Profiling visitors of a national park in Italy through unsupervised classification of mixed data

Giulia Caruso, Adelia Evangelista, Stefano Antonio Gattone

1. Introduction

The success of a tourism destination, among other things, relies on the implementation of a strategic marketing plan. Since the identification and understanding of customers features and needs are essential for a correct market segmentation, the use of inappropriate techniques could result in missing strategic marketing opportunities (Bloom, 2004, Thompson & Schofield, 2009). Furthermore, any subsequent marketing activity would incur the risk to disappoint customers' expectations, producing their dissatisfaction. Moreover, the segmentation of markets based on visitor features and their motivations enables the identification of strengths and opportunities of a market (Lee & Lee, 2001).

The main benefit of market segmentation lies in knowledge acquisition. Profiling visitor allows to identify current consumers travel behaviour and to forecast future ones (Suleiman & Mohamed, 2011), enabling to acquire a competitive advantage (Hsu & Kang, 2003; Bui & Le, 2016, Koshy et al, 2019).

The aim of our study is to determine visitors characteristics and their satisfaction toward facilities of the National Park of Majella, in Italy. The outcome of our analysis is expected to serve as a guide for tourism operators, in order to facilitate plans toward formulating robust marketing strategies aimed to enhance visitors satisfaction. Our data have been collected on-site, from a sample of park visitors, and include both continuous and categorical features. In order to cluster such kind of data, we used an unsupervised classification method, specific for mixed data.

The paper is articulated as follows: in Section 2 we explain our data and consider the main clustering approaches for mixed variables, whereas in Section 3 we show the results obtained by the application of these methods to our dataset, providing an evaluation of the clustering results, by means of internal and external validity indexes. Finally, in Section 4, we draw some conclusions and discuss some suggestions for future research.

2. Data and method

Our dataset results from a questionnaire which has been collected on-site, from a sample of visitors of the Park, during the period from July 16 until October 27, 2020. A total of 523 tourists has been interviewed.

The Majella National Park is in Abruzzo, central Italy, and incorporates the provinces of Chieti, L'Aquila and Pescara, including 39 municipalities, characterized by a high spatial heterogeneity. This natural area is crucial for the protection of the natural ecosystem and for the socio-economic development of the area.

These data allow to perform a qualitative analysis on visitors of the Majella National Park, and consequently to assess their satisfaction level on the Park services.

The variables analysed are 16 (9 numerical - 7 categorical) and the entries are 523. The numerical variables concern the visitors perceived quality (measured in a 5 point Likert scale) on the following aspects: the web site, the naturalistic heritage conservation, the adequate presence of signage, of public transport, of children amenities, of footpaths maintenance, of accommodation facilities, of restaurant services and of food and wine products. The qualitative variables, instead, involve the following variables:

Giulia Caruso, Gabriele d'Annunzio University, Italy, giulia.caruso@unich.it, 0000-0003-0236-6201
Adelia Evangelista, Gabriele d'Annunzio University, Italy, adelia.evangelista@unich.it
Stefano Antonio Gattone, Gabriele d'Annunzio University, Italy, antonio.gattone@unich.it, 0000-0002-6143-9012

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giulia Caruso, Adelia Evangelista, Stefano Antonio Gattone, *Profiling visitors of a national park in Italy through unsupervised classification of mixed data*, pp. 135-140, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-304-8.27, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-304-8 (PDF), DOI 10.36253/978-88-5518-304-8

customers' expectations, the aim of their trip, the chosen location and how they came to its knowledge, the number of overnight stays, the type of chosen accommodation and, finally, the daily average expenditure per person.

In literature, most clustering approaches are limited to numerical or categorical data only. The traditional approach, instead, when dealing with both quantitative and qualitative variables, is to convert the latter values into numerical ones, and then apply the quantitative value based clustering methods (Foss et al, 2016; Ichino et al, 1994, Caruso et al, 2018). However, this approach would ignore the similarity information enclosed in the qualitative attributes, producing a loss of knowledge (Ahmad, A. & Dey, L. 2007). Finding a unified similarity metric for both kind of data, instead, would allow to remove the metric gap between them. Therefore, in order to detect different clusters, we compared two of the most used mixed data clustering methods, namely, the methods of Huang (Huang, Z., 1997) and Cheung & Jia (Cheung, Y. & Jia, H., 2013).

