Clustering Commuter Behavior based on Automated Fare Collection (AFC)

Dwijoko Purbohadi
Department of Information Technology, Muhammadiyah University of Yogyakarta, Indonesia
dwijoko.purbohadi@umy.ac.id (corresponding author)

Laila Marifatul Azizah
Department of Information Technology, Muhammadiyah University of Yogyakarta, Indonesia
laila.m.azizah@umy.ac.id

Lilis Kurniasari
Department of Electrical Engineering, Nahdlatul Ulama University of Yogyakarta, Indonesia
lilis@unu-jogja.ac.id

Novi Diah Wulandari
Department of Management, Nahdlatul Ulama University of Yogyakarta, Indonesia
d.wulandari@unu-jogja.ac.id

Nurna Pratiwi
Department of Management, Nahdlatul Ulama University of Yogyakarta, Indonesia
nurnapratiwi@unu-jogja.ac.id

Puji Hastuti
Department of Information Technology, State Polytechnic of Jember, Indonesia
pujihastuti@polije.ac.id

Received: 3 September 2024 | Revised: 12 November 2024 and 18 November 2024 | Accepted: 21 November 2024

 DOI: https://doi.org/10.48084/etasr.8899

ABSTRACT

This paper examines the application of the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method to cluster Automated Fare Collection (AFC) transaction data from train travelers in Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek) in Indonesia. To enhance the clustering process, the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction and the DenseClus library are employed. In this study, different combinations of hyperparameters are used to identify the optimal configuration for producing distinct clusters with a high concentration and noticeable distinction. The results demonstrate that the utilization of HDBSCAN on UMAP-reduced data effectively, discerning unique trip patterns and emphasizing notable disparities in travel distance, time, and length among various clusters. The UMAP intersection method showed notable efficacy in maintaining the local structure of the data, resulting in the development of distinct and meaningful clusters. In addition, categorical data were transformed into numerical formats using hashing techniques, efficiently tackling the difficulties posed by a high number of categories and assuring efficient data processing. The results reveal vital insights into the application of density-based clustering to intricate transportation data, with major implications for enhancing route planning and capacity management for Jabodetabek commuters.

Keywords-automated fare collection; public transportation; clustering algorithms; UMAP embedding; HDBSCAN

I. INTRODUCTION

Major Indonesian cities continue to suffer significantly from traffic congestion, especially in metropolitan regions. According to the Jabodetabek Transportation Management Agency, the economic consequences of traffic congestion in Jakarta and its neighboring areas amount to almost IDR 71.4 trillion per year [1-3]. Jabodetabek, which is a combined area of the cities of Jakarta, Bogor, Depok, Tangerang, and Bekasi, faces significant challenges. Jabodetabek needs help managing traffic flow due to the high population density and extensive commuter movement. The commuter train offers mobility, reliability, and affordability for people from different social backgrounds [3, 4]. PT. Kereta Api Indonesia (PT. KAI) has become the primary transit choice. This company provides transportation services, one of which is commuting (KAI commuter), and according to PT. KAI 2020 annual report on commuter services, the number of passengers was 424,532 per day on 940 routes [5]. Commuter services are expected to make a significant contribution to reducing congestion in the Jabodetabek area.

The main obstacle for commuter train management is to provide the most efficient service to commuters. With the expansion of the commuter train network and routes, management must adopt more effective operational practices. Indeed, passengers need to transfer between distinct service lines, and previous studies suggest that coordinated operations across service lines can improve service quality and reduce delays [6]. The increasing volume of data presents two significant challenges in studying AFC travel records. These are the extraction of relevant features and the selection of a data mining algorithm. The standard data mining techniques used include classification, clustering, Decision Trees (DTs), and association rules. To identify travelers with different mobility patterns, clustering is used [7]. Numerous research projects have utilized AFC data to evaluate traveler habits, yet a significant challenge remains in understanding commuter behavior through AFC clustering data analysis [7, 8]. Studies have employed various methods: K-means clustering to identify logistics firm staff and robbers, Partial Area Clustering (PAC) to optimize transportation networks [9], and a two-level Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to extract origin-destination pairings and habitual travel times [10]. Among these, HDBSCAN was found to be superior to K-means and DBSCAN for analyzing noisy, complex, and high-dimensional AFC data [11]. Unlike K-means, which assumes spherical and homogeneous clusters, HDBSCAN can manage clusters of varying shapes and sizes and automatically identifies outliers without operator intervention [12, 13]. This method is particularly effective for exploratory research, where cluster structures are unknown, offering flexibility in analyzing data at different granularities. The AFC systems provide detailed commuter travel itineraries, aiding in the study of travel patterns and passenger transfers [14, 15]. Despite the vast amount of data collected, challenges persist in extracting detailed route information [16, 17]. Techniques, such as agglomerative hierarchical clustering and database-based approaches, have been developed to address these issues [18, 19]. These methods aim to enhance the understanding of passenger behavior and optimize the use of AFC data, though, comprehensive solutions for analyzing commuter behavior in train networks are still emerging.

