COMBINING SUPERVISED AND UNSUPERVISED METHODS IN TOURISM VISITOR DATA

Weksi Budiaji[1], Vebriana[2], Juwarin Pancawati[3]

[1] Department of Agribusiness, University of Sultan Ageng Tirtayasa, Serang, 42128, Indonesia
[2] Student of Department of Agribusiness, University of Sultan Ageng Tirtayasa, Serang, 42128, Indonesia
budiaji@untirta.ac.id

***Abstract*--Combining supervised and unsupervised method can assist in the data analysis process. This research aims to apply a supervised method, i.e. Poisson regression, that is followed by an unsupervised method, namely cluster analysis of the visitors in a tourism dataset. The samples were taken 80 persons purposively from the visitors of the Flower Garden X in Serang Regency, Banten Province. The dataset consists of the number of visits, travel cost, income/ stipend per month, gender, age, distance from the place of origin, and perception, which is formed by 11 questions of facilities and services. The Poisson regression was applied in the 30, 40, and 50 bootstrap samples resulted in the perception as the significant features. Then, medoid-based cluster analysis, i.e. pam and simple k-medoids, in the perception dataset was applied. They compared simple matching and cooccurrence distances and were validated via medoid-based shadow value. It grouped the visitors into five clusters as the most suitable number of clusters. The combined methods of supervised and unsupervised provided the cleanliness as the important indicator. The improvement of the tourism object had to be focus on the cleanliness aspect.**

***Keywords*: Bootstrap samples; Cluster analysis; Poisson regression; Supervised method; Unsupervised method.**

I. INTRODUCTION

Two general methods in statistical learning are commonly distinguished into supervised and unsupervised methods. The supervised methods are indicated by a response feature, while it is absence in the unsupervised one [1]. Combining these two methods can be performed in three different schemes. The unsupervised method is performed as a pre-processing step before applying the supervised method in both nature and social researches [2–4]. The other way around where the supervised precedes the unsupervised is also feasible [5]. Moreover, a looping algorithm by repeating the supervised and unsupervised methods is presented in a detection of cancer dataset [6].

A visitor dataset, in which it includes count data, is usually analysed via a supervised method, namely a Poisson regression for example in both health care [7,8] and tourism [9,10] visitors. These two examples contrast in the covid-19 pandemic situation with respect to the number of visitors. While the health care suffers from over-visit of infected people, the tourism is depressed with a decreasing number of visitors.

This article focuses on the latter case where the tourism object is the topic of interest. It identifies the significant factors of tourism visitors based on the number of visits. It is analysed via a supervised method. Then, an unsupervised method is applied to characterize the visitors. The characterization of visitors is important to improve the tourism object performance.

II. METHODE

Dataset was obtained from a survey of agrotourism visitors. The samples were taken 80 persons purposively from the visitors of Taman Bunga X in Serang Regency, Banten Province. The visitors were asked about their number of visits, travel cost, income/ stipend per month, gender, age, distance from the place of origin, and perception. The perception feature was indicated by 11 questions about facilities and services of the agrotourism object. The 11 indicators namely (1) access to the agrotourism object, (2) the beauty of the scenery, (3) canteen facility, (4) cleanliness of the location, (5) completeness of the facility, (6) gazebo facility, (7) mosque facility, (8) parking

spot, (9) security of the location, (10) service quality, and (11) toilet facility. All of the indicators all measured in Likert scale with 7 points [11].

The dataset has a response feature, namely the number of visits. It is count data such that a Poisson regression as a supervised method is suitable. However, an equi-dispersion assumption has to be satisfied, or a parameter of dispersion can be introduced otherwise [12]. Then, the important features from the Poisson regression are explored via cluster analysis as an unsupervised method. The more detail steps follow:

1) Data is reduced by sampling as many as 30, 40, and 50 objects randomly. Then, the reduced dataset is modelled by a Poisson regression. Suppose y indicates a random variable of the frequency of visit in a specified time interval, the distribution of y is Poisson distribution with $\mu > 0$, if
$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}, y = 0, 1, 2, \ldots$$
The mean and variance of the Poisson distribution is equal, E(y) = V(y) = μ/ ø. The expected value of the log link inverse function is $\mu = \exp(\mathbf{x}\beta)$, where **x** is the feature and β is the estimated regression coefficient. Then, the Poisson regression is modelled by
$$\mu = \exp\ (\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6),$$
where the features are travel cost, income or stipend per month, gender, age, distance from the place of origin, and perception of facilities and services that are indicated by X_1, X_2, X_3, X_4, X_5, X_6, respectively. When an equal dispersion assumption is violated i.e., an over/ under dispersion occurs, a quasi-Poisson is applied [12]. The difference between Poisson regression and quasi-Poisson regression is the dispersion parameter. The former fixes the dispersion parameter at 1 (ø = 1) and the variance function (V(y)) equals to μ, while the latter estimates dispersion (ø) from the empirical data (V(y) = øμ).
2) The modelling of the reduced dataset is repeated for 1000 times with different samples. It is a bootstrap procedure [13].
3) For each bootstrap sample, the significant features are recorded.
4) Cluster analysis is applied in the full dataset with the significant features from step 3. The dataset consists of mixed features so that medoid-based algorithm is preferable such as partitioning around medoids (pam) [14] and simple k-medoids [15].
5) The cluster result is validated by an internal criteria namely medoid-based shadow value [16]. The number of clusters is opted in this step. Medoid-based shadow value is calculated by
$$msv\ (x) = \frac{\delta\big(x, m'(x)\big) - \delta\big(x, m(x)\big)}{\delta\big(x, m'(x)\big)}$$
where δ(x,m(x)) and δ(x,m'(x)) are the distance of object to the first and second closest medoids.
6) The cluster visualization is plotted to assist the interpretation.

