Application of Agglomerative and Partitional Algorithms for the Study of the Phenomenon of the Collaborative Economy Within the Tourism Industry

Submitted: 20th December 2019; accepted: 30th March 2020

Juan Manuel Pérez-Rocha, Jorge Alberto Soria-Alcaraz, Rafael Guerrero-Rodriguez, Omar Jair Purata-Sifuentes, Andrés Espinal, Marco Aurelio Sotelo-Figueroa

DOI: 10.14313/JAMRIS/1-2020/10

Abstract: *This research discusses the application of two different clustering algorithms (agglomerative and partitional) to a set of data derived from the phenomenon of the collaborative economy in the tourism industry known as Airbnb. In order to analyze this phenomenon, the algorithms are known as "hierarchical Tree" and "K-Means" were used with the objective of gaining a better understanding of the spatial configuration and current functioning of this complimentary lodging offer. The city of Guanajuato, Mexico was selected as the case for convenience purposes and the main touristic attractions were used as parameters to conduct the analysis. Cluster techniques were applied to both algorithms and the results were statistically compared.*

Keywords: *Clustering Tools, Tourism Industry, collaborative economy*

1. Introduction

Collaborative Economy [1, 2] is an important phenomenon seen in countries that have a high use of social network platforms. This type of economy can be seen as a marketplace where consumers rely on each other instead of large companies to meet their wants and needs [3, 4]. Examples of collaborative economy sites are Etsy a general marketplace of art crafts, Uber a marketplace of peer-to-peer ridesharing site and AirBnB a marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. [4, 5].

Guanajuato is a city and municipality in central Mexico. Tourism is one of the main activities in the city because of the Spanish colonial past evidenced in its splendid architecture, with a population if 171,709 habitants, Guanajuato can be considered as a small city. The main touristic experience of Guanajuato is the Cervantino International Festival; an annual cultural event, which sponsors a large number of artistic and cultural events with artists invited from all over the world. The median number of national and international visitors to this festival has been estimated as 450,000 in 2018 [6].

Despite the number of formal lodging offers, a large number of visitors prefer to use a collaborative economy option to address his visit to a place [7]. Services like Airbnb and Booking.com are commonly used to meet the demand for logging space during the Cervantino International Festival.

The new host collaborative economies are growing up without knowing their geographic distribution and their relations with the touristic elements of the city [8].

It is necessary to know the geographic distribution of the Airbnb hosts to understand the offers dynamic and some characteristics of this kind of service [4, 5].

In order to understand this dynamic several agglomerative and partitional algorithms such as K-means [9, 10] and AGNES [11, 12] are applied to actual data of available hosts to identify the main clusters in collaborative economic offer in lodgings based on geographic distribution.

This paper is distributed as follows: Section 2 details relevant concepts as well as agglomerative and partitional algorithms. Our proposal is detailed in Section 3. Section 4 shows our experiments. Finally, Section 5 details our conclusions as well as Future work.

2. Relevant Concepts

In this section, the tools and relevant concepts used in our proposal are detailed; From agglomerative and partitional algorithms to implementation on actual data.

2.1. K-Means

K-Means [9, 10] is a method of vector quantization popular in data mining. This method aims to partition n observations into k clusters in which each observation belongs to a cluster with the nearest mean (according to a distance rule) serving as a prototype of the cluster. This results in a partitioning of the data.

Given a set of observations (x_1, x_2, x_3, ..., x_n), where each observation is a d-dimensional real vector, k-means clustering look for a partition of the observations into k sets $S = (S_1, S_2, S_3, ..., S_n)$ so as minimize the sum squares. Formally, the objective is to find:

$$\arg min \sum_{i=1}^{k} \sum_{x \in Si} \|x - \mu i\|^2 = \arg min \sum_{i=1}^{k} |Si| VarSi$$

Where μi is the mean point in S_i. This method is cataloged as an NP-Hard for 2 or more clusters [13].

2.2. Agglomerative Nesting

Agglomerative Nesting (AGNES) [11, 12] is a common type of hierarchical clustering used to group observations in clusters based on their similarity. This algorithm starts by considering each observation as a singleton cluster. Next, the clusters who are closest to each other (according to a distance metric or rule) are merged into a new cluster of 2 elements, this process stops until all observations are contained into a single cluster. The result is a tree-based representation of the objects named dendrogram. This algorithm can be stopped also when a specific number of clusters are met.