This study aims to evaluate whether the HDBSCAN algorithm can be applied to build a clustering model using large-scale AFC transaction data. Given the complexity and potential noise in travel data, the research seeks to determine how effectively HDBSCAN can identify relevant travel patterns compared to other clustering methods. The results are expected to contribute to improved transportation planning and management, particularly in the Jabodetabek area, and provide valuable insights for other metropolitan cities. The dataset used in this study was obtained from PT Kereta Commuter Indonesia's AFC system, specifically the KRL Commuter line, during the period spanning from January to March 2020. The findings demonstrated that the HDBSCAN algorithm can categorize the AFC transaction data of KRL Jabodetabek passengers into significant clusters, with the use of UMAP dimensionality reduction techniques and the DenseClus library.

II. METHODOLOGY

This research investigated the use of HDBSCAN for clustering AFC transaction data of light rail passengers in Greater Jakarta and aimed to identify the most effective clustering strategy. Using the DenseClus library, the study combined the UMAP for dimensionality reduction with the HDBSCAN for clustering, targeting both categorical and numerical data. Feature selection was performed to identify the key attributes, and HDBSCAN was chosen for its ability to handle diverse cluster densities and high-dimensional data. The integration of UMAP and HDBSCAN through DenseClus automated the clustering process, effectively resulting in homogeneous clusters, as depicted in Figure 1.

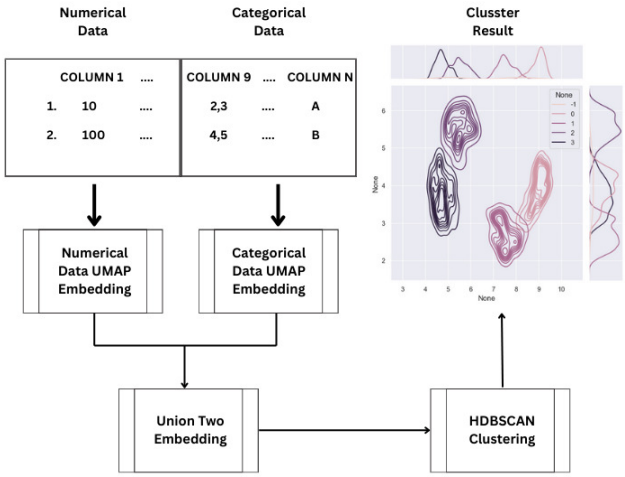

Fig. 1. DenseClus flowchart.

A. Data Gathering

This study considers the AFC KRL track Jakarta, Bogor, Depok, Tangerang, and Banten (Jabodetabek) regions. The data are private and are derived from the Kereta Commuter Indonesia (KCI) operator for a commuter line in Indonesia with masked id. Additionally, other confidential data from January

to March 2020, before the COVID-19 pandemic, are also used. Around 50 million KRL Jabodetabek travelers used cards at stations, according to AFC data. After initial data extraction, 11 AFC card properties were found and listed in Table I.

TABLE I. DATA PROPERTIES

No	Feature	Description	Data Type
1	c_csard	Card number	String
2	d_gate_in	Entry time	Datetime
3	c_station_in	Boarding station	String
4	d_gate_out	Time out	Datetime
5	c_station_out	Destination station	String
6	c_status	Transaction status	String
7	c_desc	Transaction type code	String
8	i_card_type	Card type code	Integer
9	m_balance_before	Balance before deposit	Integer
10	m_deduct	Deposit	Integer
11	m_balance	Balance after deposit	Integer

The AFC data encompass the origin and destination stations, entry and exit gates, card ID, and card type. Regarding the passengers involved, office workers, civil servants, retailers, businesspeople, and others conducted many commuter journeys from January to February 2020. This study analyzes comprehensive and long-term commuter travel patterns to help public transportation operators improve services, especially at transit station nodes with passenger congestion.

B. Data Preparation

The raw AFC data are quite large, with a file size reaching 35GB. The data are stored in a Comma-Separated Values (CSV) format, as shown in Table II. To extract the necessary information, specific columns or data fields must be selected, as detailed in Table I. During the loading of the raw data, a data cleaning process is simultaneously conducted, tailored to the data types of each column. Anomalies are then filtered out by identifying irrational or contradictory entries, particularly focusing on eliminating transactions with multiple entries and exits at the same station.

