# Assessment of visitors' perceptions in protected areas through a model-based clustering

Annalina Sarra, Adelia Evangelista,Tonio Di Battista

## 1. Introduction

Protected areas are well-defined geographical spaces that, in view of their recognized, natural, ecological or cultural values, receive protection. They have the twofold mandate of protection of natural resources and cultural values and providing a space for nature-based tourism activities, including, among others, mountain hiking, nature photography, bird and animal watching. In the last years, the nature-based tourism is experiencing a positive and sustainable growth worldwide, making it an important sector of the tourism activity, with substantial impacts on the environment, economy and local communities. As broadly highlighted in literature, visitors' experience can be deemed a complex interaction between people and their internal states, the activity they are undertaking, and the social and natural environment in which they find themselves (Leung et al., 2008). Understanding the value attached by visitors to their destination and know their assessment on various activities in which they are engaged during their stay is a key element in shaping their satisfaction. A number of studies have shown that visitor's satisfaction is essential to boost demand, since it increases intention to revisit and recommend the destination to other people (see, among others,Sangpikul (2018); Su et al. (2016)). Besides, a greater knowledge of needs and perception of different visitor-groups should lead to improved management and marketing strategies and to more targeted provision of facilities. In this study, we focus on analyzing the perceived value of visitors who had a specific experience in the Majella National Park, located in the Abruzzo region (Italy). The research data were collected by means of a structured questionnaire administrated to people who visited the sites of the protected area during the last three summer months of 2020. A total of 151 valid questionnaires were obtained and form the base of the data analysis. Our aim is to assess the views and preferences of visitors of that protected natural space in relation to specific profiles. To this end, through a Bayesian model-based clustering, better known as Bayesian Profile Regression, we partitioned visitors into clusters, characterized by similar profiles in terms of their demographic characteristics (age, gender, education attainment, professional activity, origin area), as well as, in terms of the features of their travel behaviour (accommodation, length of stay, past visitation experience). The benefit of the followed approach lies in the ability of that Bayesian technique of simultaneously estimating the contribute of all covariates to the outcome of interest. In our context, we explore the association of detected groups with visitors' satisfaction. In the survey, the global quality of tourism service is segmented into single features and respondents were asked to give their level of appreciation on a five-point Likert satisfaction scale. To estimate the latent trait measured by the items and related to the overall satisfaction, we followed an IRT modelling. The rest of the paper is organized as follows. In Section 2 we describe the study area and the data collected. Section 3 is devoted to illustrate the methodology adopted. Finally, in Section 4 are presented the main results.

Annalina Sarra, Gabriele d'Annunzio University, Italy, annalina.sarra@unich.it, 0000-0002-0974-0799
Adelia Evangelista, Gabriele d'Annunzio University, Italy, adelia.evangelista@unich.it, 0000-0002-7596-9719
Tonio Di Battista, Gabriele d'Annunzio University, Italy, tonio.dibattista@unich.it, 0000-0003-2139-7273

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

Annalina Sarra, Adelia Evangelista,Tonio Di Battista, *Assessment of visitors' perceptions in protected areas through a model-based clustering*, pp. 245-250, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.46, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site 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 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