2.3. Collaborative Economy and Tourism

Online collaborative economy lodging platforms such AirBnB are part of a growing movement in eCommerce which uses advanced technology platforms to enable new operators to compete with traditional lodging providers like hotels and resorts to meet the demand of a tourism accommodation sector. Online Platforms like AirBnB enables individuals to compete with hotel operators without major overhead or investment by connecting ordinary people who have homes or rooms to rent with tourists in ways previously not possible [14]. Online collaborative economy's pervasive marketing extends the potential reach of the sector far beyond that of traditional holiday rental homes and enables several new forms of accommodation. First, individuals can rent out a spare bed in a living area or room within their own house or apartment, remaining present during the visit. Second, people might list their homes for rent while they are away. Third, owners of holiday houses might make their property available for others when not in use. Finally, investors might use Online collaborative economy platforms to market homes that are solely reserved for short-term tourism accommodations [5].

2.4. NbClustering

NbClustering [15] is an R clustering tool that provides near 30 indices doe determining the number of clusters and proposes to use the best clustering scheme for different results obtained by varying all combinations of clusters, distance measures and grouping methods.

The main Distance metrics used by this tool are:

Euclidean distance. Square distance between two given real vectors eq.1.

$$d(x,y)=\left(\sum_{j=1}^{d}(x_j-y_j)^2\right)^{1/2} \quad (1)$$

Maximum distance. Maximum distance between two components of **x** and **y** (supreme norm) eq.2.

$$d(x,y)=\sup_{1\le j\le d}|x_j-y_j| \quad (2)$$

Manhattan distance. Absolute distance between two vectors eq. 3.

$$d(x,y)=\sum_{j=1}^{d}|x_j-y_j| \quad (3)$$

Canberra distance. Terms with zero in numerator and denominator are omitted to form the sum eq. 4.

$$d(x,y)=\sum_{j=1}^{d}\frac{|x_j-y_j|}{|x_j|+|y_j|} \quad (4)$$

Minkowski distance. The **p** norm, the **p**th root of the sum of the **p**th powers of the differences of the components eq.5.

$$d(x,y)=\left(\sum_{j=1}^{d}|x_j-y_j|^p\right)^{1/p} \quad (5)$$

3. Methodology

In this section, our proposal of an Agglomerative and partitional data analysis applied to the Collaborative economy tourism data in Guanajuato city is detailed. Guanajuato city is shown in Figure 1.

Fig. 1. Guanajuato City Map

The data obtained from each host was longitude, latitude, price, and capacity. Those data were obtained using a Web Scraping [16] technique. Fig. 2 shows the hosts on the Guanajuato City map using the longitude and latitude.

Guanajuato touristic attractors were taken from the state-art [17]. Each attractor is composed by longitude, latitude, and the description. Fig. 3 shows the attractors on the Guanajuato City map using the longitude and latitude.

It was necessary to determine the cluster number to split the data, we use NbClustering to obtain the cluster number. We applied the K-means and AGNES algorithms to the hosts with the cluster number as a parameter for each one.

