Proceedings of 5th International Conference of Economic, Business and Government Challenges 2022. Vol. 1 No. 1 August 2022: 141-151

Analysis of Freight Mode Characteristics on The Northern Coastline Route using Mixed Data based Cluster Analysis

Moch. Abdillah Nafis [1]*, and Dedy Dwi Prastyo [2]
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
manafis49@gmail.com[1*]; dedy-dp@statistika.its.ac.id [2]

Abstract. Due to the increase in the number of vehicles, the maintenance cost on north coast roads (Pantura route) increases because many of its parts are damaged, potholed, and other problems. Mode switching is expected to produce more efficiently, and grouping a mode's characteristics are proven to classify which products need to undergo a modal shift. In this study, a grouping of goods truck modes will be carried out based on the characteristics of the trip, the traveler, and the transportation system on the Northern coastline route to then provide policy recommendations for switching to another mode on the goods truck mode at the Northern coastline route with mostly possible characteristics. The method used for the grouping in this research is clustering analysis, particularly the *k*-prototype clustering method, and partitions around medoids because the data contains mixed variables, i.e., both categorical and numerical variables.
Keywords: Cluster Analysis; Freight Transport; Mixed Data; Transportation.

I. Introduction

The problem of social inequality and population growth continues to increase due to the lack of capacity to support growth in Indonesia as a developing country is more than a developed one. Activities that occur in human life that always move from one place to another cannot be separated from transportation. In addition to the increased cost of land transportation maintenance due to the increase in the number of vehicles, the cost of road maintenance as an intermediary for mobilization will also increase because there are many damaged, potholed, and so on north coast roads. This condition also impacts the saturation on the Highway, especially the Northern coastline Highway. Maintenance must be carried out regularly to anticipate accidents on the Northern coastline route. The northern coastline is a line that crosses from Merak to Banyuwangi along 1,316 km on the island of Java which crosses several provinces, namely East Java, Central Java, West Java, DKI Jakarta, and Banten (Muzaki, 2013).

Many problems still exist on the north coast highway, such as congestion, increasing accidents, increasing air pollution, and road maintenance costs. Congestion that often occurs on the main road is caused by the number of vehicles that exceed the road capacity. The number of vehicles passing by on the road, both private and public transportation, causes an increase in the number of road users. In addition, the increasing number of accidents is also a problem because of the costs incurred due to accidents. Many potholes and bumpy road surfaces are often dangerous for road users. Mode shifting is assumed to result in more efficiency, and clustering has proven to be capable of the grouping in which products are required to undergo a modal shift (Bull, 2004).

So in this study, a grouping of goods truck modes will be carried out based on the characteristics of travel, travel agents, and the transportation system on the northern coastline to then provide policy recommendations for switching to other modes of goods truck mode on the northern coast route with the most possible characteristics. The mode shift is carried out to minimize the level of passenger load inequality, which is expected to maximize the efficiency of shipping goods. Inequality in a load of intermodal passengers must be addressed comprehensively and consistently, such as by increasing the number of train trips on the Java Island double-track railway and optimizing the capacity of container ships (Paddeu *et al.*, 2019). Until now, there is no policy on the transfer of transportation modes which are generally transported by trucks on the Northern coastline, so they are transferred to containers or trains. Many factors need to be considered, such as the impact on motorists as the vanguard of the logistics movement, and companies engaged in logistics need to be considered to produce optimal policies.

The problem in this research is how to group the types of goods/cargo modes on the north coast route based on the characteristics of drivers, trucks, and goods using Mixed Data Cluster Analysis, as well as policy proposals related to the transfer of goods transportation modes based on the results of the grouping. The results of the proper characteristic grouping are expected to be one of the evaluation materials and considerations in implementing the policy formulation for selecting alternative modes of transportation other than trucks on the Pantura Highway. In addition, the results of this study are expected to become a prototype that can be developed in other alternative modes of transportation to reduce congestion on the Pantura highway.

II. Literature Review

Transportation System

The transportation system is an effort to move, transport, or divert people or goods from one place to another by using vehicles driven by humans or machines, which are carried out regularly to run smoothly. There are two most important elements in transportation: movement and physical movement of goods (commodities) and passengers to other places. Based on the type of terrain traversed, land transportation is divided into two including (Anggraini *et al.*, 2016):

1. Road Transportation
   A land vehicle operates on land to facilitate the movement of goods and passengers' transportation to arrive at the destination of goods/passengers, which is carried out safely and efficiently because it can minimize conflict in the movement (Karndacharuk & Hassan, 2017).
2. Railway Transportation
   A rail Vehicle is transportation that uses a rail to access the movement. Rail vehicles are considered a substitute for road vehicles because they are more environmentally friendly and carry more loads (Zak *et al.*, 2014).