## 2. Study case and data

Majella National Park (MNP) is a protected area located in the provinces of Chieti, Pescara and L'Aquila, in the region of Abruzzo, Italy. It was established in 1991 and it extends over an area of about 74,000 hectares, comprising the mountains of Majella and Morrone. The mountains dominate the territory of this national park: as a matter of fact, the 55% of it is over the 2,000 meters. It includes wide lands with particular wilderness aspects, the rarest and most precious part of the biodiversity national heritage. The diversity of the environments, the richness of nature, the evidences left by the human presence make Majella protected area attractive for visitors, potentially involved in different activities, ranging from visits to hamlets and hermitages, climbing and trekking excursions to participation to traditional festivals. In this study, a sample of visitors was intercepted through a non-probabilistic design. At the end of the survey period, a total of 151 valid questionnaires were returned and form the basis of the data analysis reported herein. The questionnaire, other than to include a few questions with respect to respondents' background (age, gender, education, professional activity, origin area), is organized in two sections devoted to investigate different aspects. Part 1 controls for travel behaviour characteristics of respondents (accommodation type, length of stay, past visitation experience, daily average expenditure) and their expectations. Since visitors are increasingly demand high quality recreational opportunities and the service that support them, the second section of the questionnaire deals with the satisfaction. The satisfaction scale contains 23 items, corresponding to different aspects of tourism experience (staff, food, excursions, outdoor activities, accommodation, information services, naturalistic and historical heritages, hospitality of local population, safety and sustainability, sanitation). Respondents were asked to indicate the degree to which they agree with each item on a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In our sample, the majority of the visitors surveyed were men (55%), from outside region (68%), aged between 36 and 45 (40%). Half of them had a university education. Regarding their professional activity the respondents were mainly employees (29%). The main reasons that encourage people to visit the Majella National Park are relax and contact with nature, they are in fact curious to carry out guided tours, and they make this choice thanks to the testimony of their acquaintances (word of mouth). The tourist offices to which they contacted for information are principally those of the province of Pescara. Visitors are encouraged to return to the places of the Majella National Park; in fact about half of those interviewed have already been there more than 5 times, and there spent more days (at least 1-3 nights). However, the average daily expenditure that tourists expect to expend is between 10-30 euros.

## 3. Methodology

**3.1 IRT model for polytomous response data: the Graded Response Model** Item response theory (IRT), initially developed in the 1960s, comprises a versatile family of measurement models concerned with the measurement of an individual's latent trait assessed indirectly by a group of items (de Ayala, 2009). The basic idea behind IRT is that the structure in the manifest responses is explained by assuming the existence of one or more latent traits ($\theta$). The mathematical characteristics of IRT models allow a transformation from binary or ordinal answer pattern, e.g. Likert type data, into measure on an equal-interval scale. In this study, the parametric Graded Response Model (GRM; Samejima (1969)) was applied to analyze the ordered response categories of the satisfaction scale included in the questionnaire administered to tourists. In the GRM, items are described by a discrimination parameter ($\lambda$) and two or more location parameters ($\gamma$). Graded-response model item parameters are easily interpretable: the location parameters locate the point at the latent trait continuum where is a 50% chance of scoring at or above category $c_k$ of item $k$ whereas the discrimination parameter reflects the degree to which the item is related to the underlying latent trait and can differentiate among persons with different trait levels. Specifically, in the GRM the probability that a person's response falls at or above a particular ordered category $c$ ($c = 1, \dots, C_\kappa$), given the latent trait $\theta$, may be expressed as follows:

$$Pr(Y_{ij} = c|\lambda_k, \theta_i, \gamma_k) = \Phi(\lambda_k\theta_i - \gamma_{k,c-1}) - \Phi(\lambda_k\theta_i - \gamma_{k,c}) \tag{1}$$

Eq. 1 describe the normal ogive formulation for the unidimensional two-parameter GRM model, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution functions and $c$ represent the $k$ ordered categories. GRM cannot be identified because it is overparameterized: for each item a set of $k-1$ location parameters along k item slope ($\lambda$) are to be estimated. To overcome this issue some restrictions on parameters are necessary.

**3.2 Model-based clustering: the BPR** In this work, we focused on tracing the profile of the tourists who visited the Majella National Park, considering in the analysis, as covariates, the socio-demographic background, vacation habits, typically activities at destinations. To this end, we opted for a cluster-based method, better known as Bayesian Profile Regression (BPR)(Molitor et al., 2010). Profile regression is a Bayesian cluster method that, by capturing the heterogeneity among covariates, allows both identifying specific covariate profiles that are representative of a population (i.e. cluster) and associating them with the outcome of interest (in our case tourists'satisfaction) via a regression model. The Bayesian aspect of this clustering process has some advantages over traditional clustering approaches (e.g. Latent Class Analysis) in that the number of clusters has not to be fixed in advance but it is informed by the data and the model is fitted as a unit, allowing that an individual's outcome potentially influences cluster membership, so that the outcome and the clusters mutually inform each other. Additionally, BPR provides a unified procedure in which the uncertainty associated with clustering is naturally propagated into the regression model and incorporated into posterior inference via MCMC algorithms. Formally, for each individual $i = 1, \dots, N$, $Y_i$ denotes the outcome of interest and $\mathbf{X}_i = (X_{i_1}, \dots, X_{i_P})$ represents the covariate profile. The joint probability model for the outcome $Y_i$ and the predictors $\mathbf{X}_i$ can be written as an infinite mixture model (Molitor et al., 2010; Hastie et al., 2013)

