Segmentation analysis of Bahia state in regions based on local economic activities

Daniel de Amaral da Silva[1,2]

[1]Departamento de Estatística e Matemática Aplicada (DEMA), Centro de Ciências, Universidade Federal do Ceará, Fortaleza - CE, CEP 60.440-900, Brasil.

[2]GREat, Departamento de Engenharia de Teleinformática (DETI), Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza - CE, CEP 60.455-970, Brasil.

danielamaral@alu.ufc.br

Abstract. *Bahia, is one of the most heated tourist spots in Brazil, with a welcoming population, beautiful beaches, several historic centers, and of course, one of the most outstanding cuisines in Brazil. This paper proposes a cluster analysis to define regions of the state based on local economic activities, taking as a reference each county with its estimated radius. The datasets used to come from two sources, the first is the counties and their respective areas in km^2 obtained from the database of the Brazilian Institute of Geography and Statistics (IBGE) and the second are venue data from the Foursquare API. After manipulating and grouping data, we obtain 5 clusters/regions that provide us with the main economic characteristics of each set of municipalities, such as areas of a high density of restaurants or coastal areas with a high density of beaches.*

1. Introduction

Bahia is one of the states that most receive tourists in Brazil. According to the State Tourism Secretariat (Setur), six million tourists visited the region during the summer of 2018/2019. The state has an area of approximately 567,733 km, 417 counties, and is home to more than 15 million inhabitants. It is located in the Northeast region and is the main destination for tourists due to the beautiful beaches and coastal cities [da Silva 2007].

The state and its capital (Salvador) since the first half of the 90s have been contemplated with a series of public investments directed to the tourist activity, through specific programs, such as the Tourism Development Program in the Northeast - PRODETUR - or investments made with resources from the State Treasury or other sources, such as the World Bank (IBRD), the Kreditanstalt Für Wiederaufban (KFW), the National Bank for Economic and Social Development (BNDES), the General Tourism Fund (FUNGETUR). Undoubtedly, Salvador and its surroundings have an immense tourist potential that has been attracting national and international groups, mainly interested in making investments in the area of lodging/leisure equipment, building hotels, resorts, and inns [da Silva 2007].

The Bahian region has one of the most striking cultures in the Brazilian territory, which is why it ends up producing regions with unique characteristics, its extreme south is an example [Almeida et al. 2008]. The main objective of this work is to try to group counties in that state to create profiles of specific economic activities. These profiles

can help micro-entrepreneurs looking for warm regions with a certain activity profile, guide people looking for properties with a pre-established neighborhood, for example, a couple who have children would opt for a region that has schools, parks, and supermarkets nearby.

When we consider these issues, we can create a map and try to group the counties of Bahia according to the local economic density.

2. Data

We use two data sources mainly, the first from the Brazilian Institute of Geography and Statistics (IBGE) and the second from the foursquare API.

2.1. IBGE County Data

We obtain the dataset composed of two columns of information from each county in the state of Bahia, they are County Name (Neighborhood) and Area (km^2). Table 1 shows the data distribution.

Table 1. Distribution of data counties in Bahia state

Neighborhood	Area (km^2)
Alcobaça	21,690
Almadina	8,835
⋮	⋮
Tanhaçu	1,277,514
Ubaitaba	181,102

After reading the data, we use the Geopy geolocation API to provide the central coordinates of latitude and longitude based on the names of each county. We also created a new column called $radius$, where we take the following formula.

$$radius = \sqrt{\frac{Area}{\pi}}$$

This method of obtaining the radius is a naive approach, as we consider that each county has the geometry of a perfect circle. Grouped all variables, we now have a dataset with neighborhood names, latitude, longitude, radius (km) and area (km^2). Table 2 shows the data distribution.

Table 2. Distribution of data counties in Bahia state with coordinates

Neighborhood	Latitude	Longitude	Radius (km)	Area (km^2)
Alcobaça	-17.521587	-39.196547	21.690	1,477.929
Almadina	-14.704660	-39.638199	8.835	245.236
⋮	⋮	⋮	⋮	⋮
Tanhaçu	-14.019662	-41.247271	20.165	1,277,514
Ubaitaba	-14.311777	-39.322660	7.593	181,102

2.2. Foursquare Venue Data

Through the coordinate data (Latitude and Longitude) obtained in Subsection 2.1, we can obtain venue data from the foursquare API with a radius of search pre-established. The API provides a JSON with a variety of information, such as venue types, name of the venue, score, comments by visitors, photos, etc. Table 3 shows the data distribution.

