Identification of zones with a high density of green areas in Santiago, Chile

Mauricio Santelices

July, 2020

1. Introduction

1.1 Background

Chile is among the countries with the highest obesity levels in the world, which, together with the high rate of sedentary lifestyle, represents a serious risk to the health of the population. In Santiago, the capital, lives about a third of the country's population, so it's expected that the largest number of patients with sedentary lifestyle and obesity-related diseases come from this city.

In recent times, plans have been enacted to promote a healthy lifestyle, changes in food habits and the practice of outdoor physical activity in the various parks of the country, however, the city of Santiago presents an uneven distribution of green areas (GA) that could be used for this purpose, so it is important to determine the distribution of these areas in order to plan improvements in access to parks or in the construction of new GAs for the citizens.

1.2 Problem

The purpose of this project is to identify the areas with the highest density of GAs in the Santiago Metropolitan Area (SMA), made up of 36 communes (the name that each administrative unit of the city receives) through a geospatial analysis based on the coordinates (latitude and longitude) of each important GA in the city.

1.3 Interest

Organizations that develop plans that promote healthy lifestyle can use the distribution of GA in the city in order to plan according to the conditions of access to them for residents of a given area. Entities responsible for proposing improvements to public transport or construction can use this information to improve access to existing GAs or to create new GAs in low-density areas.

2. Data acquisition and cleaning

2.1 Data sources

The list of communes that make up the SMA can be found here, section "Comunas del Área Metropolitana de Santiago de Chile". The table lists each commune, its general location on the map and its population as of the last census (year 2017), among other data, however, it does not have the coordinates of each commune. This list was obtained using scraping techniques and the coordinates of each commune through geolocation libraries.

Additionally, we proceeded to obtain a list of venues through the Foursquare API for each commune, using a wide search radius in order to collect as many venues as possible.

2.2 Data cleaning

Once the list of all possible venues for each commune was obtained, we proceeded to select only those categorized as "Plaza", "Park" or "Garden", in addition to removing any duplicate value, since in the collection phase, se chose to use a very wide radius for each commune in the city.

2.3 Feature selection

After cleaning the data, a list of 37 unique GAs registered on Foursquare with their coordinates (latitude and longitude) was obtained. Because the objective of the study is to determine high-density areas and that these areas can encompass different neighboring communes, we decided to discard the commune name, in which each GA is located, so the final features only include the name of each GA, its latitude and its longitude.

	Venue	Venue Latitude	Venue Longitude
0	Parque El Rosal	-33.485326	-70.765686
1	Running Fernández Albano	-33.523544	-70.660307
2	Viveros de Renca	-33.403115	-70.734140
3	Parque Forestal	-33.436429	-70.640126
4	Parque Bicentenario	-33.400747	-70.601693

Table 1: Extract of the final table obtained after cleaning the data.

3. Exploratory data analysis

3.1 GA locations

To visualize the GAs on the map, first it was necessary to obtain the coordinates of Santiago (-33.4377968, -70.6504451), for which the Nominatim library was used. The creation of the map was done using the Folium library.