TABLE II. RAW AFC DATA

	A	B	C	D	E	F	H	I	J	K	L
1		c_station_in	d_gate_in	c_station_out	d_gate_out	i_card_type	m_balance_before	m_deduct	m_balance	c_status	c_desc
2	0	CW	31/03/00 12.36	BJD	29/02/20 14.32	8	0	0	0	S	NT
3	1	TNT	30/03/18 14.39	PSMB	05/01/20 14.41	4	0	0	0	S	NP
4	2	TNT	30/03/18 14.39	PSMB	05/01/20 15.12	4	0	0	0	S	NP
5	3	CKI	15/02/19 08.57	CKI	02/01/20 11.46	4	0	0	0	S	NP
6	4	BKS	26/02/19 11.32	BKS	13/01/20 07.33	4	0	0	0	S	NP
7	5	THB	01/03/19 11.45	CKI	05/02/20 07.34	4	0	0	0	S	NP
8	6	THB	01/03/19 11.45	CKI	05/02/20 07.34	4	0	0	0	S	NP
9	7	CKI	31/03/19 21.27	BKS	15/01/20 07.54	4	0	0	0	S	NP
10	8	DPB	04/04/19 09.18	DP	28/01/20 22.01	4	0	0	0	S	NP
11	9	CTA	11/04/19 12.04	BKS	13/01/20 07.32	4	0	0	0	S	NP
12	10	CTA	11/04/19 12.04	BKS	13/01/20 07.33	4	0	0	0	S	NP

Following this, the AFC transaction data are analyzed for peak commuter density from Monday to Friday, with a focus on data from February 3 to 10, 2020, and were stabilized at 1 million transactions per day. The dataset is then filtered and sorted by date and time, and new columns are created to isolate specific time components for detailed analysis. Trip chains are generated to link origin and destination stations, providing insights into commuter behavior, and the dataset is enhanced with temporal and journey-specific details. The AFC data contain confidential or sensitive information related to security, which necessitates that the information be concealed or anonymized. It is crucial to recognize that data mining involves ethical issues and privacy concerns. Organizations must manage data responsibly, adhering to legal and ethical standards [20]. Finally, the frequency of unique trip chains is analyzed to identify dominant travel patterns and less common routes, which informs transportation planning and enhances system efficiency and passenger satisfaction.

C. Commuter Selection

The summary statistics are the details of passenger travel behavior for one week in February 2020. Distinct travelers, who made two or more daily travels, were chosen. Most unique travelers made round trips throughout the week. According to these data, Understanding daily passenger behavior fluctuations is crucial. Round trip consistency over weekdays suggests constant transportation demand, whereas variances may signal the necessity for service adjustments as passenger needs shift at the end of the week. Figure 2 displays hourly statistics for the utilization of daily travel cards. There is an increase in card usage between 7-9 am and 5-7 pm on weekdays. This card usage means that most of the card transactions during these hours concern a work-related travel.

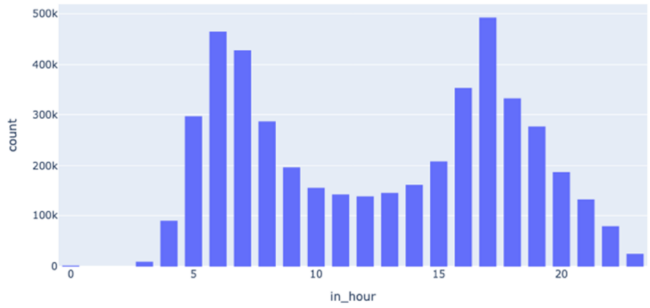

Fig. 2. Passengers per hour.

1) HDBSCAN Hyperparameter Tunning

HDBSCAN clustering was optimized by examining the min_samples and min_cluster_size hyperparameters. Using Randomized Search with Density-Based Clustering Validation (DBCV) as the evaluation metric, various combinations were tested to find the best clustering performance. The search included parameters, like min_samples, 5 to 90, min_cluster_size, 100 to 1000, cluster selection methods, and distance metrics. The RandomizedSearchCV was applied over 50 iterations to identify the optimal configuration, as displayed in Table III.

TABLE III. HYPERPARAMETER TUNNING

No	Best Parameter	Score
1	min_samples	70
2	min_cluster_size	200
3	metric	Euclidean
4	cluster_selection_method	leaf

III. RESULTS AND DISCUSSION

This study evaluates the preservation of data structure in UMAP dimensionality reduction using trust and continuity metrics, with the implementation of DenseClus to stabilize UMAP's convergence. Trustworthiness and continuity are two relevant metrics for evaluating the retention of data structure in dimensionality reduction techniques, like UMAP. This research aims to demonstrate the implementation and application of trustworthiness and continuity metrics as additional validation methods using the DenseClus package to ensure stable UMAP convergence. By understanding and applying these additional validation methods, a deeper understanding of UMAP's performance can be obtained, ensuring that the technique accurately captures the underlying data structure.