For sake of brevity, we will not describe in detail the methods we adopted to analyse the variables; the reader may consult our previous works for details (Caruso et al 2018-2019).

3. Results

We implemented a cluster analysis with a number of clusters equal to 3. Table 1 displays, for each cluster, the mean value of the 9 quantitative attributes analyzed and shows that the patterns produced by the two performed methods, specific for mixed data, are quite similar among them. The Huang one, in particular, highlights a slightly stronger clustering structure, meaning that the dissimilarity between clusters is higher.

Method	Cluster	Size	Signage	Footpaths	Restaurant services	Food & wine products	Public transports	Accommodation facilities	Web site	Children amenities	Natural heritage conservation
Huang	1	246	3.43	3.46	4.16	4.25	2.71	4.11	3.74	3.52	4.06
	2	171	4.44	4.53	4.75	4.82	3.84	4.71	4.57	4.39	4.68
	3	106	2.53	2.44	3.17	3.38	1.96	3.02	2.95	2.52	2.87
Cheung	1	201	2.96	2.91	3.69	3.85	2.31	3.54	3.36	2.95	3.44
	2	161	3.49	3.54	4.22	4.28	2.77	4.18	3.80	3.71	4.11
	3	161	4.45	4.53	4.65	4.76	3.84	4.67	4.52	4.31	4.66

Table 1. Cluster mean values for quantitative variables.

Figures 1 and 2 show the boxplot of the variables "Signage" and "Footpaths" in each cluster. The visual analysis highlights different median values in each group. Similar behaviours have been observed for the remaining quantitative variables.

Table 2 reports the results for the variable "**overnight stays**". The mode of the marginal distribution is represented by the value "1-3 nights stays" (42%). The clusters identified by the Cheung method are characterized with three different modes "1-3 nights stays" (Cluster 2), "4-7 nights stays" (Cluster 3) and "more than 7 nights stays" (Cluster 1).

The Huang method produced a slightly different result with two clusters out of three having mode "1-3 nights stays".

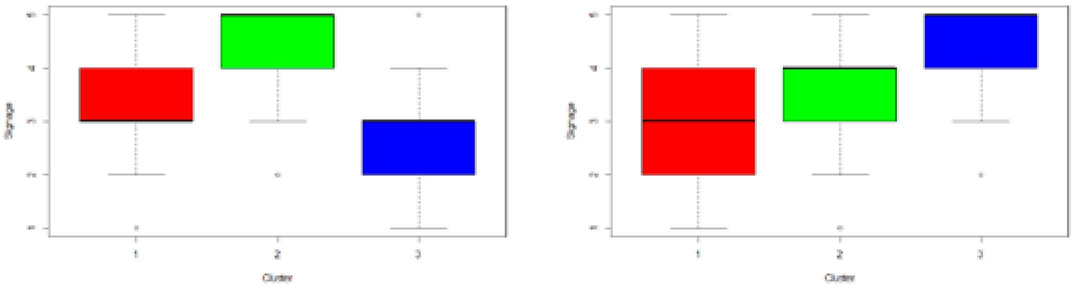

Figure 1. Quantitative variable Signage: boxplots for Huang (left panel) and Cheung (right panel) method.

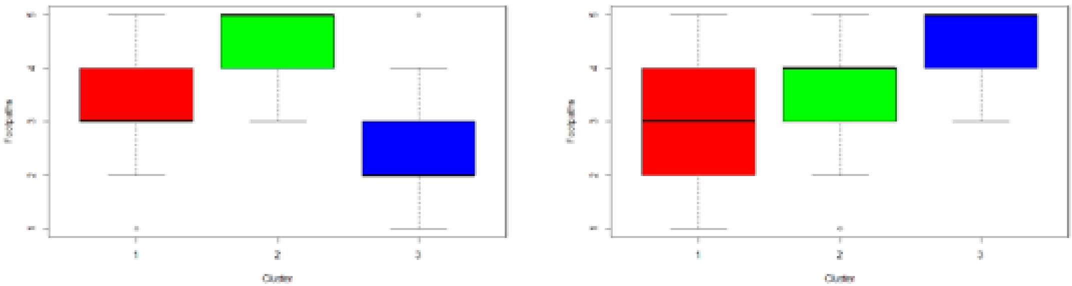

Figure 2. Quantitative variable Footpaths: boxplots for Huang (left panel) and Cheung (right panel) method.