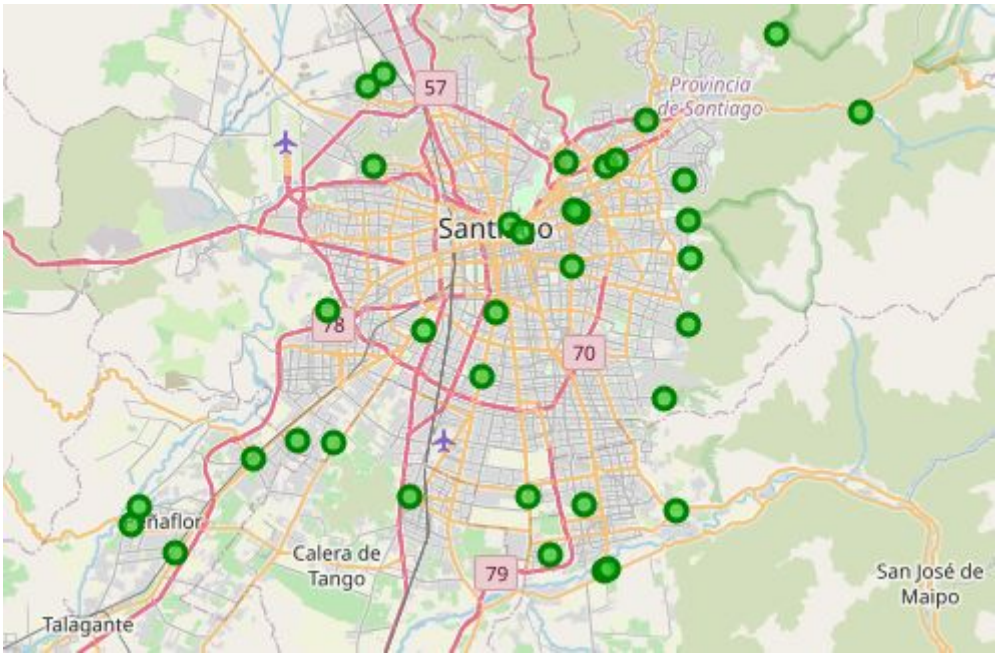

Figure 1: GA locations on Santiago

3.2 Calculation of the optimal parameters for DBSCAN

DBSCAN is a clustering algorithm that finds a number of groups based on the density of the points. The algorithm takes as parameters the epsilon value (E) and the minimum number of elements (M) that a cluster must include.

The E value represents a distance threshold between two points. If two points have a Euclidean distance less than E, they are considered as neighbors.

To calculate the distance between two points we can use the Nearest Neighbors (NN) method. In this case, the distance calculation was performed taking into account 2 neighboring points. By ordering and plotting the distances, we obtain a graph like the one observed in Figure 2, where the point of maximum curvature represents the optimal value of E. For this analysis, after different tests, it was decided to round down the value of the maximum curvature, obtaining an E of 0.4.

Regarding the M value, due to the low amount of data and by observing its distribution on the map (Figure 1), we decided on a value of 2.

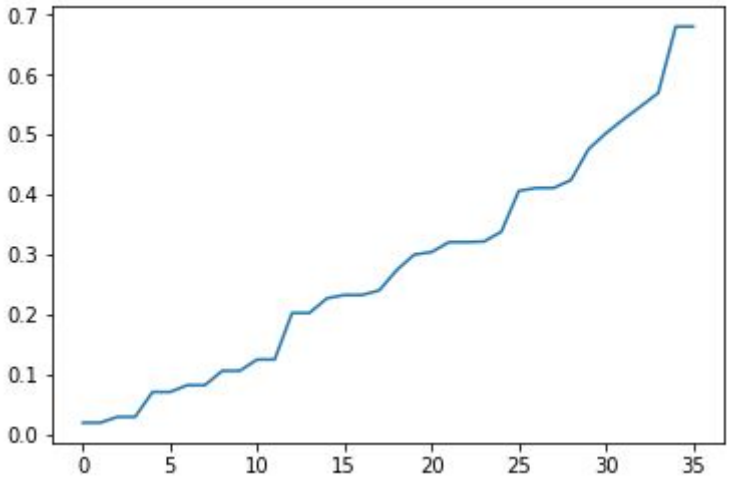

Figure 2: Distance between two neighboring points. The value corresponding to the maximum curvature was rounded to 0.4

4. Results

Applying the DBSCAN algorithm with E = 0.4 and M = 2, six clusters were obtained, each one identified by a color other than red. The red elements represent GAs that do not belong to any Cluster, that is, they are not close to another GA. The distribution of the clusters can be seen in Figure 3 and the legend in Table 2.