In Figure 3, the plot illustrates the trustworthiness score as a function of K values for numerical and categorical data in UMAP embeddings. The trustworthiness score measures how well the local structure of the original data is preserved in the UMAP-generated embedding. The value of K is set up to 50, determining the number of the nearest neighbors considered in the trustworthiness calculation. The UMAP algorithm uses the parameter K (number of neighbors) to construct a graphical representation of the data, which is crucial for the dimensionality reduction process. When K is low, for instance, 5 or 10, the algorithm primarily focuses on the local relationships between the data points. However, as K increases, UMAP begins to capture more global patterns within the data. Research indicates that with K around 50, UMAP effectively maintains a balance between the local details and global patterns without overemphasizing the distant points [21-23]. This ensures that the overall structure of the data is preserved after dimensionality reduction.

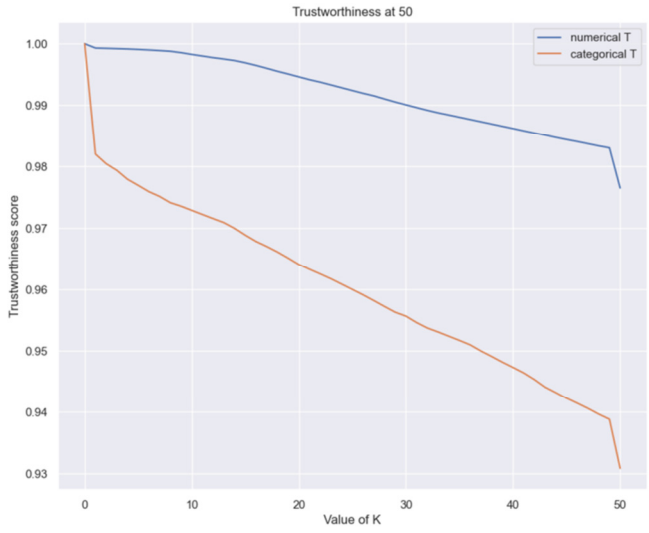

Fig. 3. Trustworthiness score versus K values.

The X-axis displays the value of K, which ranges from 0 to 50, while the Y-axis shows the trustworthiness score, which ranges from 0 to 1. Higher scores indicate that the UMAP embedding is more reliable in maintaining the local structure of the original data. The blue line represents the trustworthiness score for numerical data across different K values, while the orange line represents the trustworthiness score for categorical data. At smaller K values, the trustworthiness score for both data types, numerical and categorical, approaches 1, indicating that the local structure of the original data is well-preserved in the UMAP embedding. However, as K increases, the trustworthiness score tends to decrease, suggesting that at higher K values, the UMAP embedding may be less reliable in maintaining the nearest-neighbor relationships of the original data. The trustworthiness score for the numerical data consistently remains higher than that for the categorical data across most K values, indicating that UMAP embeddings are more effective at preserving the local structure for the numerical data compared to the categorical data.

A. *Numerical data*

The diagram in Figure 4 presents the UMAP embedding and HDBSCAN clustering results on the numerical data. The X and Y axes correspond to the two dimensions of the UMAP embedding. The main plot shows the density of HDBSCAN clusters in color and contour lines, with -1 indicating noisy data points. The marginal plots at the top and right show the data distribution along the axes, separately. These visualizations exhibit clusters of numerical data with different densities, highlighting the concentration of data points and the spread of noise.

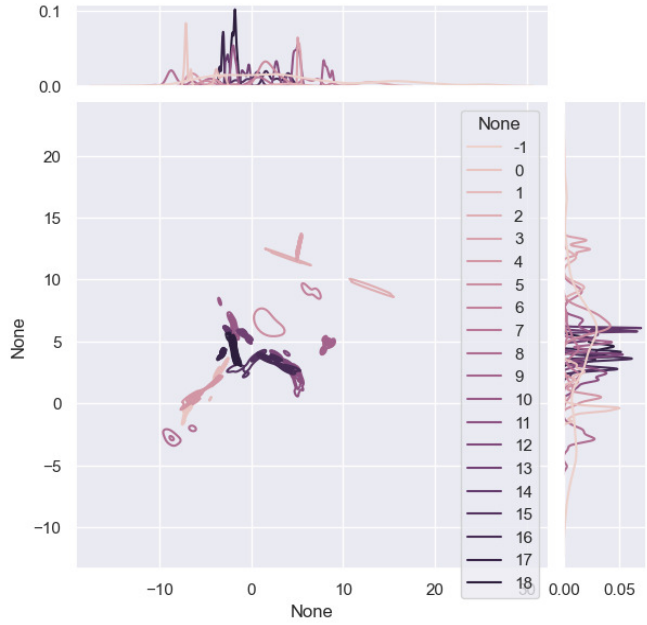

Fig. 4. UMAP embedding and HDBSCAN clustering on numerical data.

