

Passenger and Revenue Estimation for New Rail Transit Lines Under Construction: A Demographic Approach

Tarisma Dwi Putri Alifianti¹, Rifdatun Ni'mah², Regita Putri Permata³

^{1,2,3}Department of Data Science, Telkom University, Indonesia

Article Info

Article history:

Received Jun 29, 2025

Revised Jul 25, 2025

Accepted Oct 27, 2025

Keywords:

Rail-Based Transportation

Passenger Estimation

Revenue

Markov Chain

Fuzzy C-Means

ABSTRACT

This study proposes a data-driven approach to estimate passenger volume and revenue for new rail transit lines under construction, addressing the challenge of limited historical data. Principal Component Analysis (PCA) was used to reduce 29 demographic variables into three principal components, which collectively captured up to 85% of the variance. These components informed a Fuzzy C-Means (FCM) clustering process that grouped new stations with existing ones based on demographic similarity. The clustering yielded a Fuzzy Partition Coefficient (FPC) of 0.913, indicating high cluster validity and low overlap between clusters. Transition probabilities of passenger flows between stations were modeled using Markov Chains. The expanded transition matrix, incorporating new stations through demographic analogy, demonstrated rapid convergence to a stationary distribution within 5–10 iterations, validating the model's stability. Simulation results project a 57% increase in weekday passengers and a 74% increase in weekend passengers, with estimated daily revenue peaking at Rp1.216 billion. The evaluation results confirm the robustness and reliability of the combined FCM–Markov model for long-term passenger and revenue forecasting in new transit infrastructure planning.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rifdatun Ni'mah,

Department of Data Science,

Telkom University,

Ketintang Street No.156, Ketintang, Gayungan District, Surabaya, East Java 60231.

Email: rifdatun@telkomuniversity.ac.id

1. INTRODUCTION

Transportation acts as the mechanism for moving people or goods between distinct locations to meet various needs [1]. According to BPS, DKI Jakarta is the province with the highest population density in Indonesia, with a population of over 10 million people as of 2024 [2]. As the population grows, the number of private vehicles increases significantly. In 2022, the number of private vehicles in the capital city increased by 4.02% compared to the previous year [3]. This factor is one of the causes of congestion in DKI Jakarta. Based on Abdulkadir's statement in his book Commerce Transportation Law, various steps need to be taken by the government to improve public welfare, including through the transportation sector, which plays a role in developing facilities to support population productivity and reduce various environmental problems[4][5].

One of the public transportation modes in Jakarta is rail-based transportation. Railways in Jakarta have a long history that is inextricably linked to the city's development. The railway system in Jakarta, formerly known as Batavia, began in the 18th century [6]. Until now, the railway system in Jakarta continues to develop with the construction of new stations [7]. However, data on the new station infrastructure is not yet available. Therefore, a predictive approach is needed to estimate the number of passengers and revenue at these new stations.

Predicting the number of passengers and revenue in the construction phase of new stations in Jakarta is very important for several reasons [8]. First, revenue predictions are necessary to assess the feasibility of investing in the construction phase of new stations. Second, predictive data

is crucial for developing effective operational strategies. In addition, this prediction can also serve as a reference for the development of areas around new stations. And then, the estimated data generated from this study have strategic value in long-term decision-making, including infrastructure planning, service development, and marketing strategies.

One of the relevant methods for analysing passenger movement patterns between stations and overcoming data limitations at stations under construction in this study utilises a data-based approach that combines Markov chain analysis and clustering using the Fuzzy C-Means algorithm. Markov chains are stochastic processes that can be used to model the probability of passenger movement from one station to another based on transition probabilities [9][10].

Meanwhile, the Fuzzy C-Means method is used to group new stations based on the demographic characteristics of the surrounding area, which refers to the administrative area within a geographic radius defined as one sub-district. This algorithm divides the data into several clusters that represent groups with the closest degree of membership [11]. By combining these two methods, this study aims to produce estimates of the number of passengers and income that can support decision making in the second phase of development.

In the modern era, The rail network serves as the backbone constrof public transportation, connecting Jakarta with satellite cities of Jabodetabek, and continues to adapt to the increasing number of passengers and city traffic congestion [12]. The ability to accurately project revenue is essential for supporting financial planning and the development of railway services. Research related to the projection of passenger numbers and revenue in public transportation modes is still relatively scarce in the academic literature [13]; therefore, this research is expected to make a significant contribution to the development of studies in the field of urban transportation.

Previous studies have explored demand prediction in various transportation systems, such as bike-sharing systems. For instance, Zhou et al. (2018) developed a Markov Chain-based demand prediction model for bike-sharing stations, demonstrating higher forecasting accuracy and better generalization ability using field data from Zhongshan City [14]. While their focus was on bike-sharing, the principles of using Markov chains for predicting user movement and station activity are relevant and provide a valuable benchmark for our approach to rail-based systems. This study aims to adapt and evaluate a similar methodological rigor by validating our model using historical data from existing railway lines in Jakarta, where comparable transit patterns and operational characteristics can be observed, to ensure the robustness and applicability of the predictions for new station infrastructure.

2. RESEARCH METHOD

In this study, problem solving is designed systematically to ensure that each stage is carried out in a structured manner. The flowchart of the research method, shown in Figure 1, is designed to visualise the entire process.