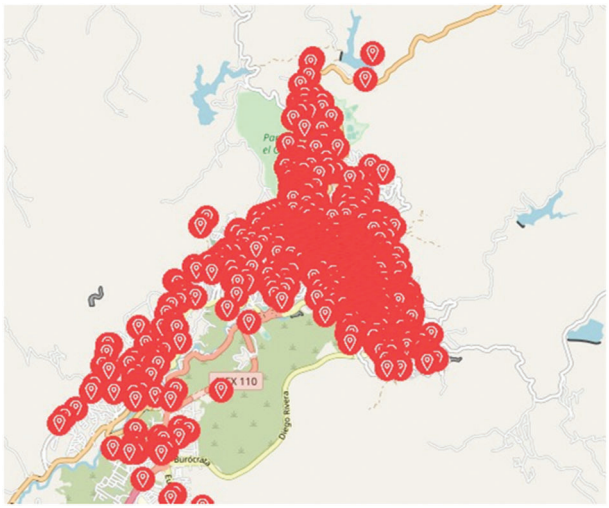

Fig. 2. Host geolocation obtained by Web Scraping tool

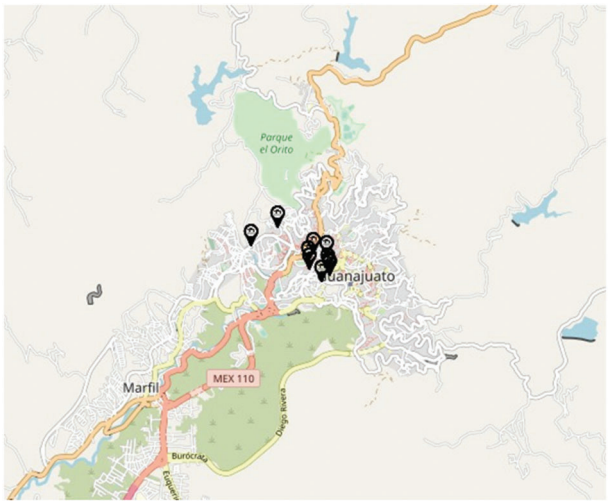

Fig. 3. Guanajuato touristic attractors

It was performed an experimental setup based on the host data combination. The results were contrasted by the cluster number generated by the k-means and AGNES algorithm.

4. Experiments

In this section, the used experiment set-up is detailed as well as the configuration of the generated.

4.1. Dataset Configuration

We use a dataset generated by the available data of lodging offers in Guanajuato City. It contains 1190 hosts obtained by web scrapping technique with the following characteristics:

- Each host has a geographic location price and capacity.
- 10 touristic attractors of Guanajuato City were taken from touristic state of art.

4.2. Experiment Configuration

In this work two clustering techniques were used; k-means based on partitioning clustering and AGNES from hierarchical approaches. Four experiments were conducted with the following configuration:

- Geographic location.
- Geographic location and capacity.
- Geographic location and price.
- Using all attributes.

4.3. Geographic Location Experiment

We applied the NbClustering tool to the host map in Guanajuato City considering only geographically data. From 24 metrics we gather the next data when using K-means as clustering tool:

- 12 experiments proposed 2 as the best number of clusters.
- 4 experiments proposed 3 as the best number of clusters.
- 2 experiments proposed 4 as the best number of clusters.
- 4 experiments proposed 7 as the best number of clusters.

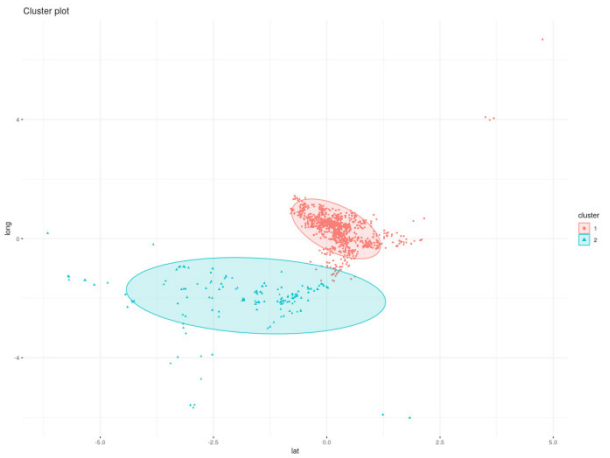

Fig. 4. Hosts grouped by geographically data with k-means and k=2

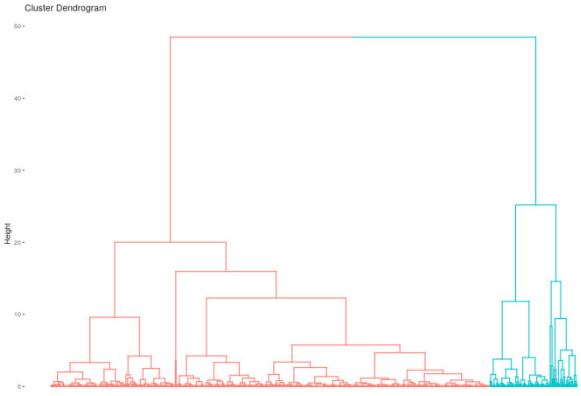

Fig. 5. Hosts grouped by geographically data with AGNES and two clusters

– 2 experiments proposed 10 as the best number of clusters.

According to the majority rule, the best number of clusters is 2. Figure 4 shows the application of K-means with three clusters over the host data in Guanajuato City. For AGNES hierarchical tree we gather the following data among all indices available in NbClustering tool:

– 6 experiments proposed 2 as the best number of clusters.
– 4 experiments proposed 3 as the best number of clusters.
– 6 experiments proposed 11 as the best number of clusters.
– 3 experiments proposed 14 as the best number of clusters.
– 1 experiment proposed 15 as the best number of clusters.

Consistently, the best number of clusters found by NbClustering was 2. Figure 5 shows the dendrogram produced by AGNES when using 2 clusters.