Table 3. Distribution of Venues by data counties in Bahia state with coordinates

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alcobaça	-17.521587	-39.196547	Farol de Alcobaça	-17.519936	-39.193818	Lighthouse
Alcobaça	-17.521587	-39.196547	Cabana Do Compadre	-17.517453	-39.192121	Seafood Restaurant
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Ubaitaba	-14.311777	-39.32266	lojas guaibin	-14.313340	-39.330338	Department Store
Ubaitaba	-14.311777	-39.32266	Baiano do Acarajé	-14.319062	-39.326327	Food Truck

After storing all the data we need, we will transform it until it is ready for clustering. We created a new sparse table that will contain the name of each 138 county/neighborhood with more than 10 venues and 265 columns that represent the average presence of the 265 unique venues categories, therefore, the data for clustering has the dimension of (137, 266). Table 4 shows the sparse data distribution.

Table 4. Sparse table of average venues presence in each county

Neighborhood	Women's Store	ATM	Acai House	⋯	Waterfront	Whisky Bar	Wine Shop	Wings Joint
Abaré	0	0	0	⋯	0.0625	0	0	0
Alagoinhas	0	0	0	⋯	0	0	0	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
Vitória da Conquista	0	0	0	⋯	0	0	0	0
Xique-Xique	0	0	0	⋯	0	0	0	0

3. Methodology

The Algorithm 1 shows the pseudocode for k-Means.

We used the statistical tool of clustering k-Means [MacQueen 1967] for the ten variables of most common venues. the k-Means algorithm is one of the most common cluster methods of unsupervised learning.

The Algorithm logic is based on starting from an array of observations $\mathbf{X}$ start k random centroids that belong to $\mathbf{X}$ (l.2), that is, they are observations of $\mathbf{X}$. Afterwards, the algorithm assigns to the Cluster Set j, C_j, the observations of $\mathbf{X}$ which has a smaller distance from the centroid j, μ_j (l.6-9) then the centroid is updated j based on the average of the observations contained in the Cluster set j, C_j (10.10-12). This process is repeated until the centroids at time t do not differ from the centroids at time $t-1$ by a tolerance constant ϵ, (l.2-13).

4. Results

The number of classes was defined by analyzing the Silhouette Score graph. A grid search was performed with k, in which $k \in \{2, ..., 10\}$, through the application of the k-Means algorithm whose results can be seen in Figure 1.

Algorithm 1: k-Means

Input: A observation matrix X and a scalar k $(\leq n)$
Output: K an agroupment of matrix X

$t = 0$;
Start randomly k centroids: $\mu_1, \cdots, \mu_k \in R$;
repeat
  $t \leftarrow t + 1$;
  $C_j \leftarrow \emptyset, \forall j = 1, \cdots, k$;
  for $\mathbf{x}_j \in X$ **do**
    $j^* \leftarrow argmin_i \{||\mathbf{x}_j - \mu_i^t||^2\}$
    $C_{j^*} \leftarrow C_{j^*} \cup \{x_j\}$
  end
  for $i = 1 \rightarrow k$ **do**
    $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
  end
until $\sum_{i=1}^{k} ||\mu_i^t - \mu_i^{t-1}||^2 \leq \epsilon$;

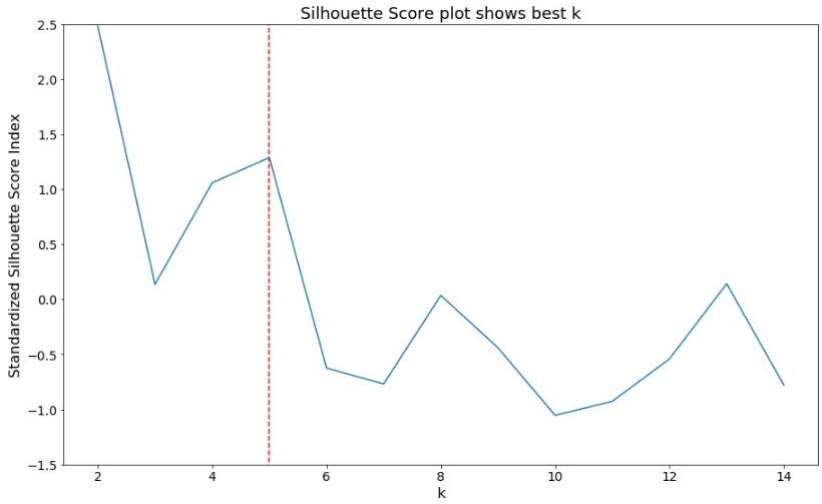

Figure 1. Silhouette Plot. In red, k = 5

It is important to note that although $k = 2$ by analyzing the graph was the most suitable k value ($max\ silhouette$), it would not provide us with a variety of clusters that would provide us with information, so we opted for the value of k = 5 which is a good trade-off between the number of clusters and inter-cluster heterogeneity. The Table shows counties and their associated labels

After choosing the k value, we perform the clustering and take the clusters. the bar chart of the venue count for each cluster is shown in Figure 2.