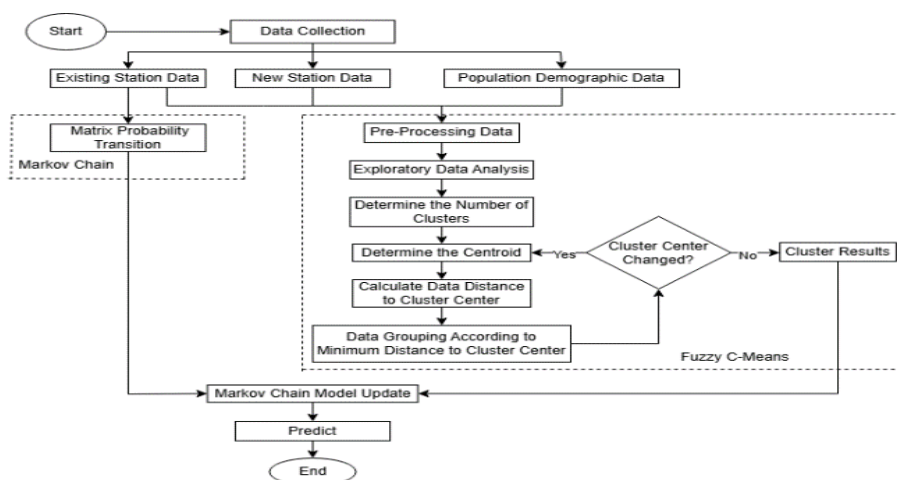


Figure 1. Research Method Diagram

This study utilises three types of data: existing station data, new station data, and demographic data. The existing station data covers a two-year period, from January 2023 to December 2024. Meanwhile, population demographic data were obtained from various official sources from the DKI Jakarta provincial government and national institutions. This dataset comprises 29 variables in 8 sub-districts and 18 urban villages surrounding the station. A detailed explanation of the variables used is presented in Table 1.

Table 1. Detailed of The Variables

Data Type	Variable	Description
Existing Station Data	Origin Station	Name of the passenger's departure station.
	Destination Station	Name of the passenger's destination station.
	Number of Passengers	Total number of passengers using a specific station.
New Station Data	Ticket Price	Price of inter-station travel tickets.
	Station Name	Names of new stations currently under construction.
	Sub-district	Name of the sub-district where the data was obtained.
	Urban Village	Name of the urban village where the data was obtained.
	Latitude	Latitude coordinate of the location.
	Longitude	Longitude coordinate of the location.
	Average Distance	Average measured distance.
	Total Indonesian Residents	Total number of Indonesian citizens.
	Total Foreign Residents	Total number of foreign citizens.
	Number of RWs	Number of Community Units (Rukun Warga).
	Number of RTs	Number of Neighborhood Units (Rukun Tetangga).
	Population Density (Persons/km ²)	Population density per square kilometer.
	Population Growth in 2024 (%)	Population growth rate in 2024.
	Proportion of Disabled Population	Proportion of the population with disabilities.
Demographic Data	Proportion of Poor Social Aid Recipients	Proportion of the poor population receiving social assistance.
	Total Area (km ²)	Total area in square kilometers.
	Distance to Sub-district Capital (km)	Distance from the location to the sub-district capital in kilometers.
	Distance to District/City Capital (km)	Distance from the location to the district/city capital in kilometers.
	Number of Places of Worship	Total number of places of worship.
	Number of Accommodation Facilities	Total number of accommodation facilities.
	Number of Bank Financial Institutions	Number of bank financial institutions.
	Number of Cooperative Financial Institutions	Number of cooperative financial institutions.
	Number of Trade Facilities	Number of trade facilities.
	Number of Educational Facilities	Number of educational facilities.
	Population Age < 5 Years	Number of residents under 5 years old.
	Population Age 5-9 Years	Number of residents aged 5 to 9 years old.
	Population Age 10-19 Years	Number of residents aged 10 to 19 years old.
	Population Age 20-59 Years	Number of residents aged 20 to 59 years old.
	Population Age > 60 Years	Number of residents over 60 years old.

After the data is collected, it will undergo the data preprocessing stage, which aims to clean, transform, and integrate data from various sources, making it ready for analysis [15]. Before entering the Markov chain, demographic data will be clustered using the Fuzzy C-Means (FCM) algorithm to obtain data for new stations that have the same demographic patterns as existing stations. FCM treats each data point probabilistically, assigning a continuous membership value to each cluster prototype and displaying a finer level of similarity between clusters [16] [17]. Consequently, FCM is often referred to as a soft clustering method [18].

The clustering process using FCM is carried out through the following series of steps [19]:

1. Set
 - C : number of clusters.
 - $m > 1$: degree of fuzziness.
 - ε : error tolerance limit value.
2. Randomly initialise the initial cluster centre position $\{c_i^{(0)}\}_{i=1}^C$.
3. Set iteration $k \leftarrow 1$.

4. Repeat until convergence¹:(a) Calculate the membership degree for each data point X_j and cluster i , calculate:

$$u_{ij}^{(k)} = \frac{1}{\sum_{l=1}^C \left(\frac{d(c_i^{(k-1)}, X_j)}{d(c_l^{(k-1)}, X_j)} \right)^{\frac{2}{m-1}}}$$

(b) Update the cluster centre for each cluster i , set:

$$c_i^{(k)} = \frac{\sum_{j=1}^n (u_{ij}^{(k)})^m X_j}{\sum_{j=1}^n (u_{ij}^{(k)})^m}$$

(c) Increase the iteration index: $k \leftarrow k + 1$ 5. After get the convergence criteria, return the final cluster centre set $\{c_i^{(k)}\}_{i=1}^C$.