4.4. Geographic Location and Capacity Experiment

We applied the NbClustering tool to the host map in Guanajuato City considering latitude, longitude capacity for each host. From 24 metrics we gather the next data when using K-means as clustering tool:

– 7 experiments proposed 2 as the best number of clusters.
– 12 experiments proposed 3 as the best number of clusters.
– 2 experiments proposed 5 as the best number of clusters.
– 1 experiment proposed 6 as the best number of clusters.
– 2 experiments proposed 10 as the best number of clusters.

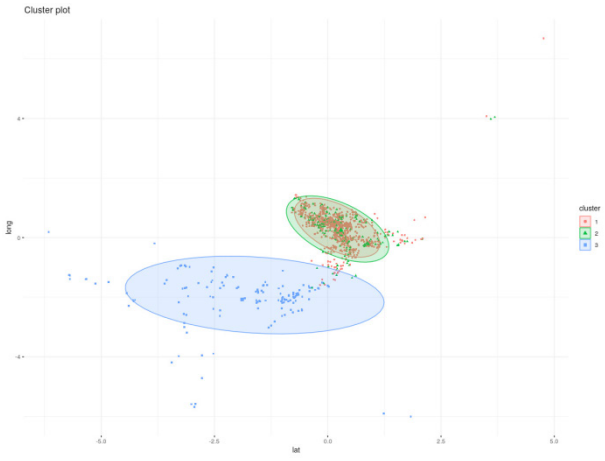

Fig. 6. Hosts grouped by latitude, longitude and capacity with k-means and k=3

According to the majority rule, the best number of clusters is 3. Figure 6 shows the application of K-means with two clusters over the host data in Guanajuato City. For AGNES hierarchical tree we gather the following data among all indices available in NbClustering tool

– 5 experiments proposed 2 as the best number of clusters.
– 11 experiments proposed 3 as the best number of clusters.
– 2 experiments proposed 4 as the best number of clusters.
– 1 experiment proposed 5 as the best number of clusters.
– 1 experiment proposed 7 as the best number of clusters.

Consistently, the best number of clusters found by NbClustering was 3. Figure 7 shows the dendrogram produced by AGNES when using 3 clusters.

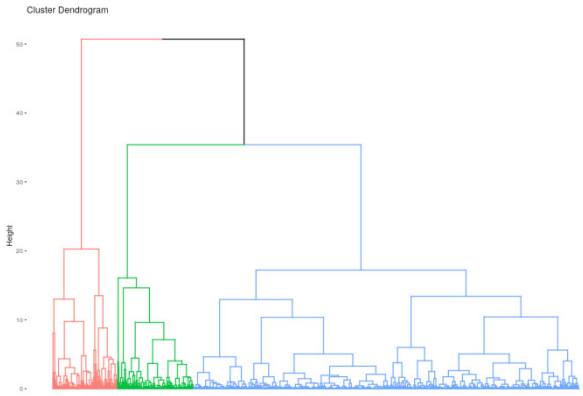

Fig. 7. Hosts grouped by latitude, longitude and capacity with AGNES and clusters=3

4.5. Geographic Location and Price Experiment

We applied the NbClustering tool to the host map in Guanajuato City considering latitude, longitude price for each host. From 24 metrics we gather the next data when using K-means as clustering tool:

– 8 experiments proposed 2 as the best number of clusters.
– 9 experiments proposed 3 as the best number of clusters.
– 2 experiments proposed 6 as the best number of clusters.
– 1 experiment proposed 8 as the best number of clusters.
– 2 experiments proposed 9 as the best number of clusters.

According to the majority rule, the best number of clusters is 3. Figure 8 shows the application of K-means with three clusters over the host data in Guanajuato City. For AGNES hierarchical tree we gather the following data among all indices available in NbClustering tool

– 8 experiments proposed 2 as the best number of clusters.
– 8 experiments proposed 3 as the best number of clusters.
– 2 experiments proposed 4 as the best number of clusters.
– 1 experiment proposed 5 as the best number of clusters.

Fig. 8. Hosts grouped by latitude, longitude and price with k-means and k=3

– 2 experiments proposed 11 as the best number of clusters.

Consistently, the best number of clusters found by NbClustering was 2. Figure 9 shows the dendrogram produced by AGNES when using 2 clusters.