$$f(Y_i, \mathbf{X_i}|\mathbf{\Theta}) = \sum_{c=1}^{\infty} \psi_c Pr(\boldsymbol{X}_i|\mathbf{\Theta}_c, \mathbf{\Theta}_0) Pr(Y_i|\mathbf{\Theta}_c, \mathbf{\Theta}_0). \tag{2}$$

The probability model in equation (2) consists of two sub-models:
the *profile sub-model*, $Pr(\mathbf{X}_i|Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Theta}_0)$, and the *response sub-model*, $Pr(Y_i|Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Theta}_0)$. For each mixture component, the probability models for the outcome $Y_i$ and the profile $\boldsymbol{X}_i$ are independent conditionally on some component specific parameters $\mathbf{\Theta}_c$ and some global parameters $\mathbf{\Theta}_0$. In the BPR approach, in order to make inference, an additional allocation parameter $Z_i$ is introduced such that $Z_i = c$ indicates that individual $i$ is assigned to component $c$ and $Pr(Z_i = c) = \psi_c$, with $\psi_c$ the mixture component weight. As pointed out in Molitor et al. (2010), it is possible to approximate the infinite mixture model with a finite one, by specifying a maximum number $C$ of components. The mixture weights, $\boldsymbol{\psi} = (\psi_1, \dots, \psi_C)$, are modelled according to a stick-breaking prior (Ishwaran and James, 2001).

Table 1: Two-parameter GRM estimates

| | **Discrimination parameter** | **Thresholds** | | | |
|---|---|---|---|---|---|
| ***Item*** | $\lambda_j$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_4$ |
| Cleaning | 2.20 | -1.90 | -1.19 | -0.37 | 0.53 |
| Signposting | 2.92 | -1.63 | -1.09 | -0.35 | 0.53 |
| Places accessibility | 3.14 | -1.83 | -1.11 | -0.41 | 0.56 |
| Web site | 3.52 | -1.81 | -1.07 | -0.36 | 0.46 |
| Park info | 3.92 | -2.13 | -1.10 | -0.31 | 0.33 |
| Guided tours | 3.90 | -1.71 | -1.06 | -0.29 | 0.40 |
| Public transport services | 2.27 | -1.09 | -0.15 | 0.42 | 1.29 |
| Cultural events | 2.25 | -2.12 | -1.00 | 0.03 | 1.01 |
| Catering quality | 1.85 | -3.08 | -1.90 | -0.95 | 0.28 |
| Food & wine products | 2.17 | -2.12 | -0.97 | -0.57 | 0.05 |
| Children services | 2.88 | -1.73 | -0.79 | 0.08 | 0.90 |
| Path maintenance | 3.38 | -1.66 | -0.98 | -0.12 | 0.54 |
| Accommodation facilities | 2.80 | -2.25 | -1.27 | -0.54 | 0.50 |
| Hospitality of local population | 2.07 | -2.84 | -1.89 | -0.89 | -0.01 |
| Naturalistic heritage | 4.08 | -1.86 | -1.31 | -0.59 | 0.38 |
| Historical cultural heritage | 3.33 | -1.82 | -1.31 | -0.43 | 0.56 |
| Environmental education center | 3.23 | -1.82 | -1.03 | -0.11 | 0.66 |
| Sanitation | 2.84 | -1.66 | -0.95 | -0.19 | 0.92 |
| Reception center | 5.09 | -2.05 | -1.25 | -0.56 | 0.26 |
| Promotion park activities | 4.15 | -1.69 | -1.17 | -0.43 | 0.39 |
| Park staff | 2.69 | -2.05 | -1.55 | -0.96 | -0.12 |
| Park staff's knowledge of foreign languages | 2.25 | -2.57 | -1.59 | -0.47 | 0.42 |
| Information material area | 2.92 | -2.31 | -1.61 | -0.91 | -0.07 |