Both road and rail transportation have their advantages and disadvantages. Rail vehicles also have several advantages, namely, carrying more cargo than road vehicles and accommodating them, thereby saving on shipping costs. Moreover, the accident rate is lower compared to road vehicles, which can go to the city center that cannot be reached by large vehicles such as trailers and trailer trucks. The absence of traffic disturbances that make deliveries tends to be disciplined. On the other hand, the lack of trains, i.e., low flexibility, makes it difficult to change lines if there are obstacles on the main line. Rail maintenance is also relatively high because trains' main components are metal, requiring additional vehicles (roads and sea lanes) to reach consumers (Sarder, 2021).

Factors Affecting Mode Selection

The mode selection model aims to determine the proportion of people who will use each mode. This process is carried out to calibrate the model selection model in the base year by knowing the attribute changes that affect the mode selection. After the calibration process has been carried out, the model can be used to predict the determination of modes by using attribute coefficients for the future. There are four groups of factors that are considered to affect travel behavior (Miro, 2005). Each of these factors is further divided into several variables that can be identified. This variable can be assessed quantitatively and qualitatively. These factors or variables are:

1. Travel characteristics factor
2. Transportation user characteristics
3. Transportation system characteristics
4. City and zone characteristics

In this study, only travel characteristics, travel agents, and transportation systems were used. These three factors can be implemented at the level of the implementation of goods transportation in the field, with truck drivers as the front line. These variables are expected to represent transportation conditions in Indonesia and be able to differentiate between groups.

Gower Distance

The method of measuring similarities by assigning weights to qualitative and dichotomous data types (has a level for each category). The easiest way to assign weights is to provide a constant value of wk with the following formula (Gower, 1971):

$$S_{ij} = \sum_{k=1}^{v} S_{ijk} w_k(x_{ik}, x_{jk}) \,/\, \sum_{k=1}^{v} \delta_{ijk} w_k(x_{ik}, x_{jk}) \quad \text{.....(1)}$$

with S_{ij} denotes the similarities for all datasets, S_{ijk} denotes the similarities for every cluster, δ_{ijk} means the dissimilarities for every cluster, and $w_k(x_{ik}, x_{jk})$ denotes the weight for the value x_{ik} and x_{jk} for every value *i* and *j*, with $i \neq j$.

K-Prototypes

It is a non-hierarchical type of cluster that covers the shortcomings of the *K*-Means method, which can only be used on numeric data. The *K*-prototypes method eliminates the limitations of the *K*-Means method, which can group non-numeric data because the *K*-Prototype optimizes cost functions such as

Euclidean distance, which measures between points and the cluster average. Minimizing the cost function by calculating the average is limited to numerical data. So it needs particular optimization for non-numeric data. The resulting algorithm from the k-prototype method is identical to the K-Means when applied to numerical data. So that the *K*-prototype is a *K*-Means without the limitation of variable type. The following is the cost function obtained to find similarity (Huang, 1998).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{m_r} a_{ij}^{p} d\left(x_r(i), \mu_r(j)\right) + \gamma \sum_{j=1}^{k} \sum_{i=1}^{m_c} a_{ij}^{p} \delta\left(x_c(i), \mu_c(j)\right) \ldots\ldots(2)$$

with J denotes cost function, a_{ij}^{p} denotes fuzzy value $(a_{1j} + a_{2j} + \cdots + a_{mj} = 1)$, p denotes the weighted parameter a_{ij}. The $x_r(i)$ denotes the *i*-th numeric data, $\mu_r(j)$ denotes the *j*-th numeric cluster, $x_c(i)$ denotes the *i*-th categoric dataset, $\mu_c(j)$ denotes the *j*-th categoric cluster, γdenotes the balancing the influence between categorical and numerical variables, d denotes the distance balancing the influence between categorical and numerical variables, and δ denotes dissimilarities.

Partition Around Medoids

It is a clustering algorithm applying the *K*-means method with the medoid shift algorithm. The two algorithms go hand in hand to minimize the errors of the two algorithms. What distinguishes it from k-means is that the Partition Around Medoids method uses medoids as data entities that can represent the specified group (Han & Kamber, 2006). The following is the similarity formula for the Partition Around Medoids method (Rousseeuw, 1987).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{m} d(x_i, \mu_j) \ldots\ldots(3)$$

with J denotes cost function, x_i denotes the *i*-th data, μ_j denotes the center of cluster j, and $d(x_i, \mu_j)$ is the distance between x_i and μ_j.