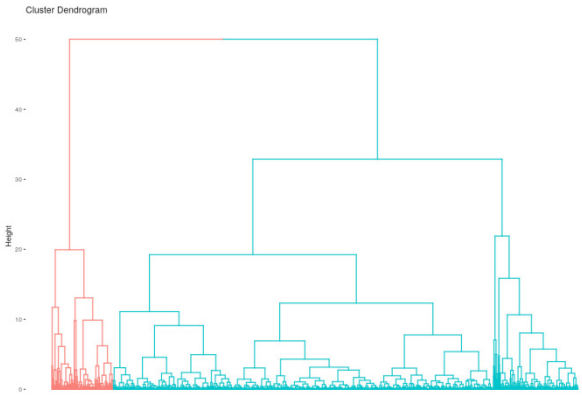

Fig. 9. Hosts grouped by latitude, longitude and price with AGNES and clusters=2

4.6. Geographic Location, Capacity and Price Experiment

We applied the NbClustering tool to the host map in Guanajuato City considering all variables latitude, longitude, capacity and price for each host. From 24 metrics we gather the next data when using K-means as clustering tool:

– 4 experiments proposed 2 as the best number of clusters.
– 14 experiments proposed 3 as the best number of clusters.
– 2 experiments proposed 7 as the best number of clusters.
– 1 experiment proposed 8 as the best number of clusters.
– 1 experiment proposed 9 as the best number of clusters.

According to the majority rule, the best number of clusters is 3. Figure 10 shows the application of K-means with three clusters over the host data in Guanajuato City. For AGNES hierarchical tree we gather the following data among all indices available in NbClustering tool

– 6 experiments proposed 2 as the best number of clusters.
– 8 experiments proposed 3 as the best number of clusters.
– 4 experiments proposed 4 as the best number of clusters.
– 1 experiment proposed 5 as the best number of clusters.
– 1 experiment proposed 6 as the best number of clusters.

Consistently, the best number of clusters found by NbClustering was 3. Figure 11 shows the dendrogram produced by AGNES when using 3 clusters.

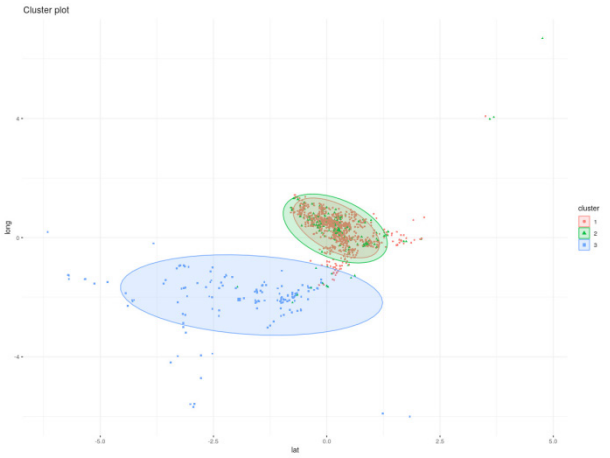

Fig. 10. Hosts grouped by latitude, longitude, capacity and price with k-means and k=3

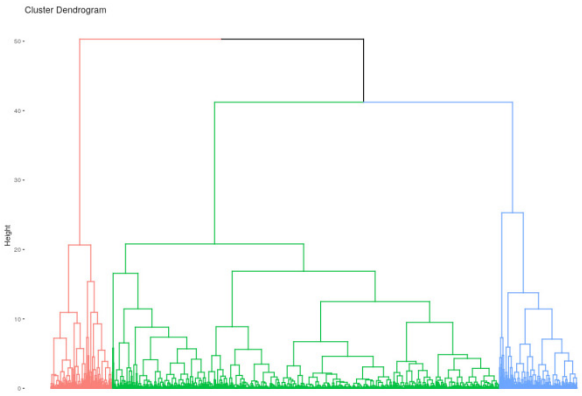

Fig. 11. Hosts grouped by latitude, longitude, capacity and price with AGNES and clusters=2

5. Conclusion

We have applied two agglomerative and partitional techniques to perform a study of the phenomenon of the collaborative economy within the tourism industry in Guanajuato City using real-world data. This data was obtained through Web Scraping to build a dataset of lodging hosts. This dataset contains the latitude, longitude capacity, and price of each available host.

Two clustering techniques were applied; K-means and AGNES with the number of groups determinate

by the NbClustering tool. With this data, the best number of groups found by this analysis were 2 and 3.

As future work can be made comparative between clusters using cluster statistics like hosts distance, the average price, the average rooms, the distance between clusters and other impact values to the tourism.