After clustering, the historical passenger data will be modelled to obtain estimates at new stations using the Markov chain model, to model passenger movement between stations. A Markov chain is a mathematical model that represents a stochastic system in which the probability of transitioning from one state to another depends only on the current state, not on the sequence of previous states, which states that the future of a process is determined solely by its current state, regardless of its history [20] [21].

Mathematically, this Markov property can be formulated as follows [22]:

$$P(S_{t+j} | S_t, S_{t-1}, \dots, S_{t-(n-1)}) = P(S_{t+j} | S_t)$$

The equation above illustrates that the probability of event S_{t+j} depends only on S_t , which is the current state.

A Markov process can be illustrated as shown in figure below, which depicts three states with transition probabilities between them.

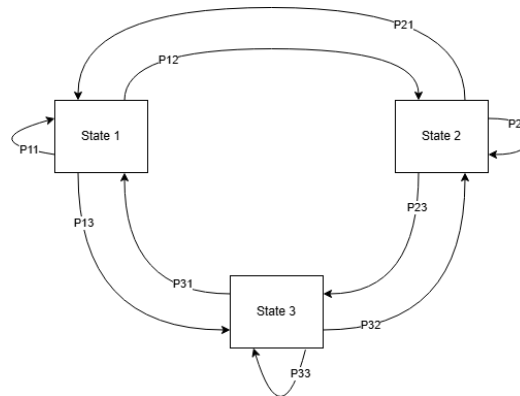


Figure 2. Illustrates of Markov Chain

The diagram in the figure illustrates a system with three states: State 1, State 2, and State 3. Each state represents the system's status at a given time, while the arrows connecting the states show the probability of transitioning from one state to another. These probabilities are expressed as transition probabilities P_{ij} , which represent the probability of transitioning from state i to state j , where i and j are elements of $\{1, 2, 3\}$.

- If the system is in State 1:
 - It has a probability of remaining in State 1, denoted as P_{11} .
 - It has a probability of moving to State 2, denoted as P_{12} .
 - It has a probability of moving to State 3, denoted as P_{13} .

¹ Convergence occurs when $\max_i \|c_i^{(k)} - c_i^{(k-1)}\| < \varepsilon$.

- If the system is in State 2:
 - It has a probability of remaining in State 2, denoted as P_{22} .
 - It has a probability of moving to State 1, denoted as P_{21} .
 - It has a probability of moving to State 3, denoted as P_{23} .
- If the system is in State 3:
 - It has a probability of remaining in State 3, denoted as P_{33} .
 - It has a probability of moving to State 1, denoted as P_{31} .
 - It has a probability of moving to State 2, denoted as P_{32} .

Transition probability describes the likelihood of moving from one state to another in the next time step. This transition probability can be expressed in a matrix like this:

$$P(i, j) = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}$$

Where:

- P_{ij} denotes the probability of moving from state i to state j .
 - $i, j = 1, 2, 3$, corresponding to the states shown in the diagram.
- In the diagram, the interstate transition probabilities include:
- P_{12}, P_{13} for transitions from State 1 to State 2 and State 3.
 - P_{21}, P_{23} for transitions from State 2 to State 1 and State 3.
 - P_{31}, P_{32}, P_{33} for transitions from State 3 to State 1, State 2, and remaining in State 3.

These transition probabilities represent the dynamics of a stationary process in a Markov chain, where the probability of each matrix element P_{ij} can be calculated using the formula:

$$P_{ij} = \frac{n_{ij}}{n_i}$$

with:

- n_{ij} : The number of transitions from state S_i to S_j .
- n_i : The total number of transitions from state S_i .

The main challenge is to expand the transition matrix to 20x20, incorporating seven new stations with no historical data. Transition probabilities for these new stations will be estimated based on demographic classification. It is assumed that new stations will have similar passenger generation and withdrawal patterns to those of old stations in the same demographic cluster. This demographic analogy approach enables the estimation of transition probabilities, yielding a comprehensive 20x20 matrix as a basis for prediction.

1. Transition Probability Matrix of Existing Stations

A fundamental step in this analysis is the construction of a transition probability matrix based on historical passenger data of existing stations. The raw data are aggregated into seven matrices representing the average daily passenger flows (Monday to Sunday) between stations, producing a transition probability matrix T^{hist} , where each element T_{ij} represents the historical probability of passenger movement from origin station i to destination station j .

2. Initial Estimates of Initial Passenger Numbers at Existing and New Stations

The integration of new stations into the model requires initial parameter estimation. The average daily passenger number for each new station is estimated by mapping the new stations into relevant clusters based on the results of the Fuzzy C-means analysis. Next, the average passenger value of the corresponding cluster is adopted as the estimated value for the new station. This process produces initial passenger population data for all 20 stations.

3. Formation of Passenger Flow Probability Estimates from/to Clusters

The passenger flow probabilities between clusters are estimated by combining the historical transition probability matrix T^{hist} with the sum-product of cluster membership degrees

from the Fuzzy C-means (FCM) analysis. This process enables the model to estimate passenger movements at a broader cluster level, providing a more generalized picture of travel patterns, rather than focusing only on individual stations.

4. Formation of Similarity Matrix

A key element in estimating passenger movement patterns at new stations is measuring their characteristic similarity to existing stations. This is achieved through the formulation of a proximity matrix, which quantifies the degree of similarity between each pair of stations based on their fuzzy membership profiles. The proximity value between station i and station j (S_{ij}) is calculated using the fuzzy scalar product approach [23]:

$$S_{ij} = \sum_{k=1}^c u_{ik} \cdot u_{jk}$$

Where u_{ik} is the membership degree of station i in cluster k , and c is the total number of clusters. This matrix S is the basis for transferring known transition patterns from old stations to new stations.