Analyzing each cluster by the types and quantities of venues found in each, we can interpret and relabel each cluster.

- Cluster 0: We have the presence of BBQ joint, squares and pizza places, and more social establishments, but with little density. We can characterize this cluster as

Table 5. Counties and cluster labels

Neighborhood	Cluster Label
Alcobaça	3
Aurelino Leal	4
⋮	⋮
Ubaitaba	4

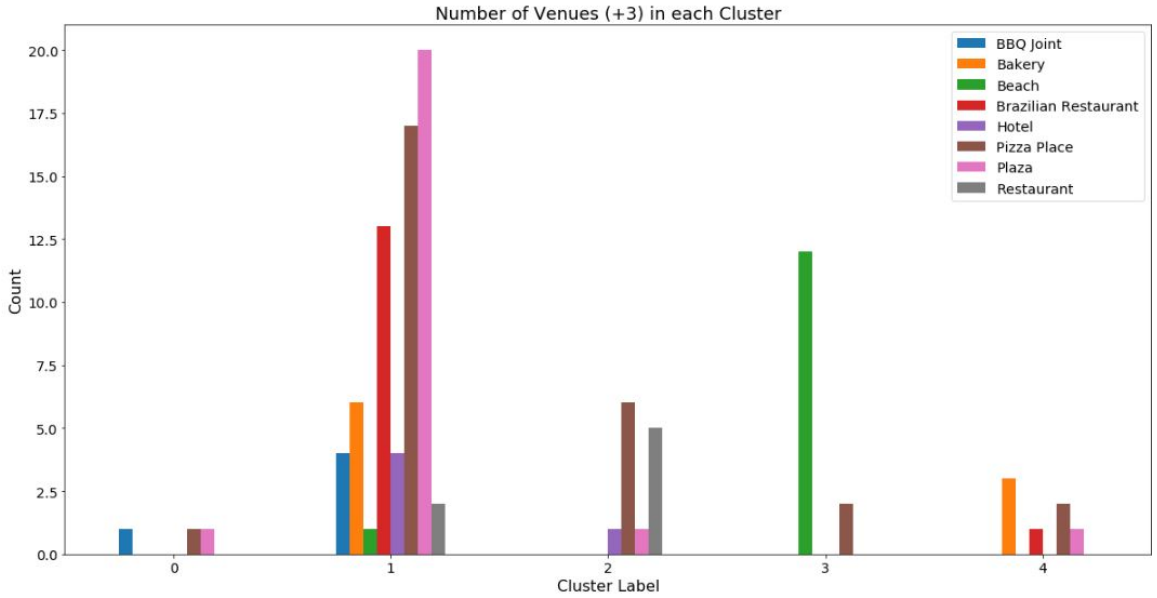

Figure 2. Bar chart showing the venues with minimun presence of 4 in each cluster.

"Multiple Social Venues (Low Density)".

- Cluster 1: Similar to cluster 0, we have the presence of several social establishments, such as most squares, pizza places, restaurants, steakhouses, etc. however, we have a high density of characteristic venues in this cluster, showing high restriction in the types of locations. We can characterize this cluster as **"Multiple Social Venues (High Density)"**.
- Cluster 2: We have the presence of hotels, social centers and various markets. We can characterize this cluster as **"Center Area"**.
- Cluster 3: Here curiously we have the presence of the coastal part of the bay, this can be better observed through the visualization on the map. We can characterize this cluster as **"Coastal Area"**.
- Cluster 4: we have several bakeries, restaurants, squares and pizza places, this combination is usually found in residential neighborhoods or sub centers. We can characterize this cluster as **"Residential Social Area"**.

One of my aims was also to show cluster venue information on the map. Thus, i grouped each borough by the specific cluster label and i combined that information in a Map showed in Figure 3. The colors chosen to identify each cluster are:

- Scarlet = Cluster 0: "Multiple Social Venues (Low Density)"

- Purple Heart = Cluster 1: ”Multiple Social Venues (High Density)”
- Cornflower Blue = Cluster 2: ”Center Area”
- Malibu = Cluster 3: ”Coastal Area”
- Aquamarine = Cluster 4: ”Residencial Social Venues”.