## 4. Results and conclusion

In this section, we first present the parameter estimates obtained by fitting the GRM to the items of visitors' satisfaction scale. All data were analyzed in the R programming environment (R Core Team, 2020) with mirt package (Chalmers, 2012). Table 1 shows the estimates for discrimination and threshold parameters. Discrimination estimates for the items ranged from 1.85 to 5.09, indicating that all items discriminate well between low and high levels of satisfaction: higher values indicate better discrimination. Specifically, the inspection of discrimination parameter estimates suggests that the key indicators in distinguishing visitors with different satisfaction levels are those related to appropriate promotion and guided tours. Also, play a fundamental role in discriminating visitors scoring high and low on the latent satisfaction trait, the item ascertaining the appreciation of the park naturalistic heritage. Likewise, the items that had a smaller discriminative power are those linked to the quality of catering, food and wine products, the hospitality of local population, park staff's knowledge of foreign languages. As for threshold parameters, it is worth noting that they reflects the cut-points between the five item categories. Each of them, mirrors the probability of scoring above or below a given cut-point. In other terms, the thresholds can be thought of as being on the same scale as the z-scale, where a normal distribution is centered at zero with a unit standard deviation metric. By comparing thresholds values across items, we see that, for example, item related to "catering quality" has the lowest initial threshold value of -3.08 and item referring to "pubblic transport services" has the largest initial threshold value of -1.09. This result indicated that fewer people endorsed the first item response category for the item related to "catering quality" compared to the item concerning "pubblic transport services". After estimating the visitors' satisfaction by means of IRT modelling, the next step in our data analysis was the identification of specific visitor-groups. The BPR was fitted using the R package PReMiuM (Liverani et al., 2015). The BPR has pro-

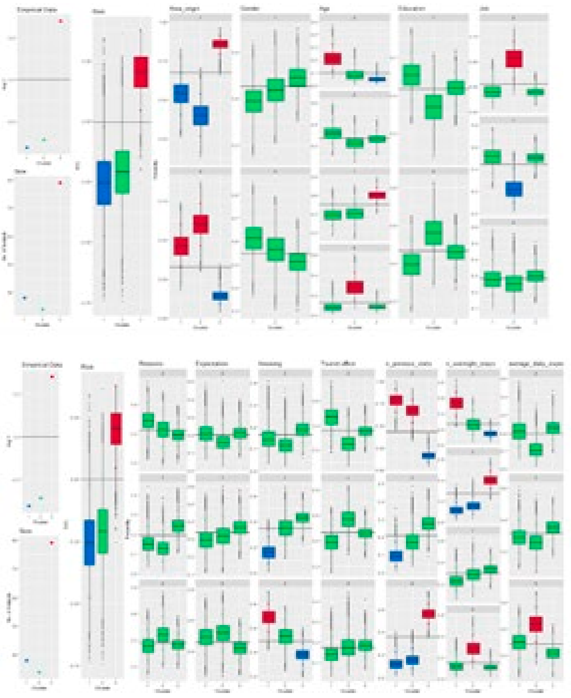

Figure 1: Characterization of visitors'satisfaction profiles associated to each cluster