Previous Research

Based on the journal entitled "Analysis of the Transfer of Goods Transport Modes on Pantura Highway, Java Island (Case Study: Surabaya-Jakarta Corridor)," it is known that by comparing the components of transactional and non-transactional costs, transport capacity, and the burden of public costs that arise from activities transportation of goods of each mode. The results obtained from the research, namely container trains and container ships, are alternative modes of transportation that can be used to parse the road load by applying the concept of multimodal transportation so that the density of the northern coastline can be reduced by 47.97% in the first year by operating ten series of container trains and four container ships measuring 538 TEUS (Prasetyo & Hadi, 2013).

The research entitled "A Travel Mode Choice Model Using Individual Grouping Based on Cluster Analysis." The study proposes a predictive model to estimate travel mode choice by grouping individuals based on cluster analysis. The results showed that the accuracy levels in the three modal groups were 89.8%, 85.6%, and 78.2%. This result is much higher than without grouping, which is 65.5%. This shows that clustering will increase cluster accuracy. Based on these results, it is known that clusters will make it easier to predict mode choices for tourists so that they can reduce the intensity of congestion and not reduce tourist satisfaction (Ding & Zhang, 2016).

The research entitled “Improving Inland Freight Logistic Efficiencies: Is There any Ideal Modal Split?” related to increasing the efficiency of land transportation logistics in Pakistan by moving some rail transport (railways). It was explained with several considerations: the objective function, repair function, fuel function, timer function, and demand function. By changing the mode from road vehicle to rail mode, the efficiency obtained was much more significant by obtaining the ratio of goods transported (by weight). It is known that the ratio of 96: 4 (left for the ratio of goods transported by train, right for the ratio of goods carried by road vehicles) results in cost-efficiency of 17.08 million Rupee. So it is better for logistics transportation in Pakistan which is dominated by rail, because the efficiency is very high (Ali *et al.*, 2022).

III. Research Method

Type, Population, and Variable

This research is applied research that leads to a quantitative approach to producing the proper grouping for the movement of goods transportation on the Pantura Line. The population to be studied is every actor in the delivery of goods by using land transportation modes. The modes that are focused on in this research are road and rail modes. Samples were taken using the random purposive sampling method,

namely by doing random for several population groups selected according to the researcher's needs. The source of data used in this study is based on a direct survey of land transport freight forwarders. The variables used are the latest education, age, income, income, type of commodity, cost of savings, mileage, travel time, number of breaks, length of rest, overnight / not, and number of monthly trips.

Data Analysis Technique

The cluster analysis method used in this study is the *K*-Prototype and Partition Around Medoids method using Gower's Distance as similarity measurement. The advantage of using the *K*-Prototype method is that it can overcome numeric and non-numeric (mixed) variables. However, the algorithm of the k-prototype is identical to the k-means. As for Partition Around Medoids, it is more directed to the use of medoids than centroids. The open-source Python is the computational tool used to analyze the data in this research.

Research Steps

The following are the steps of this research:

1. Obtaining the data of driver of cargo delivery truck for research material.
2. Pre-processing data by looking for missing values and outliers, followed by measuring data adequacy.
3. If there are outliers or missing values, they can be removed or maintained.
4. If the sample does not meet the adequacy, then add new data and repeat step 2.
5. Perform cluster analysis on mixed data with the *K*-prototype method, and partition around medoids, with Gower distance.
6. Determine the optimal number of clusters using the elbow and silhouette methods.
7. After obtaining the optimal number of clusters with the method that produces the highest silhouette value, information is extracted from the clustering results by grouping each variable against each cluster resulting from the best clustering method between k-prototype and partition around medoids, then performs visualization with boxplot and barplot to obtain variables that distinguish between clusters in the clustering method.
8. After finding the variables that distinguish each cluster, a group with characteristics that have more potential for mode switching will be determined compared to other groups.
9. Conclude and provide recommendations based on extracting information in step 8 and determining clusters that require mode shifting from trucks to another mode (e.g., trains).

IV. Results and Discussion

Before the cluster analysis is carried out, a descriptive analysis will be done by making comparison charts between the variables, as shown in Figure 1. Based on Figure 1, it is known that there is a significant positive relationship between Transportation Costs, Mileage, and Travel Time on the route. The three variables have a graphic pattern that tends to show a negative relationship to the number of trips per driver in one month, but not so significant. It is also known that the distribution of data from these six numerical variables tends to be positive skewness, and most data tend to have values below the average. These empirical facts mean fewer drivers with higher distance and travel time, transportation costs, number of breaks, and number of monthly trips.