Figure 5 displays a condensed tree visualization generated by the HDBSCAN algorithm, illustrating the hierarchical structure of the clusters at varying density levels. The Y-axis represents the λ value, with lower values indicating denser clusters, while the horizontal lines show the number of data points in each cluster, and a purple-to-yellow color gradient reflects cluster magnitude. The vertical lines depict cluster divisions at specific density thresholds, and the red ellipses

highlight the significant clusters within the hierarchy. The analysis reveals that larger clusters tend to divide at lower densities, while smaller clusters merge at various thresholds, providing valuable insights into the data structure and demonstrating HDBSCAN's effectiveness in identifying meaningful clusters. In this study, all non-numerical columns were removed from the dataset, leaving only numerical features, with a "segment" column containing HDBSCAN cluster labels. The mean and median of each numerical characteristic within each cluster were calculated, revealing significant disparities. The analysis focused on features, like geographical coordinates, travel distance, time, and trip duration. From the calculation, it was found that the highest average trip distance was 48.664171 km and the longest average travel time was 121.989831 minutes, while the shortest average trip distance was 29.700604 km and the shortest travel time was 107.535836 minutes. These findings highlight the effectiveness of the HDBSCAN algorithm in identifying meaningful clusters, providing valuable insights into the data distribution and patterns within each cluster.

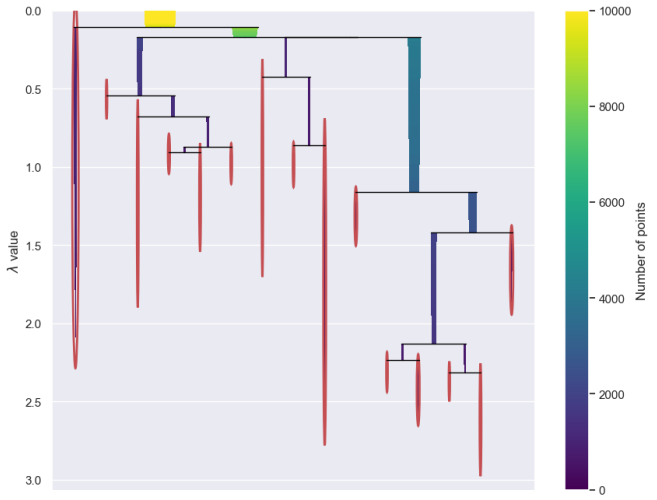

Fig. 5. Tree condensation numerical clustering.

B. Categorical Data

A data frame with just categorical features was obtained from the primary dataset using the extract categorical function. The categorical data were embedded utilizing improved UMAP parameters for memory, economy, and reproducibility. During the investigation, one categorical characteristic had a maximum value of 723 above the 25 criteria. Hashing with a hash function, converted category features into numbers. The HDBSCAN algorithm clustered the embeddings, revealing nine categorical data clusters. The results indicate that the combination of UMAP and HDBSCAN successfully lowers the dimensionality of categorical data and clusters them into meaningful groups, yielding reliable and reproducible outcomes. This investigation emphasizes the significance of employing hashing techniques to tackle the issue of high category counts in categorical features. It also demonstrates the efficiency of the HDBSCAN algorithm in clustering data that have been decreased in dimensions using UMAP. Figure 6 shows a combined plot of the UMAP embedding results for categorical data clustered using the HDBSCAN algorithm. The contour lines illustrate the density of the data points within each cluster, while the X and Y axes represent the two dimensions of the resulting embedding. By plotting marginal plots along the X and Y axes, further insights can be provided into the distribution of the data in each dimension. This visualization demonstrates that the UMAP and HDBSCAN could decrease the dimensionality of categorical data. Furthermore, the UMAP and HDBSCAN categorize the data into significant clusters, facilitating a visual depiction that enables a more comprehensive analysis of the data's structure.