5. Estimation and Normalization of Flow Probability Matrices of Existing and New Stations

The proximity matrix and the historical transition probability matrix are combined to form a combined flow probability matrix, including both old and new stations. The flow probability from station i to j is calculated as:

$$T'_{ij} = \sum_{p=1}^{N_{F1}} S_{ip} \cdot T_{pj}^{hist}$$

Where N_{F1} is the number of existing stations, this equation effectively projects the travel pattern of the new station based on its similarity to the existing stations. Each element in row i of the matrix T' is divided by the total sum of all elements in that row. This process produces the final normalized transition probability matrix (T^{norm}).

The normalized transition matrix will be used to make predictions with Markov chains. The prediction results are used to estimate the number of passengers and revenue at the new station.

$$P_{t+1} = P_t \cdot T$$

Where P_t is a vector of the number of passengers at time t , and T is a transition matrix that describes the probability of passenger transfers between stations. Based on the results of this multiplication, the estimated number of passengers in the next period is obtained. The estimated revenue on each route between stations i and j can be calculated as follows:

$$Revenue_{ij} = Number\ of\ Passengers_{ij} \times Fare_{ij}$$

Where $Revenue_{ij}$ is the revenue generated from the route between stations i and j , $Number\ of\ Passengers_{ij}$ is the projected number of passengers for the route, and $Fare_{ij}$ is the fare charged for travel between the two stations.

3. RESULTS AND DISCUSSION

3.1. Dimensionality Reduction

To overcome high dimensionality and multicollinearity in demographic data, Principal Component Analysis (PCA) is used as a dimension reduction method [24]. PCA transforms the original correlated variables into linearly independent principal components, while maintaining the data variance [25]. The data is standardized using Standard Scaler to ensure that all variables are on the same scale. Three principal components are selected because they collectively explain approximately 80% to 85% of the data variation.

The PCA results show that the three principal components represent the data structure as follows:

1. PC1: Describes the demographic scale and basic infrastructure, distinguishing areas based on population and basic facilities.

2. PC2: Shows the contrast between dense urban areas and sprawling suburbs, serving as an indicator of the level of urbanization.
3. PC3: Reveals the dimensions of regional growth and development, indicating areas that are developing demographically and economically.

3.2. Clustering New Stations

After obtaining the three main components from the PCA dimension reduction results, the next stage is to apply the Fuzzy C-Means (FCM) algorithm. Fuzzy C-Means (FCM) is a clustering method that is conceptually developed from K-Means by incorporating the concept of degree of membership (fuzziness) into determining the ownership of data points in clusters [26]. FCM is a soft clustering method, where each data point not only belongs to one cluster but also has a degree of membership in all clusters [27]. This approach is more flexible and suitable for data with unclear boundaries between groups. The results of this process are expected to form a typology of areas around existing stations and new stations, which will be used as input for modelling passenger numbers and revenue projections. FCM is applied with a total of four clusters.

The results of FCM clustering are visualized in Figure 2. From this visualization, the characteristics of each cluster can be identified as follows:

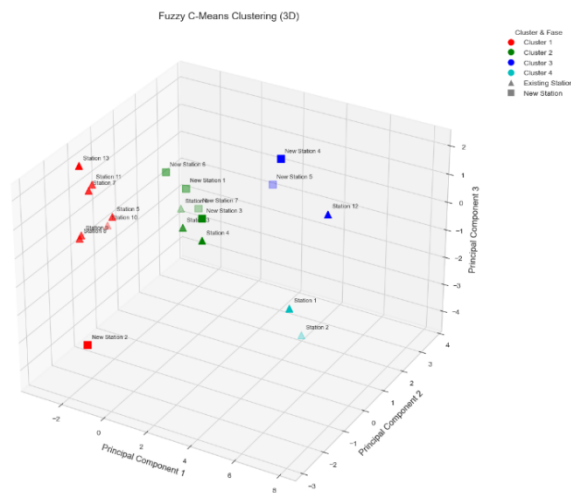


Figure 3. Result of Cluster Fuzzy C-Means

1. Cluster 1 represents commercial and government centres, which have high economic and development activities but low residential populations.
2. Cluster 2 represents established suburban settlements, characterized by large populations, extensive areas, and relatively stable growth patterns.
3. Cluster 3 encompasses the main urban areas, including dense urban centres and modern growth centres, reflecting the highest urbanization intensity.
4. Cluster 4 represents a transitional urban area that is multifunctional, stable, but less dense and active than the primary urban centre.

The performance of the FCM model was evaluated using the Fuzzy Partition Coefficient, which yielded a value of 0.913. The FPC value approaching 1 indicates that the clustering results are pretty straightforward, with a low level of overlap between clusters. The convergence process was achieved after going through 13 iterations.