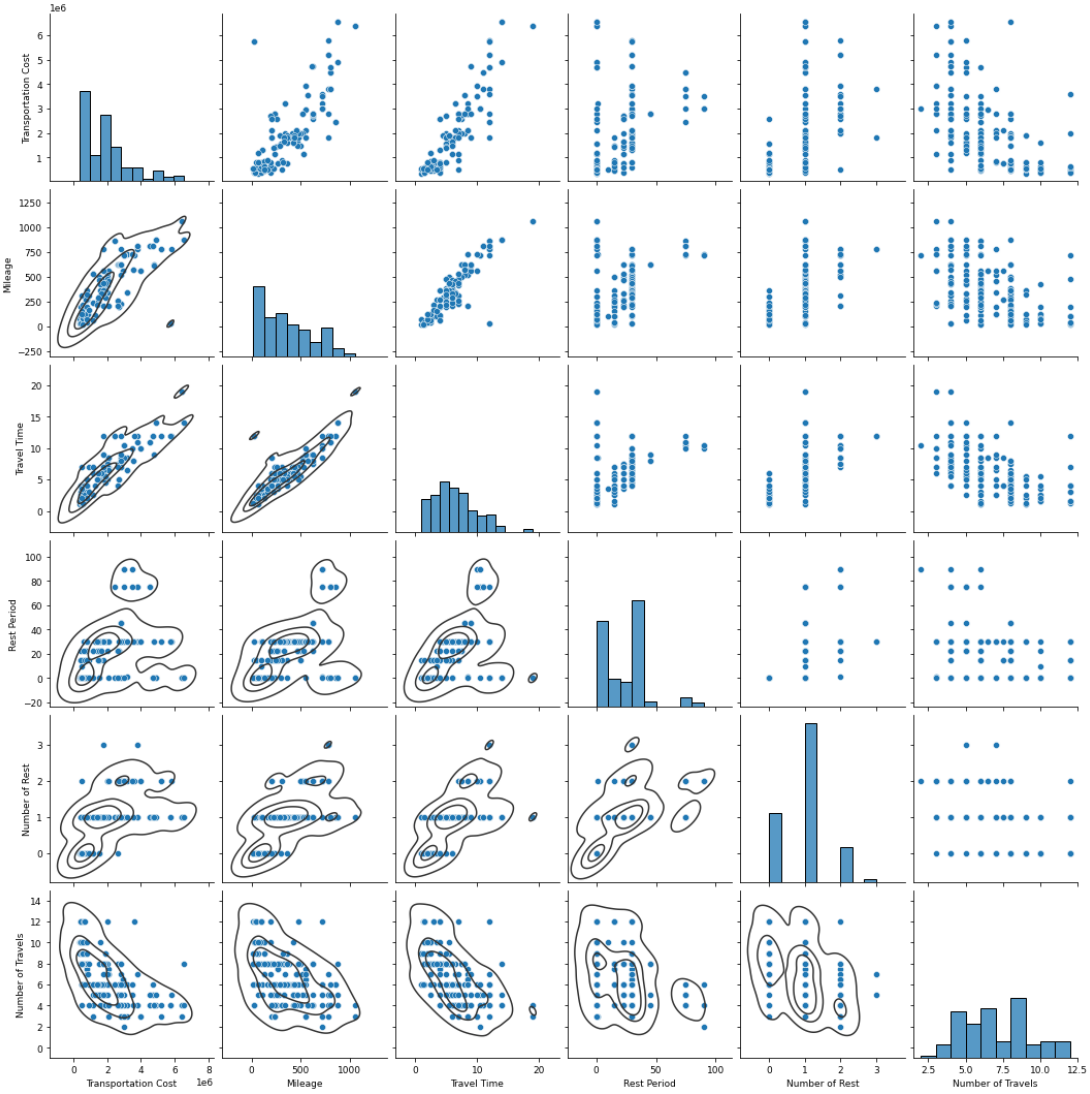

Figure 1. Graphs for Comparison Between Variables

Characteristics of Categorical Variables

Here are some graphs for each categorical variable:

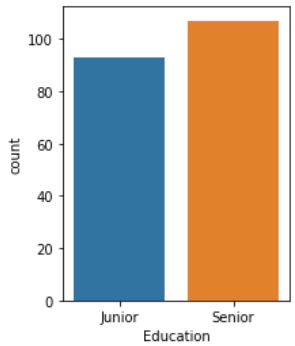

Figure 2. Graphs for Education Variable

Based on Figure 2, it is known that the respondent's education tends to be high school education. However, there are quite a number of respondents with a junior high school education level, but not as many as high school students. The following are also the characteristics of the age variable.