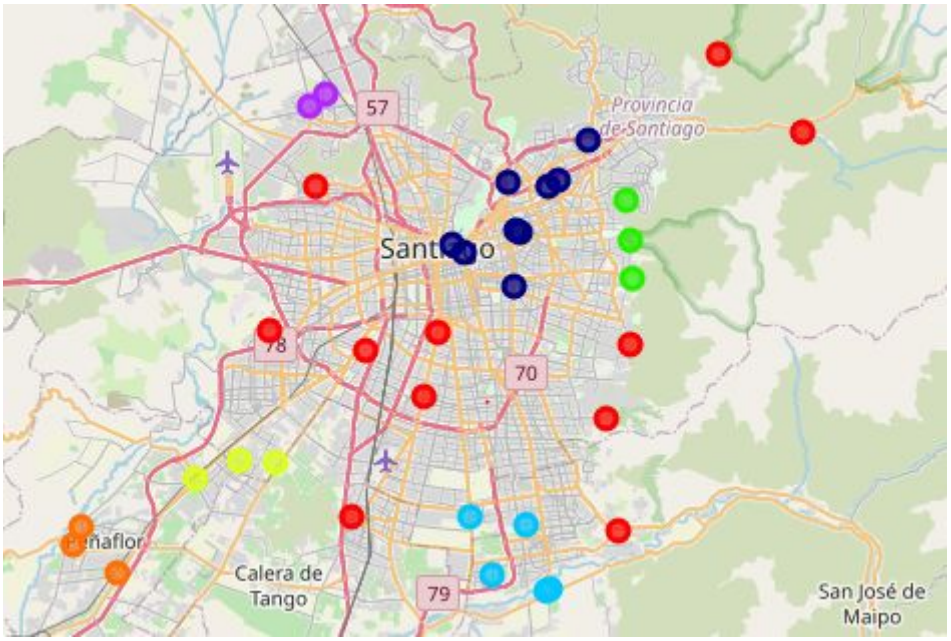

Figure 3: GA clusters in Santiago. Items in red are not part of any cluster.

Cluster	Color	Location
1	Blue	Northeast
2	Light Blue	Southeast
3	Green	East
4	Yellow	Southwest
5	Orange	Southwest
6	Violet	Northwest

Table 2: Legend of the distribution observed in Figure 3

5. Discusión

After applying the DBSCAN algorithm, some expected results are observed, specifically Clusters 1, 4, 5 and 6, however, Clusters 2 and 3 were expected to be more extensive, also, we expected the appearance of a Cluster in the central area of the city.

Regarding Cluster 1, it is a totally expected result, considering the country situation. In the northeastern sector of the city reside the citizens with the highest income level and at the same time, these communes have the largest amount of funds to invest. This is a trend that is repeated throughout the country, the higher income sectors harbor a greater number of GAs, which in turn are well used by the community.

At the same time, several GA clusters are observed on the outskirts of the city, however, it is known that in these areas, GAs are not well used, which may be due to several reasons. First, these communes represent lower income areas and, administratively, they have fewer funds to invest in the community, which leads to a lack of knowledge of the possibility of using the GAs, lack of incentive and lack of activities to carry out by the residents. Another important factor to consider is the higher crime rate observed in these communes, making some of the GA dangerous for the community.

The downtown area has a series of GAs but not close enough to form a cluster, one of the possible reasons is having favored residential and commercial solutions, since they are highly residential communes. Lack of access to GA can promote a sedentary lifestyle, stress and other problems.

6. Conclusión

The identification of GAs based on Clusters allows to visualize the current situation of the city regarding the distribution of the GAs. It allows to implement plans for its use or plans for the construction of new areas in communes with low GA density. The peripheral communes have a high density of GAs, however, these are not well used, so these communes are good candidates for plans that promote the use of GAs. On the other hand, the communes of the center of the city present a low density of GAs, being candidates for future plans for the construction of new GAs.