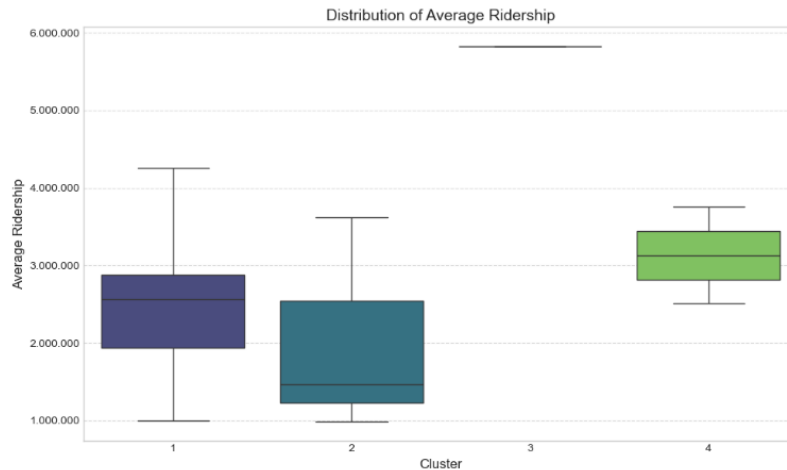


Figure 4. Distribution of Average Ridership in Existing Station

Based on the graph above, it can be observed that the stations operating in each cluster are assumed to have a similar number of passengers. This is because each cluster is assumed to have the same demographic characteristics, resulting in a similar or close number of passengers generated in the estimation. Cluster 1 shows that most stations have an average ridership between around 2 million to 2.8 million with some higher outliers. Cluster 2 has a lower ridership, primarily in the range of 1.2 million to 2.5 million. Cluster 3 stands out with the highest and most consistent average ridership at around 5.8 million indicating minimal variation between stations in the cluster. Meanwhile, Cluster 4 shows a medium average ridership between 2.7 million and 3.4 million. Overall, despite variations within each cluster, this graph supports the assumption that stations within the same cluster tend to exhibit similar ridership patterns, which can serve as a basis for validation of estimates at new stations.

3.3. Passenger and Revenue Estimation

The Markov Chain Model is used as a methodological framework to project the number of passengers and estimate the daily revenue of rail-based transportation modes, especially at existing and new stations. This approach combines historical passenger movement data with the results of station clustering using the Fuzzy C-Means method, allowing new stations to be mapped based on their similarity to existing stations in terms of characteristics.

The first stage involves preparing the Transition Probability Matrix between existing stations, T^{hist} , which measures 13×13 (Table 2). This matrix describes the probability of passenger transfers from one origin station to a destination station, with a diagonal value of zero, indicating that there is no transfer to the same station.

Table 2. Matrix Transition of Existing Stations

Destination Station Origin Station	Station 1	Station 2	Station 3	...	Station 11	Station 12	Station 13
Station 1	0	0.0097	0.0183	...	0.1138	0.098	0.2069
Station 2	0.0143	0	0.015	...	0.1029	0.1133	0.1793
Station 3	0.0773	0.0211	0	...	0.0975	0.1429	0.1732
...	...	...	...	...	...	...	...
Station 11	0.1593	0.1066	0.0551	...	0	0.2476	0.0783
Station 12	0.0649	0.0651	0.0367	...	0.1166	0	0.0838
Station 13	0.2111	0.1417	0.0682	...	0.0408	0.1304	0

The second stage estimates the initial number of passengers at the new station by mapping it into clusters based on the average daily volume. The third stage forms a Proximity Matrix to measure the similarity between stations based on fuzzy membership degrees (Table 3).

Table 3. Proximity Matrix

Station Name	Station 1	Station 2	Station 3	...	New Station 5	New Station 6	New Station 7
Station 1	0	0.9997	0.0005	...	0.0003	0.001	0.0004
Station 2	0.9997	0	0.0005	...	0.0003	0.001	0.0004
Station 3	0.0005	0.0005	0	...	0.0216	0.7131	0.9438
...	...	...	...	...	...	...	...
New Station 5	0.0003	0.0003	0.0216	...	0	0.0365	0.0399
New Station 6	0.001	0.001	0.7131	...	0.0365	0	0.6948
New Station 7	0.0004	0.0004	0.9438	...	0.0399	0.6948	0

The fourth stage integrates and normalises the Transition Probability Matrix for existing and new stations, and then ensures that each row adds up to one to form the stochastic matrix (Table 4).

Table 4. Matrix Transition of Existing and New Station

Destination Station Origin Station	Station 1	Station 2	Station 3	...	New Station 5	New Station 6	New Station 7
Station 1	0	0.0096	0.0181	...	0	0.0001	0
Station 2	0.0142	0	0.0148	...	0	0.0001	0
Station 3	0.0493	0.0134	0	...	0.0018	0.0488	0.0646
...	...	...	...	...	...	...	...
New Station 5	0	0	0.0079	...	0	0.0648	0.0707
New Station 6	0.0001	0.0001	0.0281	...	0.0089	0	0.116
New Station 7	0.0001	0.0001	0.0438	...	0.0115	0.1363	0

Evaluation of the Markov chain model projection results can be seen through the convergence of the stationary distribution. The convergence graph is shown in the figure below.