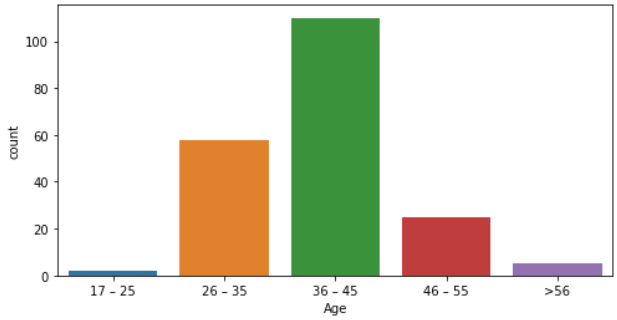

Figure 3. Graphs for Age Variable

Based on Figure 3, it is known that the respondent's age forms a normal distribution pattern, in which the majority of respondents are 36-45 years old. In addition to the age of 36-45 years, there are fewer. It can be concluded that the respondents (drivers), on average, tend to be young (under 45 years), which is explained by the skewness of the graph, which tends to be heavier to the left (more frequency on the left side of the average). After the age variable, the respondent's income characteristics are as follows.

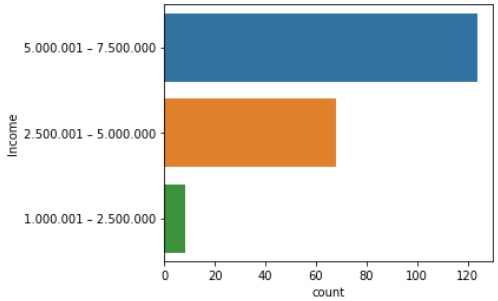

Figure 4. Graphs For Income Variables

Based on Figure 4. It is known that respondents tend to have a reasonably high income, in the range of 5-7 million Rupiah. However, this can also indicate that the income also includes travel costs, ranging from toll fees and fuel to urgent repairs.

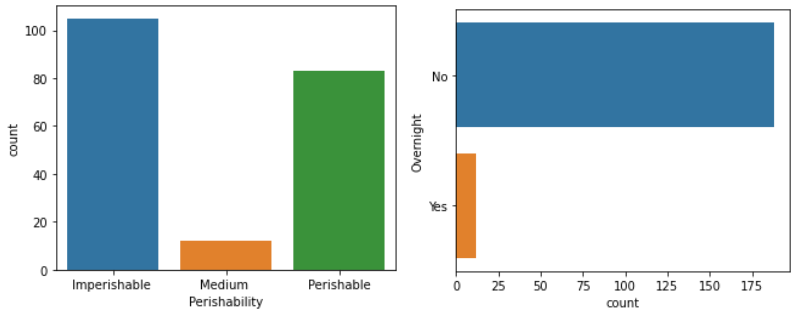

Figure 5. Graphs for Variables of Savings and Overnight / Whether Respondents in Delivery

Based on Figure 5. It is known that most deliveries are goods that are difficult to rot, and there is a difference with perishable. Since perishable goods do not appear to be in a rush in delivery, long-distance travel and/or long delivery times (which allow drivers to stay overnight, even though the data show only a few stays) can be allocated to trains.

Cluster Analysis

After getting the characteristics of each variable, cluster analysis was obtained using the k-prototype method and partition around medoids using Gower's distance. The following is an elbow and a silhouette of each method used.