,

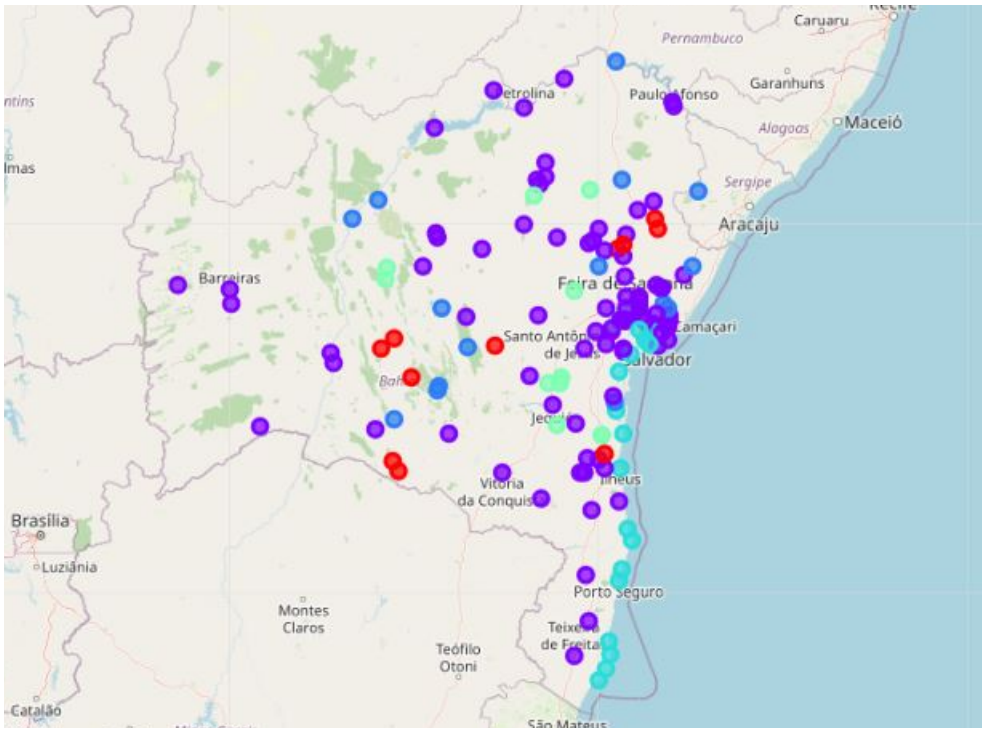

Figure 3. Map of Bahia state that contains the identification of the 138 counties with their respective clustering labels.

Note how areas of multiple social venues in low density (Scarlet) are on average further away from the tourist centers and Areas touristic. Tourist areas, with high density of social venue (Purple Heart), are close to the beach and close to the center. and tourist areas, with a high density of social venue, are on average closer to the beach and closer to the center.

5. Discussion

Bahia is a very big tourist with a high population density. The total number of measurements and population densities in the 417 counties can vary. Because of this, very different approaches can be tried in cluster and classification studies. Our approach is just one of a world of possibilities.

Regarding the k-Means clustering algorithm, when I tested the silhouette index method, I set the optimal value of k to 5. However, only 137 counties, a 67% reduction of the total, were used due to the lack of information recorded in the foursquare API. The data set can be expanded and the details of the neighborhood or street can also be detailed with the addition of data provided by other APIs.

The data was taken from the IBGE website [4] and stored in my Github. In future studies, this data can be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and grouping the information on the map of Bahia. In future studies, web or phone applications can be made to target stakeholders.

6. Conclusion

Tourism in the state of Bahia has made the region very receptive and with a high density of establishments dedicated to cooking and located in specific regions, such as coastal regions and surroundings.

Knowing where to start a business or simply choosing the neighborhood that best fits the family profile, can be a much more important milestone than just worrying about the size and physical characteristics of the property first.

Not only for investors, but city managers can also manage the city more regularly using similar types or platforms of data analysis.

References

Almeida, T. M. d., Moreau, A. M. S. d. S., Moreau, M. S., Pires, M. d. M., Fontes, E. d. O., and Góes, L. M. (2008). Reorganização socioeconômica no extremo sul da Bahia decorrente da introdução da cultura do eucalipto. *Sociedade Natureza*, 20:5 – 18.

da Silva, R. S. (2007). A ECONOMIA DO TURISMO NA BAHIA – FOCO NO SEGMENTO DOS CRUZEIROS MARÍTIMOS. *UNIFACS*, 11.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press.