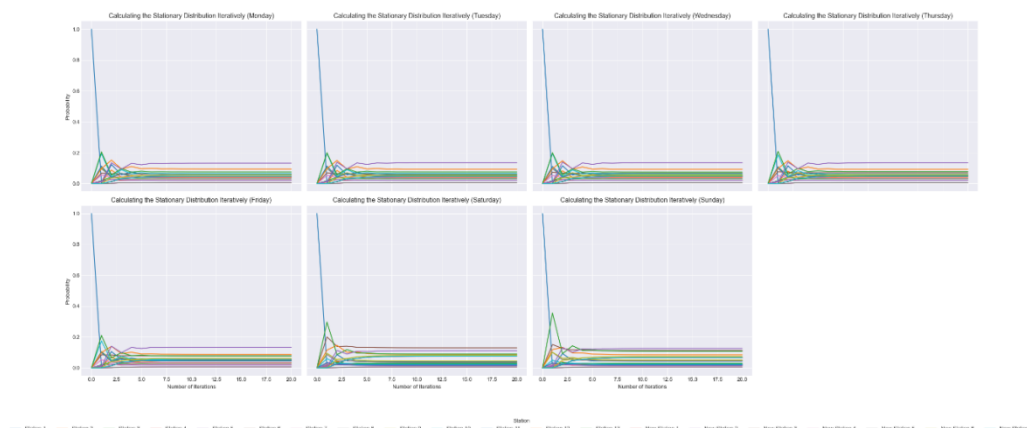


Figure 5. Calculating the Stationary Distribution Iteratively

The graph above, showing the probability curves for each day of the week, consistently indicates that after a small number of iterations (approximately 5-10 iterations), the curves for both old and new stations tend to flatten and approach stable values. This straightforward and rapid convergence is a strong indicator that the model has reached a stable state, where the probability of passenger presence at each station no longer changes significantly over time or with increasing iterations. The stability of these predictions indicates that the model is capable of producing consistent and reliable long-term probability estimates, thereby reducing its dependence on initial conditions. Once the stationary probabilities for each station are obtained, these values can be used as a solid basis for projecting actual passenger numbers and revenue, thus supporting the overall reliability of the projections.

The calculation of the number of passengers is generated from $P_{t_1} = P_t \cdot T$, where P_t is the initial estimate of the number of passengers for existing stations and new stations generated from stage 2, and T is the normalized matrix of transition probabilities from the fourth stage (Table 4).

An example of calculating the estimated number of passengers for Monday's data is as follows:

$$P_{t+1} = [11407 \quad 8313 \quad \cdots \quad 5759 \quad 5023]_{1 \times 20} \begin{bmatrix} 0 & \cdots & 0.0005 \\ \vdots & \ddots & \vdots \\ 0.00007 & \cdots & 0 \end{bmatrix}_{20 \times 20}$$

The estimated number of passengers on that day is:

$$P_{t+1} = [9306 \quad 6707 \quad 4806 \quad \cdots \quad 2919 \quad 7853 \quad 6688]_{1 \times 20}$$

Thus, the resulting matrix value is in the form of a 1×20 vector. The value of 9306 ... in the vector result represents the number of passengers at station 1 ... on Monday. This calculation will be done seven times, according to the number of days in a week.

Based on the transition matrix, the projected total estimated number of daily passengers on all rail-based transportation routes is shown in Table 5. It can be seen that the peak passenger volume occurs on Tuesday, with 179,516 passengers per day. In contrast, on weekends (Saturday and Sunday), there is a significant decrease, averaging 115,551 passengers per day. This pattern is consistent with the characteristics of commuters who have higher activity on weekdays.

Table 5. Ridership by Day

Day	Total Ridership
Monday	164179
Tuesday	179526
Wednesday	176157
Thursday	173630
Friday	173436
Saturday	123711
Sunday	107388

With the addition of new lines, there is a significant increase in the number of passengers at all stations. On weekdays, there is a 57% increase, and on holidays, the increase is even higher at 74%. Table 6 presents the results of the estimated number of daily passengers for each new station.

Table 6. Estimate Ridership for New Station

Station Name	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
New Station 1	6069	6548	6674	6732	7165	7884	6035
New Station 2	30909	34503	33564	32865	32390	18077	17793
New Station 3	5556	5999	6104	6156	6594	7237	5549
New Station 4	2809	2952	2867	2847	2756	2082	1773
New Station 5	2927	3072	2994	2965	2884	2222	1882
New Station 6	7842	8486	8614	8659	9191	9396	7319
New Station 7	6684	7200	7338	7407	7851	8811	6708

From Table 6 it can be seen that New Station 2 has the highest estimate, most likely due to its proximity to the business centre and tourist destinations.

To generate daily revenue projections, the fare per passenger is calculated based on the distance between existing stations and new stations. By multiplying the average fare per trip by the estimated number of daily passengers in Table 7.

Table 7. Travel Fare

Station Name	Station 1	Station 2	Station 3	...	New Station 5	New Station 6	New Station 7
Station 1		4,000	5,000	...	18000	19000	20000
Station 2	4,000		4,000	...	16000	17000	18000
Station 3	5,000	4,000		...	15000	16000	16000
...	...	...	...	...	...	...	...
New Station 5	18000	16000	15000	...		3000	3000
New Station 6	19000	17000	16000	...	3000		3000
New Station 7	20000	18000	16000	...	3000	3000	

Based on the estimated daily number of passengers that have been calculated and the average fare per passenger, the total estimated daily revenue for each day of the week is presented in Table 8.

Table 8. Total Revenue

Day	Total Revenue
Monday	Rp 1.117.583.000
Tuesday	Rp 1.216.164.000
Wednesday	Rp 1.190.097.000
Thursday	Rp 1.171.427.000
Friday	Rp 1.162.578.000
Saturday	Rp 813.463.000
Sunday	Rp 706.604.000

The estimated daily revenue results show a pattern that is almost parallel to passenger volume. The average weekday revenue is Rp 1.172 billion. Revenue drops sharply on weekends, resulting in a loss of Rp 760.25 million. The difference in revenue between weekdays and weekends suggests that the main revenue comes from commuter activity. This fluctuation suggests an opportunity to implement flexible fares or weekend-specific promotions to attract more passengers during periods of lower revenue.

To see the contribution of each station to total revenue, the total estimated daily revenue per new station is presented in Table 9.