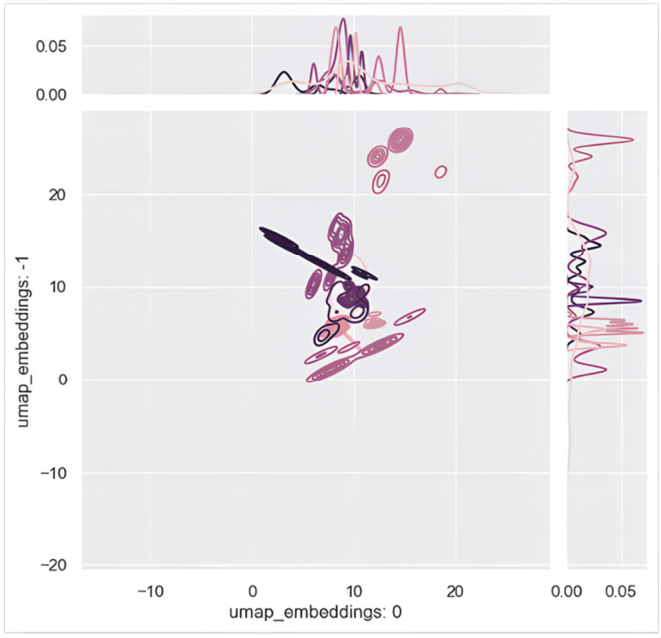

Fig. 6. UMAP embedding and HDBSCAN clustering on categorical data.

C. DenseClus for Numerical and Categorical Data

In this investigation, the HDBSCAN algorithm was used to assess the impact of four UMAP combination methods - intersection, union, contrast, and intersection union mapper - on clustering outcomes. The HDBSCAN parameters were set with a minimum cluster size of 200, a minimum sample count of 70, and employed the Euclidean metric and EoM technique for cluster selection. Each UMAP combination method was applied sequentially, followed by the clustering process and evaluation to determine the number of clusters generated. The analysis, as observed in Figure 7, indicated that the intersection and union methods delivered superior clustering performance, with coverage values of 0.9512 and 0.9083, respectively, and high Calinski-Harabasz scores of 6742.023 and 6810.469, suggesting well-defined and clearly separated clusters. Figure 8 displays the clustering results of the AFC KRL data using the UMAP intersection method combined with the HDBSCAN algorithm, showing the identification of significant clusters with varying distribution and density. The well-defined clusters, particularly on the right side of the image, exhibit high density, indicating the success of this method in preserving the local structure of the data. Conversely, the more dispersed

clusters on the left suggest the presence of heterogeneity within the dataset, with more significant variation in data characteristics. The presence of noise data, which are not included in any cluster, highlights the method's sensitivity in separating data that do not conform to the dominant patterns, which is an essential aspect for further analysis. Overall, the intersection method proves effective in identifying significant clusters with clear structures while maintaining the flexibility to capture variations within the data.

```
Running combination method: intersection
Max of 723 is greater than threshold 25
Hashing categorical features
Coverage 0.9512
Calinski-Harabasz Score: 6742.023141930125
Number of clusters: 20
------------------------------
Running combination method: union
Max of 723 is greater than threshold 25
Hashing categorical features
Coverage 0.9083
Calinski-Harabasz Score: 6810.468857751236
Number of clusters: 23
------------------------------
Running combination method: contrast
Max of 723 is greater than threshold 25
Hashing categorical features
Coverage 0.8013
Calinski-Harabasz Score: 1246.1560997847548
Number of clusters: 16
------------------------------
Running combination method: intersection_union_mapper
Max of 723 is greater than threshold 25
Hashing categorical features
Coverage 0.8959
Calinski-Harabasz Score: 2973.4526848469573
Number of clusters: 8
------------------------------
```

Fig. 7. UMAP combination methods.

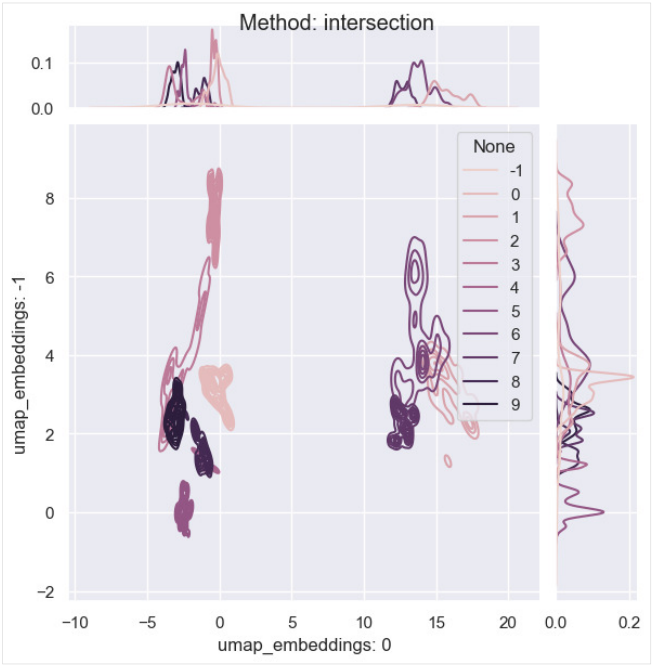

Fig. 8. Intersection clusterring.