OVERNIGHT STAYS	**Huang**			**Cheung**			
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**Marginal**
1-3 nights stays	**0.54**	**0.35**	0.24	0.23	**0.71**	0.35	**0.42**
4-7 nights stays	0.26	0.33	0.15	0.17	0.23	**0.40**	0.26
More than 7 nights stays	0.20	0.32	**0.61**	**0.60**	0.06	0.25	0.32

Table 2. Categorical variable Overnight stays: marginal and conditional distribution for each cluster.

A similar pattern can be observed with regards to the variable “Accommodation” (Table 3). The clusters identified by the Cheung method have different modes, i.e. “Other” (Cluster 3), “Second house” (Cluster 1) and “Hotel” (Cluster 2) while Clusters 2 and 3 of Huang have the same mode “Second house”.

ACCOMMODATION	**Huang**			**Cheung**			
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**Marginal**
Other	0.22	0.22	0.17	0.17	0.17	**0.29**	0.21
Rented apartment	0.07	0.11	0.14	0.13	0.02	0.12	0.09
B&B/ rented rooms	0.20	0.23	0.16	0.13	0.25	0.24	0.20
Second house	0.22	**0.30**	**0.43**	**0.53**	0.06	0.21	**0.29**
Hotel	**0.31**	0.13	0.09	0.03	**0.50**	0.14	0.21

Table 3. Categorical variable Accommodation: marginal and conditional distribution for each cluster.

With regards to the variable "Expenditure" (Table 4), the mode of the marginal distribution is represented by "10-30 Euros" (36%). The same result is observed in two out of three clusters for both methods.

EXPENDITURE	**Huang**			**Cheung**			
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**Marginal**
10-30 €	0.26	**0.46**	**0.42**	**0.42**	0.14	**0.50**	**0.36**
30-50 €	0.33	0.29	0.35	0.41	0.23	0.30	0.32
Less than 10€	0.06	0.08	0.05	0.06	0.04	0.08	0.06
More than 50€	**0.35**	0.17	0.18	0.10	**0.60**	0.11	0.26

Table 4. Categorical variable Expenditure: marginal and conditional distribution for each cluster.

With regards to the variable "Expectation" (Table 5), most tourists visited the park in order to take "guided tours for environmental education" (45%). This result is in line with all clusters produced by the Huang method and by two clusters obtained by the Cheung method.

EXPECTATION	**Huang**			**Cheung**			
Cluster	**1**	**2**	**3**	**1**	**2**	**3**	**Marginal**
Other	0.31	0.20	0.27	0.23	**0.43**	0.15	0.27
Flora observation	0.29	0.30	0.22	0.22	0.38	0.25	0.28
Guided tour for environmental education	**0.40**	**0.50**	**0.51**	**0.54**	0.19	**0.60**	**0.45**

Table 5. Categorical variable Expectation: marginal and conditional distribution for each cluster.

INTERNAL INDEXES	Huang	Cheung
CH	**189.4317**	106.4502
SHI	**0.1853808**	0.07521456
H	1.062639	**1.00835**

Table 6: Internal indexes for each method.

Synthetizing, by using the **Huang** method, cluster 1 differs from the others because it is characterized by tourists which stay in hotel, from 1 up to 3 nights, with an average daily expenditure of Euro 50,00. Cluster 2, instead, includes visitors which choice falls on B&B or rented rooms, for a period from 1 to 3 nights and which the average daily expenditure ranges from Euros 10 to 30. Visitors belonging to cluster 3, instead, choose their second house and they stay for more of 7 nights and with an average daily expenditure which ranges from 10 to 30 Euros.

When using the **Cheung** method, cluster 1 includes tourists which stay in their second houses, for more than 7 nights, and which daily expenditure ranges from 10 and 30 Euros. The aim of their visit is to take guided tours for the environmental education and their final goal is relaxation. Tourists inside cluster 2, instead, choose to stay in hotel, from 1 up to 3 nights, and they spend more than 50 Euros per day. Both in case of expectation and motivation they selected the option “other”. The tourists of cluster 3 choose an alternative kind of accommodation and they stays from 4 to 7 nights. Their daily expenditure goes from 10 to 30 Euros. Their expectation is to take guided visits for the environmental education and their aim is to relax.