Table 9. Estimate Revenue for New Station

Station Name	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
New Station 1	Rp 60.435.000	Rp 65.560.000	Rp 66.683.000	Rp 67.172.000	Rp 72.033.000	Rp 77.491.000	Rp 59.794.000
New Station 2	Rp 199.568.000	Rp 222.045.000	Rp 216.422.000	Rp 211.721.000	Rp 209.308.000	Rp 120.420.000	Rp 115.595.000
New Station 3	Rp 55.019.000	Rp 59.795.000	Rp 60.765.000	Rp 61.159.000	Rp 65.987.000	Rp 70.258.000	Rp 54.442.000
New Station 4	Rp 76.085.000	Rp 81.434.000	Rp 79.615.000	Rp 79.168.000	Rp 77.602.000	Rp 62.008.000	Rp 51.290.000
New Station 5	Rp 78.619.000	Rp 84.143.000	Rp 82.244.000	Rp 81.623.000	Rp 80.045.000	Rp 64.084.000	Rp 52.938.000
New Station 6	Rp 86.958.000	Rp 95.045.000	Rp 95.476.000	Rp 95.230.000	Rp 100.768.000	Rp 91.249.000	Rp 73.646.000
New Station 7	Rp 72.371.000	Rp 78.511.000	Rp 79.678.000	Rp 80.149.000	Rp 85.635.000	Rp 89.644.000	Rp 69.770.000

Of all the new stations, Station 2 remains the most prominent. This emphasises the station's role as a centre of business and tourism activities, which triggers higher service demand compared to other stations on the new line.

4. CONCLUSION

This study presents a predictive framework for estimating passenger volume and revenue on newly constructed rail transit lines using demographic clustering and Markov chain modeling. Dimensionality reduction via Principal Component Analysis (PCA) effectively compressed 29 demographic features into three principal components, preserving over 80% of the variance. The Fuzzy C-Means (FCM) algorithm produced well-defined clusters (FPC = 0.913), while the Markov Chain model achieved rapid convergence within 5–10 iterations, confirming stable long-term forecasting performance.

The model projects a 57% increase in weekday ridership and a 74% increase on weekends, with the highest estimated daily revenue reaching Rp 1.216 billion. These results are promising for informing investment feasibility, service planning, and Transit-Oriented Development (TOD) around new stations.

However, this study has certain limitations. First, the assumption that new stations exhibit travel behavior similar to that of demographically clustered existing stations may not fully capture real-world variations in land use, accessibility, or future mobility shifts. Second, the model relies on static demographic data and does not account for temporal changes or external factors such as special events, economic shocks, or policy changes. Moreover, fare assumptions were based on

average distances without elasticity modeling, which could affect revenue sensitivity in real scenarios.

Future research can address these limitations by integrating real-time mobility data (e.g., from mobile phone traces or smart card systems), testing the model in different urban contexts, or incorporating dynamic demographic projections. Additionally, combining this model with other machine learning techniques (e.g., time-series forecasting, agent-based simulations) and multimodal transport data may further improve accuracy and adaptability for large-scale transportation planning.

REFERENCES

- [1] I. Dostál and V. Adamec, "Transport and its Role in the Society," *Transactions on Transport Sciences*, vol. 4, no. 2, pp. 43–56, Jun. 2011, doi: 10.2478/V10158-011-0006-8.
- [2] Badan Pusat Statistik Indonesia, "Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk, Kepadatan Penduduk, Rasio Jenis Kelamin Penduduk Menurut Provinsi, 2024," Badan Pusat Statistik Indonesia. Accessed: Jul. 02, 2025. [Online]. Available: <https://www.bps.go.id/id/statistics-table/3/V1ZSbFRUY3ITbFpEYTNsVWNGcDZjek53YkhsNFFUMDkjMw==/penduduk--laju-pertumbuhan-penduduk--distribusi-persentase-penduduk--kepadatan-penduduk--rasio-jenis-kelamin-penduduk-menurut-provinsi--2024.html?year=2024>
- [3] Badan Pusat Statistik Provinsi Dki Jakarta, "Jumlah Kendaraan Bermotor Menurut Jenis Kendaraan (unit) di Provinsi DKI Jakarta," Badan Pusat Statistik Provinsi Dki Jakarta. Accessed: Jul. 02, 2025. [Online]. Available: <https://jakarta.bps.go.id/id/statistics-table/2/Nzg2IzI=/jumlah-kendaraan-bermotor-menurut-jenis-kendaraan-unit-di-provinsi-dki-jakarta.html>
- [4] E. Ratnawati, "MRT as an Alternative Transportation Solution which is environmentally friendly and traffic jam free in Jakarta," *IOP Conf Ser Earth Environ Sci*, vol. 819, no. 1, Jul. 2021, doi: 10.1088/1755-1315/819/1/012040.
- [5] A. N. Hakimi Ibrahim and M. N. Borhan, "The Interrelationship Between Perceived Quality, Perceived Value and User Satisfaction Towards Behavioral Intention in Public Transportation: A Review of the Evidence," *Int J Adv Sci Eng Inf Technol*, vol. 10, no. 5, pp. 2048–2056, 2020, doi: 10.18517/IJASEIT.10.5.12818.
- [6] D. Bandarsyah, A. Abdulhadi, and S. Sulaeman, "The Development and Economic Impact of Railway in Batavia, 1873-1930," *Paramita: Historical Studies Journal*, vol. 32, no. 2, pp. 159–170, Sep. 2022, doi: 10.15294/PARAMITA.V32I2.31683.
- [7] K. M. Kasikoen, S. H. Mukti, A. Fauzi, Suprajaka, and E. Martini, "Dynamic Modeling Impacts of Inter-Urban Railway Transportation on Regional Development: A Case Study of Sub-Urban Greater Jakarta, Indonesia," *International Journal of Sustainable Development and Planning*, vol. 18, no. 11, pp. 3419–3428, Nov. 2023, doi: 10.18280/IJSDP.181107.
- [8] V. Padmanabhan and S. Nirmala, "PREDICTABILITY OF TOTAL REVENUE IN GLOBAL LARGE COMPANIES FROM FINANCIAL AND ECONOMIC PARAMETERS."
- [9] C. B. Gardner, S. D. Nielsen, M. Eltvéd, T. K. Rasmussen, O. A. Nielsen, and B. F. Nielsen, "Calculating conditional passenger travel time distributions in mixed schedule- and frequency-based public transport networks using Markov chains," *Transportation Research Part B: Methodological*, vol. 152, pp. 1–17, Oct. 2021, doi: 10.1016/J.TRB.2021.06.020.
- [10] T. Spaninger, B. Büchel, and F. Corman, "Train Delay Predictions Using Markov Chains Based on Process Time Deviations and Elastic State Boundaries," *Mathematics 2023, Vol. 11, Page 839*, vol. 11, no. 4, p. 839, Feb. 2023, doi: 10.3390/MATH11040839.
- [11] M. S. Alam and S. Paul, "A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh," *J Appl Stat*, vol. 47, no. 8, pp. 1460–1481, Jun. 2020, doi: 10.1080/02664763.2019.1675606.