The analysis was run in an Intel i5 8GB RAM using software R [17], by which the run time of three bootstrap schemes in 1000 replicates was eight seconds. The supervised method i.e Poisson regression or quasi-Poisson regression and bootstrap procedure can be easily run with the basic package. However, medoid-based algorithm, internal validation, and visualization required cluster [18] and kmed [19] packages. Both cluster and kmed packages are applicable for medoid-based clustering algorithms. For a reproducible purpose [20], the R code and the data were deposited with DOI: 10.5281/zenodo.5573218, DOI: 10.5281/zenodo.5089156 for the visitors dataset and DOI: 10.5281/zenodo.5089158 for the perception dataset.

III. Result and Discussion

A. Bootstrap samples

The data consist of 80 samples of visitors. With the three schemes of bootstrap samples replicated 1000 times, the perception feature is the only feature that is important in every bootstrap sample (Table 1). It is also found that in each bootstrap sample, the quasi-Poisson regression is always opted because of the over-dispersion.

TABLE 1
Important features and regression types of 1000 bootstrap samples

Bootstrap Samples	Poisson Regression	quasi-Poisson Regression	Important Features
30	0%	100%	Perception
40	0%	100%	Perception
50	0%	100%	Perception

B. Cluster Validation

Based on the result of bootstrap replicates, medoid-based clustering algorithms, which are partitioning around medoids (pam) and simple k-medoids (skm), are applied in the perception features. As a distance-based method, a distance input is required in pam and skm algorithms. Due to the categorical class of all indicators, moreover, the distance applied are simple matching [21] and cooccurrence [22] distances.

Then, the number of clusters (k) is varied from 2 to 6. All of the clustering results are validated via medoid-based shadow value where the value closes to 1 indicates having the best cluster separation [16]. Table 2 shows the result of medoid-based shadow value from all possible combination of the clustering algorithms, distance types, and number of clusters. It indicates 5 clusters via pam algorithm in cooccurrence distance has the highest medoid-based shadow value. Thus, it is opted as the final result.

TABLE 2
Medoid-based shadow value of pam and skm algorithms with k = 2, 3, 4, 5, and 6

Number of clusters	Pam		Simple k-medoids	
	Simple matching	Cooccurrence	Simple matching	Cooccurrence
2	0.38	0.41	0.38	0.41
3	0.38	0.41	0.38	0.39
4	0.38	0.38	0.40	0.38
5	0.35	**0.50**	0.34	0.34
6	0.37	0.47	0.38	0.47

C. Cluster Interpretation

To interpret the clustering results, a barplot is produced (Figure 1). In the barplot, a mean score test can be applied, in which the mean score of each question within a cluster is compared to the mean score of total samples, i.e 80 persons. The interpretation for each cluster follow.

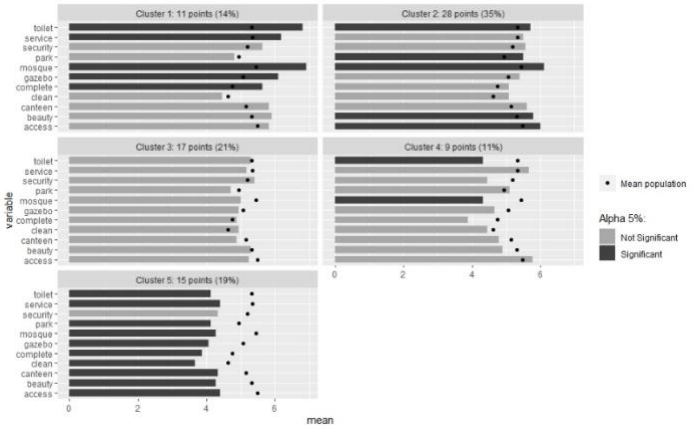

Fig. 1. Barplot of cluster results in the perception indicators

Cluster 1 consisting of 14% samples has satisfied in five aspects, namely the gazebo, mosque, and toilet facilities. They perceive that the facilities are complete and the service quality is above average implied that they are very satisfied. Cluster 2 (35%), on the other hand, has the biggest members. The facilities of mosque, parking spot, and toilet are above average. Moreover, the access to the tourism object and beauty of the scenery are also above average. Cluster 4 as the smallest members, however, perceives the mosque and toilet facilities being lack of concern by the organizer such that they are unsatisfied.