ACKNOWLEDGEMENTS

The authors want to Guanajuato University for the support provided for this research.

AUTHORS

Juan Manuel Pérez-Rocha – Administrative Information Systems, DCEA-University of Guanajuato, Guanajuato, México, e-mail: jm.perezrocha@ugto.mx.

Jorge Alberto Soria-Alcaraz – Organizational Studies Department, DCEA-University of Guanajuato, Guanajuato, México, e-mail: jorge.soria@ugto.mx.

Rafael Guerrero-Rodríguez – Management and Business Management Department, DCEA-University of Guanajuato, Guanajuato, México,
e-mail: r.guerrero@ugto.mx.

Omar Jair Purata-Sifuentes – Organizational Studies Department, DCEA-University of Guanajuato, Guanajuato, México, e-mail: opurata@yahoo.com.

Andrés Espinal – Organizational Studies Department, DCEA-University of Guanajuato, Guanajuato, México, e-mail: aespinal@ugto.mx.

Marco Aurelio Sotelo-Figueroa* – Organizational Studies Department, DCEA-University of Guanajuato, Guanajuato, México, e-mail: masotelo@ugto.mx.

* Corresponding author

REFERENCES

[1] D. Dredge and S. Gyimóthy, "The collaborative economy and tourism: Critical perspectives, questionable claims and silenced voices", *Tourism Recreation Research*, vol. 40, no. 3, 2015, 286–302, DOI: 10.1080/02508281.2015.1086076.

[2] D. Dredge and S. Gyimóthy, "Collaborative Economy and Tourism", 2017, 1–12, DOI: 10.1007/978-3-319-51799-5_1.

[3] "The Collaborative Economy". J. Owyang, C. Tran and C. Silva, https://www.slideshare.net/altimeter/the-collaborative-economy. Accessed on: 2020-05-28.

[4] D. Dredge and Sz. Gyimóthy, "The collaborative economy and tourism: Critical perspectives, questionable claims and silenced voices", *Tourism Recreation Research*, vol. 40, no. 3, 2015, 286–302, DOI: 10.1080/02508281.2015.1086076.

[5] N. Gurran and P. Phibbs, "When Tourists Move In: How Should Urban Planners Respond to Airbnb?", *Journal of the American Planning Association*, vol. 83, no. 1, 2017, 80–92, DOI: 10.1080/01944363.2016.1249011.

[6] Annual Activities Report; Secretaria de Turismo del Estado de Guanajuato, SECTUR. 2018. https://sectur.guanajuato.gob.mx. Accessed on: 2020-06-24.

[7] G. Zervas, D. Proserpio and J. W. Byers, "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry", *Journal of Marketing Research*, vol. 54, no. 5, 2017, 687–705, DOI: 10.1509/jmr.15.0204.

[8] D. Guttentag, S. Smith, L. Potwarka and M. Havitz, "Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study", *Journal of Travel Research*, vol. 57, no. 3, 2018, 342–359, DOI: 10.1177/0047287517696980.

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, 1967, 281-297.

[10] A. K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, 2010, 651–666, DOI: 10.1016/j.patrec.2009.09.011.

[11] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods", *Journal of Classification*, vol. 1, no. 1, 1984, 7–24, DOI: 10.1007/BF01890115.

[12] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou and A. Song, "Efficient agglomerative hierarchical clustering", *Expert Systems with Applications*, vol. 42, no. 5, 2015, 2785–2797.

[13] M. Garey, D. Johnson and H. Witsenhausen, "The complexity of the generalized Lloyd – Max problem (Corresp.)", *IEEE Transactions on Information Theory*, vol. 28, no. 2, 1982, 255–256, DOI: 10.1109/TIT.1982.1056488.

[14] D. Guttentag, "Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector", *Current Issues in Tourism*, vol. 18, no. 12, 2015, 1192–1217, DOI: 10.1080/13683500.2013.827159.

[15] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set", *Journal of Statistical Software*, vol. 61, no. 6, 2014, DOI: 10.18637/jss.v061.i06.

[16] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato and F. Fdez-Riverola, "Web scraping technologies in an API world", *Briefings in Bioinformatics*, vol. 15, no. 5, 2014, 788–797, DOI: 10.1093/bib/bbt026.

[17] R. Guerrero Rodríguez, "Estudiando la relación del turismo con el desarrollo humano en destinos turísticos mexicanos", *Acta Universitaria*, vol. 28, 2018, 01–06, DOI: 10.15174/au.2018.1886.