**Legend: modalities code of each categorical variable included in the profile sub-model**

Area of origin (0= Abruzzo; 1= Other Regions); Gender (0= Male; 1= Female); Age (0= less than 25 years, 1= 25-45 years; 2= 46-65 years; 3= more than 65 years); Education (0= upper secondary school; 1= Degree); Job (0= Self-employed; 1= Public-employee; 2= Other professions); Reasons (0= Contact with nature; 1= Visit to the historical, artistic and cultural heritage; 2= Relax); Expectation (0=; 1=; 2=); Knowing (0= Personal reccomandation; 1=Park website; 2=Guidebook); Tourist office (0= Chieti; 1= Pescara; 2= L'Aquila); Number of previous visits (0= Once; 1= 2-5 times; 2= more than 5 times); Number of overnight stays (0= None; 1= 1-3 overnights; 2= 4-7 overnights; 3= more than 7 overnights); Average daily expenditures (0= less than 30€, 1= 30-50€; 2= more than 50€).

duced a partition of visitors into 3 clusters: each of them is characterized by similar covariates profile as well as the same satisfaction level. In order to delineate the visitors' specific profile we can refer to the characterization of each cluster in terms of covariates, as illustrated in Figure 1. On the left panel of each figure, the MCMC posterior draws of the satisfaction level of the identified clusters are given; conversely, for categorical variables such those considered in this study, the right panel of each figure shows the posterior distributions of the probability that an explanatory variable appears with one of the discrete categories across the identified groups. Note that each column corresponds to one covariate and cluster labels are specified on horizontal axis. The different colours of box-plots (blue, green and red) refer to a $95\%$ credible interval respectively under, within or upper the global average on all visitors (whatever the cluster). The order of cluster representation follows the order of the associated estimated visitors'satisfaction level of each cluster. A close analysis of Figure 1 reveals that in the typical profile of cluster associated with the highest satisfaction level there is a prevalence of visitors coming from other regions, aged 25-45 years, who had never been before in the Majella National Park area and who have decided to stay overnight from 4 to 7 days. On the other hand, among the visitors who exhibited lower level of appreciation of the natural area, we find a greater number of Abruzzo resident people, for whom the word of mouth has had a key role in their decision making process to choose that tourist destination. Additionally, both the number of previous visits (more than 5) and the overnight stays (more than seven) have contributed to negatively shape the visitors'satisfaction. The results of this study might have practical implications for managers of protected areas giving them useful insights on how elaborate programs according to visitors' profile. To our knowledge this research represents the first attempt of identifying clusters of visitors with similar covariate profiles through a Bayesian approach based on Dirichlet modeling mixture techniques. Along this benefit, it is important to stress the major limitation of this work concerning the selection of sample units, intercepted through a non-probabilistic design. As a result, we are not able to infer the actual visitors' flows over all the different seasons of the year.

## References

Leung, Y., Marion, J., Farrell,T. (2008). Recreation ecology in sustainable tourism and ecotourism: A strengthening role, in *Tourism, Recreation and Sustainability: Linking Culture and the Environment, 2nd Edition*.Wallingford, UK, pp. 19–37.

Sangpikul, A. (2018). The effects of travel experience dimensions on tourist satisfaction and destination loyalty: the case of an island destination, *International Journal of Culture, Tourism and Hospitality Research*, **12** (1), pp. 106–123.

Su, L., Swanson, S. R., Chen, X. (2016). The effects of perceived service quality on repurchase intentions and subjective well-being of Chinese tourists: The mediating role of relationship quality, *Tourism Management*, **52**, pp. 82–95.

de Ayala, R. J. (2009). *The theory and practice ofitem response theory*.New York: Guilford Press.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, **34** (17), pp. 100-114.

Molitor, J., Papathomas, M., Jerrett, M., Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health, *Biostatistics*, **11**, pp. 484–498.

Hastie, D., Liverani, S., Azizi, L., Richardson, S., Stücker, I. (2013). A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer, *BMC Medical Research Methodology*, **13** (129).

Ishwaran, H., James, L. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association*, **96** (453), pp. 161–173.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org/

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment, *Joournal of Statistical Software*, **48** (6), pp. 1–29.

Liverani, S., Hastie, D., Papathomas, M., Richardson, S. (2015). PReMiuM: An R Package for Profile Pegression Mixture Models Using Dirichlet Processes, *Journal of Statistical Software*, **64** (7), pp. 1–30.