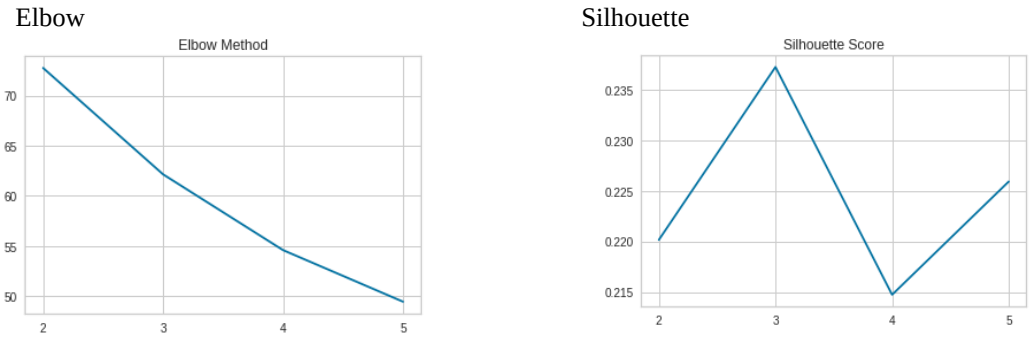

Figure 6. *Elbow* and *Silhouette* from *K*-Prototype Methods

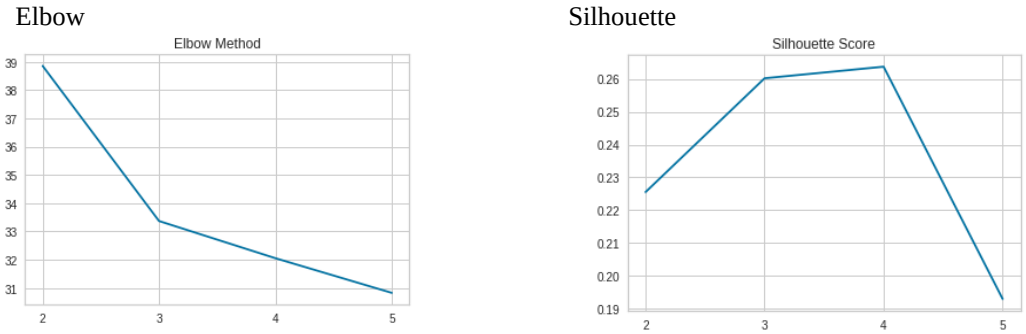

Figure 7. *Elbow* and *Silhouette* from Partition Around Medoids Methods

Based on Figures 6 and 7, the PAM (Partition Around Medoids) method was chosen because it has a higher silhouette score than the K-Prototype. Although the PAM method with four groups gives a slightly better silhouette score, patterns made for 3 clusters in overall evaluation methods are more consistent in both PAM and K-Prototype sections, both elbow method and silhouette score. So for the results of this study, the PAM clustering method was taken with three groups. The following is the division of the three clusters from the PAM method.

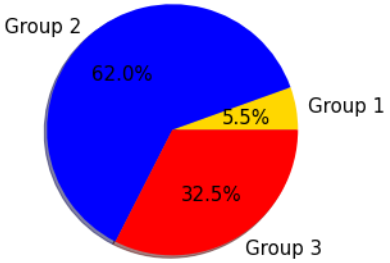

Figure 8. Cluster Grouping with PAM Method

Based on Figure 8, it is known that the number of cluster 1 is very small, with only 11 respondents. Cluster 2, which dominates the obtained cluster, which is more than 60%, is about 124 respondents. Cluster 3 consists of 65 respondents. The following are the characteristics of each cluster.

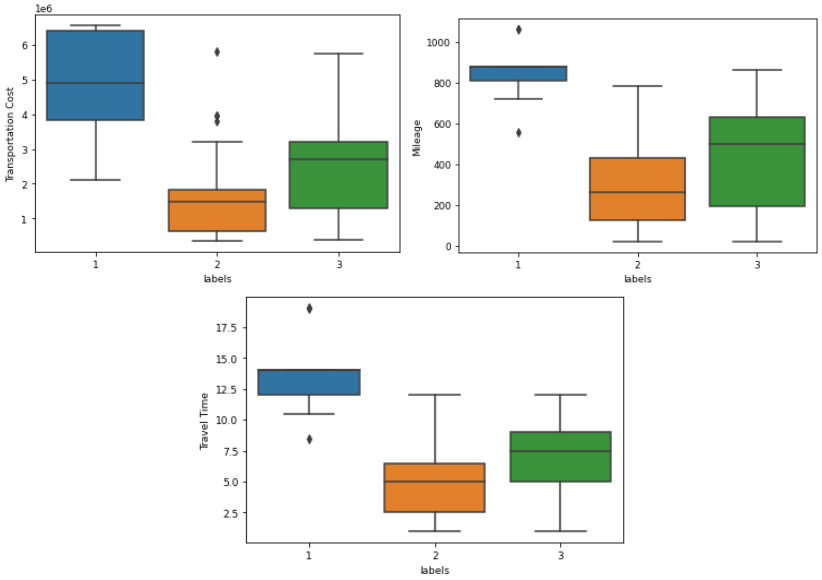

Figure 9. Characteristics of Transportation Costs, Mileage, and Travel Time of Each Cluster

Based on Figure 9, it is clear that transportation costs, distance traveled, and travel time have the same pattern. Cluster 1 is bigger, far, and longer than other clusters. There is also a tendency for cluster 3 to have higher transportation costs, mileage, and travel time than cluster 2, although several routes from cluster 2 have higher transportation costs, distance traveled, and travel time than cluster 3.

