairdressers etc. All four neighborhoods are places with all you need to live and, at the same time, in most of them there are locations were people from other neighborhoods may eventually go shopping.

Next cluster is formed by the neighborhoods Est and Sud. They are very different neighborhoods in term of socioeconomic landscape. While Est is a low-income neighborhood, Sud has areas with people with very high income. However, the venues profile is similar and it is basically characterized by almost no retailers nor personal services, and the presence of some municipal equipment.

Finally, we have three clusters with only one neighborhood. Montjuïc is a residential area with some gardens. Nord is a mixture of working-class immigration and working-class native population which is located a bit far away from the city center. It is by no means a commercial place, but there you can find some specialized businesses. The las one, Mas Xirgu, is a very special one, we see it coming alone in every clustering because it is basically an industrial area with special business like car dealers, garages, pet shops and clinics, and industrial providers.

Neighborhoods by education level

When looking at Girona by its venues profile the main feature was the presence of a great center comprising four neighborhoods. When we look at the education level, the city is split into three clear areas and the center is also split. Now, Montjuïc joins the Centre, Eixample and Sud neighborhoods. These are the places where people with the highest education is living. Next, with middle education levels we find the west crown of the city, Santa Eugènia, Oest and Nord. Finally, neighborhoods Est and Mas Xirgu cluster together even if their profile is not exactly the same. Both have a lot of people without primary studies or just with primary studies. This is probably what makes them clustering together. But Mas Xirgu

has half of people under 16 as compared with Est and Mas Xirgu. Instead Mas Xirgu has almost 16% of people with university studies while in Est neighborhood it is a bare 3%.

Neighborhoods by number of people in households

Clustering by the number of people in households is often a good way to reveal socioeconomic features of neighborhoods and a good way to reveal familiar structures if any. Clustering Girona according to number of people in households yields a rather atomized map, with three clusters with two neighborhoods in it and three neighborhoods that come alone. First of all, the neighborhood Centre is a very special one, one of those which has experienced a process of gentrification due to the pressure of tourism and university students. Here 65% of houses have one or two residents, while only 7% of the houses have more than 4 residents. Montjuïc is another singularity, it has the lowest proportion of houses with a single resident and 75% of houses have between 2 and 4 residents, it is the prototypical middle-class family neighborhood. Mas Xirgu comes again as a special case, it has the highest number of houses with only one resident and together with houses with two residents make up 75% of households. Then there is a 22% of households with four or five residents, showing the presence of two very different areas in the same neighborhood.

Neighborhoods Est and Santa Eugènia probably cluster together because they are the two neighborhoods with the highest proportion of houses with 5 residents or more with more than 16% of households.

Neighborhoods Sud and Oest are family neighborhoods with more than 90% of households with 1 to four residents, the class with two residents per house being the one with more occurrences.

In neighborhoods Nord and Eixample the leading categories are the hoses with one or two residents and then, at a certain distance, families with 3 or 4 members.

In summary, the profile of Girona is that of a modern city were families are getting smaller and most people live alone or just in a couple. However, slight differences are relevant and show that the map of the city is quite diverse.

CONCLUSIONS

- Foursquare might not be the most comprehensive database to make a venues profile of the city of Girona. In the future this should be complemented with data from other databases like google maps.
- Girona is a small town of 102 thousand people but it is a quite diverse city we can say that neighborhoods matter, they matter in terms of venues and in terms of

socioeconomic profile. Clustering them by different criteria gives different results showing that having the right information might be crucial for city management.

- Future directions.
    - Complementing Foursquare data with data from other databases
    - Adding data from other socioeconomic variables.
    - Define a socioeconomic index might help a lot on delivering a very focused information to stakeholders.
    - Tracking venues to follow food traffic in time series would be of great interest for stakeholders but this will be done after confinement measures are relieved.