Cluster 3 (21%) and 5 (19%) are different to the other clusters. These cluster members have low score of satisfaction in almost all indicators. The organizer has to considerate these clusters because they composed of 40% samples. Cluster 3 has no significant indicators to the total average, while cluster 5 only has the security indicator that is insignificant. The cleanliness has the lowest average score. With the covid-19 pandemic situation, it has to be the top priority for the organizer to improve the cleanliness of the tourism object in order to improve the tourism object.

IV. CONCLUSION

The bootstrap samples of supervised method in the tourism data resulted in the perception feature as the important feature. Then, the unsupervised method in the perception indicators yielded five cluster as the suitable number of clusters via the cooccurrence distance and pam algorithm. The

combination of supervised and unsupervised method presented an important indicator of the tourism object. To increase the tourism object performance, it suggests that the cleanliness indicator has to be improved.

V. Acknowledgment

The authors would like to thank to the owner of Flower Garden X for the support of our research.

VI. References

[1] James G, Witten D, Hastie T, Tibshirani R, "An Introduction to Statistical Learning with Applications in R" 1st ed. New Jersey, US, Springer, 2013.

[2] Omta WA, van Heesbeen RG, Shen I, de Nobel J, Robers D, van der Velden LM. (2020). Combining Supervised and Unsupervised Machine Learning Methods for Phenotypic Functional Genomics Screening. *SLAS Discovery 25 (6)* 655–64.

[3] El Aissaoui O, El Alami El Madani Y, Oughdir L, El Allioui Y. (2019). Combining Supervised and Unsupervised Machine Learning Algorithms to Predict the Learners' Learning Styles. *Procedia Computer Science 148* 87-96.

[4] Perea-Ortega JM, Martinez-Camara E, Martin-Valdivia MT, Urena-Lopez LA. (2013). Combining Supervised and Unsupervised Polarity Classification for non-English Reviews. *International Conference on Intelligent Text Processing and Computational Linguistics* 63–74.

[5] Lee C. (2021). Predicting Land Prices and Measuring Uncertainty by Combining Supervised and Unsupervised Learning International. *Journal of Strategic Property Management 25 (2)* 169–78.

[6] Krueger R, Beyer J, Jang WD, Kim NW, Sokolov A, Sorger PK. (2020). Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data. *IEEE Transactions on Visualization and Computer Graphics 26 (1)* 227–37.

[7] Bhowmik KR, Das S, Islam MA. (2020). Modelling The Number of Antenatal Care Visits in Bangladesh to Determine the Risk Factors for Reduced Antenatal Care Attendance. *Plos One 15 (1)* 1–17.

[8] Uti U, Essi ID. (2020). Poisson Regression Model with Application to Doctor Visits. *International Journal of Applied Science and Mathematical Theory 6 (1)* 48–68.

[9] Purbowo P, Daroini A. (2021). Determining Tourist Visits and Economic Valuation of Natural Attraction of Tretes Waterfall of Wonosalam. *Agriscience 1 (3)* 625–37.

[10] Hendarto KA, Hasan RA, Yumantoko Y, Nur A, Ariawan K. (2019). The Economic Value of Recreational Benefit of Aik Nyet Nature Tourism, KPHL Rinjani Barat, An Application of The Travel Cost Method. *Jurnal Penelitian Sosial dan Ekonomi Kehutanan 16 (1)* 43–54.

[11] Budiaji W. (2013). Skala Pengukuran dan Jumlah Respon Skala Likert (The Measurement Scale and The Number of Responses in Likert Scale). *Jurnal Ilmu Pertanian dan Perikanan 2 (2)* 127–33.

[12] Zeileis A, Kleiber C, Jackman S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software 27 (8)* 1–25.

[13] Efron B, Tibshirani R, "An Introduction to the Bootstrap", New York, London, Chapman and Hall, 1993.

[14] Kaufman L, Rousseeuw P, "Finding Groups in Data", New York, US, John Wiley and Sons, 1990.

[15] Budiaji W, Leisch F. (2019). Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. *Algorithms 12* 177.

[16] Budiaji W. (2019). Medoid-Based Shadow Value Validation and Visualization. *International Journal of Advances in Intelligent Informatics 5* 76–88.

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

[18] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. (2019). cluster: Cluster Analysis Basics and Extensions [Online] (R package version 2.1.0 --- For new features, see the "Changelog" file (in the package source)) Available from: https://CRAN.R-project.org/package=cluster

[19] Budiaji W. (2021). kmed: Distance-Based k-Medoids [Online] (R package version 0.4.0) Available from: https://CRAN.R-project.org/package=kmed

[20] Budiaji W. (2019). Penerapan Reproducible Research pada RStudio dengan Bahasa R dan Paket Knitr. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika 5 (1)* 1–5.

[21] Huang Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values The First Pacific-Asia Conference on Knowledge Discovery and Data Mining 21–34.

[22] Ahmad A, Dey L. (2007). A K-mean Clustering Algorithm for Mixed Numeric and Categorical Data. *Data and Knowledge Engineering 63* 503–27