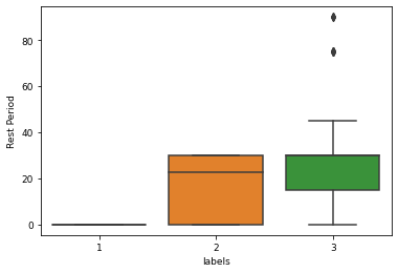

Figure 10. Characteristics of Rest Period of Each Cluster

Based on Figure 10, it is known that cluster 1 is the respondent who stays during the delivery, so they do not rest during the trip except at night to sleep. Cluster 3 tends to rest longer than cluster 2, although there is no significant difference between the two groups.

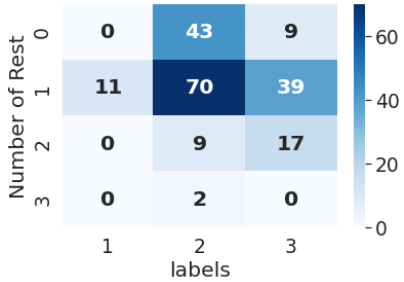

Figure 11. Frequency of Rest Number of Each Cluster

In Figure 11, for cluster 1, there is only one break (during an overnight stay) for as many as 11 respondents. Group 3 rested slightly more than group 2, although there was no significant difference between the two groups. The 120 out of 200 drivers take only one trip break.

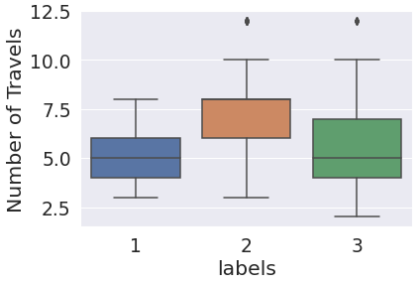

Figure 12. Travel Characteristics in 1 Month Each Cluster

Based on Figure 12, it is known that cluster 2 travels more than groups 1 and 3 because the travel time tends to be shorter. Uniquely, there is no significant difference between 1 and 3 even though the distance and travel time of 1 is higher than 3. Cluster 3 is slightly more trips but not significant.

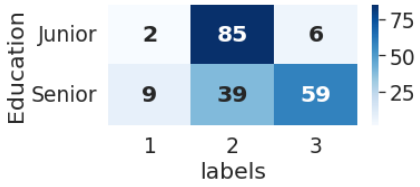

Figure 13. Frequency of Respondent Education in Each Cluster

In Figure 13, it is known that high school education is more dominant in clusters 1 and 3, while cluster 2 is dominated by respondents with a junior high education level. So it can be concluded that respondents with a junior high school education level are prioritized for short-distance travel, given that Figure 9 shows that cluster 2 is a cluster with short-distance travel and less transportation costs.

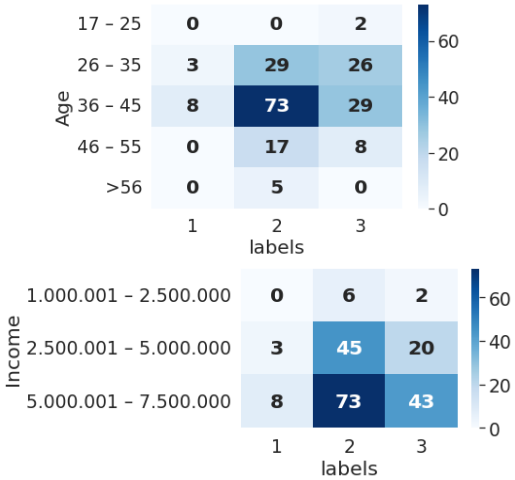

Figure 14. Frequency of Respondents' Age and Income in Each Cluster

It is known that in Figure 14 that the entire cluster is dominated by a single type of each category mentioned, such as the age of 36-45 years and 5.000.001-7.500.000 in income. So the variable will not be used as a differentiator between clusters.

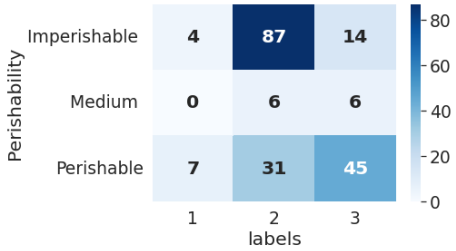

Figure 15. Frequency of Product Shelf Life Transported by Respondents in Each Cluster

Based on Figure 15, it is known that cluster 1 cannot explain the tendency of the storage capacity of the transported goods because the difference is not much different, and there are only 11 respondents in this group. However, in terms of the mode of transportation of goods, delivery that can stay overnight is delivery that carries goods that are not perishable. Cluster 2 tends to carry a lot of imperishable items, while cluster 3 tends to carry a lot of perishable items.