Figure 9 presents a density distribution plot illustrating the multimodal distribution of the data groups, with distinct peaks showing variations in density and data distribution. Some groups have concentrated data at specific values, while others are more dispersed, with overlapping curves indicating shared values or transitions between the groups. This complexity suggests multiple clusters or subgroups within the data. An analysis of the KRL travel data reveals notable commuter patterns across clusters, with travel distances ranging from 23.7 km to 58.1 km and varied departure times between 4:00 and 5:30 am. The travel durations span from 107 to 126 minutes, influenced by the route length. Key routes, like Bogor to Jakarta, and critical stations, such as Tigaraksa, Bogor, Tanah Abang, and Jakarta Kota, are highlighted, emphasizing their role in the daily travel patterns. These findings are crucial for optimizing KRL services and effectively managing peak-hour passenger flows.

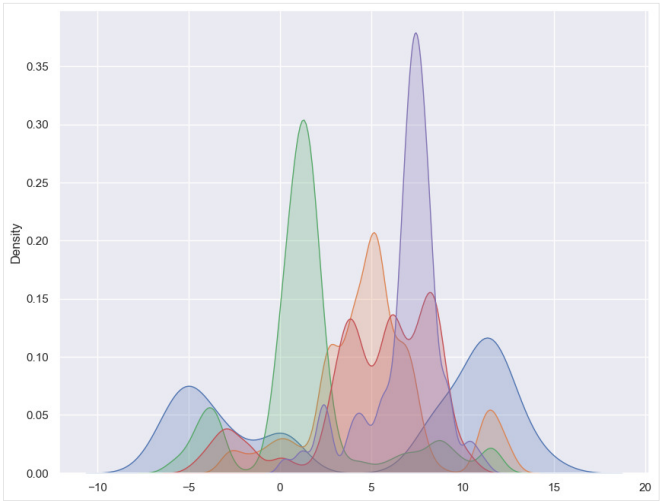

Fig. 9. Density distribution plot.

IV. CONCLUSION

This study examines the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm for clustering Automated Fare Collection (AFC) transaction data of the train passengers in Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek commuter), Indonesia, by using the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction technique and the DenseClus library. Various combinations of hyperparameters were tested to find the best settings that produce clusters with strong density and precise separation. The HDBSCAN was applied to the UMAP-reduced data and successfully identified distinct travel patterns among Jabodetabek commuter train passengers. These clusters exhibited significant differences in the distance, time, and duration of travel. The UMAP crossover method successfully maintained the data structure and produced well-defined clusters. A hashing technique was also deployed to convert the categorical data into a numerical form so that the data could be efficiently processed by machine learning algorithms. This study provides insights into how density-based clustering techniques, such as HDBSCAN, can

be applied on complex commuter train travel data. The findings are essential for planning and managing commuter train transportation networks so that the routes and capacity can be optimized to better meet passenger needs.

REFERENCES

[1] "Kerugian Ekonomi Akibat Macet Jabodetabek Capai Rp71,4 T." [Online]. Available: https://www.cnnindonesia.com/ekonomi/20210428120006-92-635840/kerugian-ekonomi-akibat-macet-jabodetabek-capai-rp714-t.

[2] A. M. H. Sitorus, "Sistem Transportasi Terintegrasi di DKI Jakarta: Analisis Transformasi Berkeadilan Sosial," *Jurnal Sosiologi Andalas*, vol. 8, no. 1, pp. 31–41, Apr. 2022, https://doi.org/10.25077/jsa.8.1.31-41.2022.

[3] A. I. Wiyogo, S. Budi, and H. Toba, "Ekstraksi Perilaku Komuter Pada Commuter Line Menggunakan Rule-Based Machine Learning," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 1, pp. 154–166, Apr. 2023, https://doi.org/10.28932/jutisi.v9i1.6133.

[4] J. Ning, Q. Peng, Y. Zhu, Y. Jiang, and O. A. Nielsen, "A Bi-objective optimization model for the last train timetabling problem," *Journal of Rail Transport Planning & Management*, vol. 23, Sep. 2022, Art. no. 100333, https://doi.org/10.1016/j.jrtpm.2022.100333.

[5] *Annual report 2020: Spinning the Limit to Win from Pandemic.* Jakarta, Indonesia: PT KAI Commuter, 2021.