Internal validity Indexes were computed in order to evaluate the quality of the cluster solutions. Results are shown in Table 6.

For numerical variables, the Calinski-Harabasz and the Silhouette Indexes are reported. Higher values correspond to better results; thus, the method of Huang is the one performing better when it comes to quantitative variables. With regards to the Internal Index for categorical variables, we used the Entropy Index. In this case a lower value of H corresponds to the best clustering result. The best (lowest) result for Entropy is obtained by using the Cheung method.

4. Conclusions

In order to detect clusters in a more efficient way, it is very useful to dispose also of qualitative variables. Our main aim was to observe the results of each method and to detect which one performs better. From our analysis it appears clearly that it corresponds to the Huang one as for the numerical variables, whereas the method of Cheung allows to obtain better results when it comes to qualitative ones.

Our objective for the future research is to develop new clustering analysis techniques for mixed data, which will consider an interesting insight provided by the work of Diday & Govaert, proposing an adaptive dynamic clustering procedure useful to calibrate the weights between qualitative and quantitative variables.

References

Basak, J., De, R., Pal, S. (1998). Unsupervised feature selection using a neuro-fuzzy approach, *Pattern Recognition Letters*, **19** (11), pp. 997-1006.

Bloom, J. Z. (2004). Tourist market segmentation with linear and non-linear techniques. *Tourism Management*, **25**, pp. 723–733.

Bui, H. T., Le, T. A. (2016). Tourist satisfaction and destination image of Vietnam's Ha Long Bay. *Asia Pacific Journal of Tourism Research*, **21**(7), pp. 795-810.

Caruso, G., Gattone, S.A., Balzanella, A., Di Battista, T. (2019). Cluster analysis: an application to a real mixed-type data set, in *Models and Theories in Social Systems. Studies in Systems, Decision and Control*, eds. C. Flaut, S. Hoskova-Mayerov´a and D. Flaut, vol. **179**, Springer, Heidelberg, pp. 525–533.

Caruso, G., Gattone, S.A., Fortuna, F., Di Battista, T. (2018). Cluster analysis as a decision-making tool: a methodological review, in *Decision Economics: In the Tradition of Herbert A. Simon's Heritage. Advances in Intelligent Systems and Computing*, eds. E. Bucciarelli, S. Chen, J.M. Corchado, vol. **618**, Springer, Heidelberg, pp. 48–55.

Cheung, Y., Jia, H. (2013). Categorical and numerical attribute data clustering based on a unified similarity metric without knowing cluster number, *Pattern Recognition*, **46** (8), pp. 2228-2238.

Diday, E., Govaert, G. (1997). Classification automatique avec distances adaptatives. R.A.I.R.O. *Informatique Computer Science,* **11** (4), pp. 329–349.

Foss A., Markatou, M., Ray, B., Heching, A. (2016). A semiparametric method for clustering mixed data, *Machine Learning*, **105**, pp. 419-458.

Hsu, C. H., Kang, S. K. (2003). Profiling asian and western family independent travelers (FITS): An exploratory study. *Asia Pacific Journal of Tourism Research*, **8** (1), pp. 58-71.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values, in: *Proceedings in the First Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Singapore: World Scientific, pp. 21–34.

Ichino, M., Yaguchi, H. (1994). Generalized minkowski metrics for mixed feature type data analysis, *IEEE Transactions on Systems, Man and Cybernetics,* **305** (24), pp. 698-708.

Lee, C. K., Lee, T. H. (2001). World culture EXPO segment characteristics. *Annals of Tourism Research*, **28** (3), pp. 812–816.

Nitanan Koshy, M. et al. (2019). Profiling the segments of visitors in adventure tourism: comparison between visitors by recreational sites. *International Journal of Business and Society*, **20** (3), pp. 1076-1095.

Suleiman, J. S., & Mohamed, B. (2011). Profiling visitors to Palestine: the case of Bethlehem city. *The Journal of Tourism and Peace Research,* **1** (2), pp. 41-52.

Thompson, K., Schofield, P. (2009). Segmenting and profiling visitors to the Ulaanbaatar Naadam festival by motivation. *Event Management*, **13**, pp. 1-15.

Yeung, D., Wang, X. (2002). Improving performance of similarity-based clustering by feature weight learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24** (4), pp. 556-561.