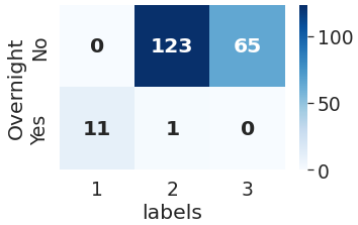

Figure 16. Frequency of Overnight/No Respondents in Each Cluster

In Figure 16, it is known that in cluster 1, all respondents stay overnight because the average trip is above 700 Km and the travel time without a break is more than 10 hours, and for clusters 2 and 3, they do not stay overnight (only rest a few times along the way).

V. Conclusion

Based on the results of clusters using the Partition Around Medoids (PAM) method of 3 clusters, the characteristics and recommendations of researchers for each cluster are as follows:

1. Cluster 1 tends to travel long distances with high transportation costs. It can be proposed to provide subsidies to logistics companies if they are willing to switch to the rail mode. Because apart from the limited allocation of funds for subsidies due to the infrequent travel frequency, the distribution of goods can be faster. The provision of subsidies can be more for companies engaged in commodity goods with a perishable category.
2. Although the cost, distance, and time are lower than in other clusters, the frequency of trips in cluster 2 is high. In addition, this cluster has the most members, especially in transporting imperishable commodities, so this group has the most influence on the density of the north coast highway and the most potential for modal shifts. Based on these empirical findings, several bold policies might be needed for companies with a pattern of distribution of goods that are close to the characteristics of this group 2, especially for imperishable commodities. It should be noted, in some cases of distribution of goods over short distances, using the train mode is cheaper than by road.
3. With characteristics that are classified as moderate compared to the previous two groups, with a fairly low frequency of trips and dominated by perishable goods, in cluster 3, it seems that the transportation of goods by road is still an optimal choice for this group.

References

Ali, Y., Sabir, M., Abubaker, A., Saad, H., & Ali, S. (2022). Improving Inland Freight Logistic Efficiencies: Is There any Ideal Modal Split? *Elsevier*.

Anggraini, E. (2016). *Manajemen Transportasi Barang.* Jakarta: Sekolah Tinggi Manajemen Transportasi Trisakti.

Bull, A. (2004). *Traffic Congestion: The Problem and How to Deal With it.* Santiago: United Nations Publication.

Ding, L., & Zhang, N. (2016). A Travel Mode Choice Model Using Individual Grouping Based on Cluster Analysis. *Procedia Engineering 137*, 786-795.

Gower, J. (1971). *A General Coefficient of Similarity and Some of Its Properties.* International Biometric Society.

Han, J., & Kamber, M. (2006). *Data Mining: Concept and Techniques, 2nd edition.* San Fransisco: Morgan Kauffman.

Huang, Z. (1998). Clustering large data sets with mixed numeric and categorical. *In Proceedings of the 1st Pacific-Asia conference on knowledge discovery* (pp. 21 - 34). PAKKD.

Institut Teknologi Sepuluh Nopember. (n.d.). *Share ITS.* Retrieved from Share ITS: http://share.its.ac.id/pluginfile.php/582/mod_resource/content/1/Penyediaan_Sarana_dan_Prasarana_Transportasi_Jalan_Raya_dan_Jalan_Rel_.pdf

Karndacharuk, A., & Hassan, A. (2017). *Road Transport Management Framework and Principles.* Sydney: Austroads Ltd.

Miro, F. (2005). *Perencanaan Transportasi untuk Mahasiswa, Perencana, dan Praktisi*. Jakarta: Erlangga.

Muzaki, L. (2013, November 25). *JALUR PANTURA JAWA BARAT SELALU MACET PADA SAAT LIBUR BESAR*. Retrieved from Dinas Perhubungan Provinsi Jawa Barat: http://dishub.jabarprov.go.id/artikel/view/352.html

Paddeu, D., Calvert, T., Clark, B., & Parkhurst, G. (2019). *New Technology and Automation Freight Transport and Handling System. Future of Mobility: Evidence Review*. Bristol: Foresight: Government Office for Science.

Prasetyo, A., & Hadi, F. (2013). Analisis Pemindahan Moda Angkutan Barang di Jalan Raya Pantura Pulau Jawa (Studi kasus: Koridor Surabaya - Jakarta). *Jurnal Teknik POMITS*, Vol. 2, no. 1.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 53 - 65.

Sarder, M. (2021). *Logistic Transportation System.* Cambridge: Joe Hayton: Elsevier.

Zak, J., Jacyna-Golda, I., Merkisz-Guranowska, A., Lewczuk, K., Klodawski, M., Pyza, D., . . . Wasiak, M. (2014). The Role of Railway Transport in Designing a Proecological Transport System. *Computers in Railways XIV Special Contributions*, Vol. 155.