[6] X. Guo, H. Sun, J. Wu, J. Jin, J. Zhou, and Z. Gao, "Multiperiod-based timetable optimization for metro transit networks," *Transportation Research Part B: Methodological*, vol. 96, pp. 46–67, Feb. 2017, https://doi.org/10.1016/j.trb.2016.11.005.

[7] Y. Chen, Y. Zhao, and K. L. Tsui, "Clustering-based Travel Pattern Recognition in Rail Transportation System Using Automated Fare Collection Data," in *2019 Prognostics and System Health Management Conference*, Qingdao, China, 2019, pp. 1–7, https://doi.org/10.1109/PHM-Qingdao46334.2019.8943009.

[8] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017, https://doi.org/10.1109/TITS.2017.2679179.

[9] J. Pei, K. Zhong, J. Li, and Z. Yu, "PAC: Partial Area Clustering for Re-Adjusting the Layout of Traffic Stations in City's Public Transport," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 1251–1260, Jan. 2023, https://doi.org/10.1109/TITS.2022.3179024.

[10] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger Segmentation Using Smart Card Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015, https://doi.org/10.1109/TITS.2014.2368998.

[11] T. H. Tran, T. D. Cao, and T. T. H. Tran, "HDBSCAN: Evaluating the Performance of Hierarchical Clustering for Big Data," in *Soft Computing: Biomedical and Related Applications*, N. H. Phuong and V. Kreinovich, Eds. Cham, Switzerland: Springer International Publishing, 2021, pp. 273–283.

[12] G. Stewart and M. Al-Khassaweneh, "An Implementation of the HDBSCAN* Clustering Algorithm," *Applied Sciences*, vol. 12, no. 5, Mar. 2022, Art. no. 2405, https://doi.org/10.3390/app12052405.

[13] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach," *Journal of Marine Science and Engineering*, vol. 9, no. 6, Jun. 2021, Art. no. 566, https://doi.org/10.3390/jmse9060566.

[14] X. Guo, D. Z. W. Wang, J. Wu, H. Sun, and L. Zhou, "Mining commuting behavior of urban rail transit network by using association rules," *Physica A: Statistical Mechanics and its Applications*, vol. 559, Dec. 2020, Art. no. 125094, https://doi.org/10.1016/j.physa.2020.125094.

[15] K. Lu, J. Liu, X. Zhou, and B. Han, "A Review of Big Data Applications in Urban Transit Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2535–2552, May 2021, https://doi.org/10.1109/TITS.2020.2973365.

[16] X. Wu, H. Dong, S. Gao, W. Li, and Q. Zhang, "Extracting Metro Passengers' Route Choice via AFC Data Utilizing Gaussian Mixture Clustering," in *2018 21st International Conference on Intelligent Transportation Systems*, Maui, HI, USA, 2018, pp. 1933–1938, https://doi.org/10.1109/ITSC.2018.8569403.

[17] Y. Sun and R. Xu, "Rail Transit Travel Time Reliability and Estimation of Passenger Route Choice Behavior: Analysis Using Automatic Fare Collection Data," *Transportation Research Record*, vol. 2275, no. 1, pp. 58–67, Jan. 2012, https://doi.org/10.3141/2275-07.

[18] Jiang Zhibin and Liao Shenmeihui, "A Method for Extracting Passenger Flow Time Series Feature of Urban Rail Transit," in *ICTE 2019*, X. Liu, Q. Peng, and K. C. P. Wang, Eds. Reston, VA, USA: American Society of Civil Engineers, 2020, pp. 861–869.

[19] Z. Chen and W. Fan, "Extracting Bus Transit Boarding and Alighting Information Using Smart Card Transaction Data," *Journal of Public Transportation*, vol. 22, no. 1, pp. 40–56, Jan. 2020, https://doi.org/10.5038/2375-0901.22.1.3.

[20] D. J. I. Raj, V. S. Radhakrishnan, M. R. Reddy, N. S. Selvan, B. Elangovan, and M. Ganesan, "The Projection-Based Data Transformation Approach for Privacy Preservation in Data Mining," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15969–15974, Aug. 2024, https://doi.org/10.48084/etasr.7969.

[21] "Basic UMAP Parameters — umap 0.5 documentation." [Online]. Available: https://umap-learn.readthedocs.io/en/latest/parameters.html.

[22] C. J. Nolet *et al.*, "Bringing UMAP Closer to the Speed of Light with GPU Acceleration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 418–426, May 2021, https://doi.org/10.1609/aaai.v35i1.16118.

[23] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey." arXiv, Aug. 25, 2021, https://doi.org/10.48550/arXiv.2109.02508.