- [12] Badan Pusat Statistik Provinsi DKI Jakarta, "Statistik Transportasi Provinsi DKI Jakarta 2024 ," 5, May 2025. Accessed: Jul. 02, 2025. [Online]. Available: <https://jakarta.bps.go.id/id/publication/2025/05/23/7454766f878f35a5fa126dc5/statistik-transportasi-provinsi-dki-jakarta-2024.html>
- [13] R. I. Whitfield and A. H. B. Duffy, "Extended revenue forecasting within a service industry," *Int J Prod Econ*, vol. 141, no. 2, pp. 505–518, Feb. 2013, doi: 10.1016/J.IJPE.2011.11.015.
- [14] Y. Zhou, L. Wang, R. Zhong, and Y. Tan, "A Markov Chain Based Demand Prediction Model for Stations in Bike Sharing Systems," 2018, doi: 10.1155/2018/8028714.
- [15] A. P. Joshi and B. V. Patel, "Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Oriental journal of computer science and technology*, vol. 13, no. 0203, pp. 78–81, Jan. 2021, doi: 10.13005/OJCST13.0203.03.
- [16] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput Geosci*, vol. 10, no. 2–3, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.
- [17] D. Krasnov, D. Davis, K. Malott, Y. Chen, X. Shi, and A. Wong, "Fuzzy C-Means Clustering: A Review of Applications in Breast Cancer Detection," *Entropy 2023, Vol. 25, Page 1021*, vol. 25, no. 7, p. 1021, Jul. 2023, doi: 10.3390/E25071021.
- [18] O. N. Purba, D. N. Sitompul, T. H. Harahap, S. R. D. Saragih, and R. F. Siregar, "Application of Fuzzy C-Means Algorithm for Clustering Customers," *Hanif Journal of Information Systems*, vol. 1, no. 1, pp. 26–36, Aug. 2023, doi: 10.56211/hanif.v1i1.8.
- [19] H. Abdellahoum, N. Mokhtari, A. Brahimi, and A. Boukra, "CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach," *Expert Syst Appl*, vol. 166, p. 114063, Mar. 2021, doi: 10.1016/J.ESWA.2020.114063.
- [20] J. I. . Agbinya, "Applied data analytics : principles and applications," p. 336, 2020, Accessed: Jul. 02, 2025. [Online]. Available: https://www.researchgate.net/publication/343810250_Markov_Chain_and_its_Applications_an_Introduction
- [21] D. W. Hogg and D. Foreman-Mackey, "Data Analysis Recipes: Using Markov Chain Monte Carlo*," *Astrophys J Suppl Ser*, vol. 236, no. 1, p. 11, May 2018, doi: 10.3847/1538-4365/AAB76E.
- [22] L. Abd. Razid, "Prediksi Curah Hujan Menggunakan Rantai Markov Untuk Perencanaan Musim Tanam Tanaman Pangan," 2019.
- [23] J. Guevara, R. Hirata, and S. Canu, "Cross product kernels for fuzzy set similarity," *IEEE International Conference on Fuzzy Systems*, Aug. 2017, doi: 10.1109/FUZZ-IEEE.2017.8015459;PAGE:STRING:ARTICLE/CHAPTER.
- [24] N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Comput Sci*, vol. 163, pp. 292–299, 2019, doi: 10.1016/J.PROCS.2019.12.111.
- [25] I. T. J. Springer, "Principal Component Analysis, Second Edition".
- [26] S. Messaoud, A. Bradai, S. H. R. Bukhari, P. T. A. Quang, O. Ben Ahmed, and M. Atri, "A survey on machine learning in Internet of Things: Algorithms, strategies, and applications," *Internet of Things*, vol. 12, p. 100314, Dec. 2020, doi: 10.1016/J.IOT.2020.100314.
- [27] J. Petrus, Ermatita, and Sukemi, "Soft and Hard Clustering for Abstract Scientific Paper in Indonesian ," 2019 *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2019.