JURNAL ECONOMINA
e-ISSN: 2963-1181
Volume 5, Number 10, October 2026
Homepage: ejournal.45mataram.ac.id/index.php/economina

Household Segmentation Based on Housing Conditions and Productive Economic Activities

Farida Nur Hayati[1*], Mega Silfiani[2], Diana Nurlaily[3], Surya Puspita Sari[4]
[1*234]Departement of Data Science And Analytics, Kalimantan Institute of Technology, Indonesia

**Corresponding Author: farida.nur.h93@gmail.com*

Article History

Received: 12-10-2026
Revised: 19-10-2026
Published: 30-10-2026

Keywords: Cluster, Economy, K-Prototype, Regional Profile, Social.

ABSTRACT

Poverty remains a real concern in East Kalimantan. According to Statistics Indonesia (BPS), the province's poverty rate stood at 5.78 percent in March 2024. Behind that figure are roughly 221,340 people living with limited means, struggling to afford basic needs such as adequate food, education, and healthcare. That is why understanding the socio-economic conditions of households matters. By looking closely, this understanding can be used to better assess community welfare and to identify groups whose circumstances differ and who need more targeted policy support.This study aims to describe the socio-economic profiles of households in East Kalimantan as a reflection of community welfare. These profiles are examined through several aspects: the type of area where households live, the materials used to build their homes, access to sanitation and electricity, ownership of assets, and ownership of micro and small businesses. Analyzing these variables is expected to provide a clearer picture of household conditions and to help form groups with distinct characteristics. The analysis employs the K-Prototype clustering method, a technique designed for data containing both numerical and categorical variables. The results show that Cluster 1 requires priority attention for welfare improvement, as this group exhibits relatively lower-quality housing and lower per capita income than Cluster 2.

INTRODUCTION

Disparities in socio-economic characteristics across regions constitute a significant issue in development planning. (Cartone, Panzera, & Postiglione, 2021). Disparities in welfare levels linked to economic factors are also reflected in various other indicators, such as housing quality, sanitation facilities, access to electricity, land and property ownership, and economic activity. East Kalimantan has a poverty rate of 5.78 percent, indicating that poverty remains a socio-economic issue requiring attention. This situation is evidenced by a poor population of approximately 221,340 people in 2024 who face limitations in meeting basic needs such as food, education, and healthcare. Consequently, an analysis is needed to map households and serve as a foundation for effective policy implementation. Poverty is a fundamental issue demanding government attention across various regions. Poverty levels in East Kalimantan's rural and urban areas differed in 2024; in September 2024, the rural poverty rate stood at 8.00%. This figure contrasts sharply with the urban poverty rate of only 4.41%, highlighting an

economic disparity between rural and urban areas. (Badan Pusat Statistika, Indikator Kesejahteraan Rakyat, 2025). In the effort to formulate effective poverty alleviation strategies, the analysis of poverty data is a crucial aspect that can serve as a solid foundation for the government and stakeholders to identify the living conditions of the community. (Badan Pusat Statistika, Laporan Perekonomian Provinsi Kalimantan Timur 2025, 2025).

Cluster analysis is one method that can be used to group households based on socioeconomic aspects, aiming to categorize objects according to the similarity of their characteristics. This method allows households to be sorted into groups that are relatively homogeneous internally, while exhibiting distinct characteristics between groups. Grouping based on various household indicators can provide more specific information regarding household profiles, thereby facilitating the identification of groups with differing socioeconomic characteristics and housing conditions. Meanwhile, cluster analysis a data mining technique has been extensively developed to process mixed data types by accounting for differing data characteristics when measuring distances and determining cluster centers. For numerical data, Euclidean distance can be used to measure the degree of similarity between objects, whereas categorical data can be analyzed using Hamming distance (Rezaei & Daneshpour, 2023). For data containing a combination of numerical and categorical variables, both approaches are integrated into the clustering process for instance, through the k-prototypes algorithm (Kuo & Wang, 2022).

One study employing cluster analysis was conducted by Korir, who examined health and socioeconomic indicators in Kenya. The analysis results demonstrate that cluster analysis methods can be useful for examining and explaining socioeconomic and health disparities (Korir, 2024). Yusran et al. also conducted research on the clustering of households in West Java based on stunting risk factors using the K-Modes and K-Prototypes algorithms. The analysis revealed that, among the two methods tested, the K-Prototypes algorithm demonstrated superior performance, yielding a more balanced cluster distribution and a higher Silhouette Score. (Yusran, et al., 2025). Other research on k-prototype also demonstrates that, whether dealing with numerical, categorical, or mixed datasets, the k-prototype method is capable of achieving optimal clustering performance with lower time complexity (Wang & Mi, 2025). In this study, household characteristics are analyzed based on several aspects, including residential area classification, housing construction materials, sanitation and electricity facilities, asset ownership, and ownership of micro or small enterprises. These variables are expected to provide an overview of household conditions and the formation of groups with distinct characteristics. The resulting groupings can subsequently be used to classify households based on housing conditions, asset ownership, and productive economic activities.

RESEARCH METHODS

Cluster

Cluster analysis or simply "clustering" is the process of partitioning a set of data objects (or observations) into several groups or subsets. Each subset is referred to as a cluster. Clustering operates by grouping data objects such that objects within the same cluster exhibit a high degree of similarity, while objects in different clusters exhibit a high degree of dissimilarity. These levels of similarity and dissimilarity are assessed based on attribute values that describe the characteristics of each object, often involving the use of distance measures. (Han, Kamber, & Pei, 2012). The most popular distance measure is Euclidean distance. Let's assume $i=(x_{i1}, x_{i2},$

…, x_{ip}) and j=(x_{j1}, x_{j2}, …, x_{jp}) For two objects with p attributes, the Euclidean distance can be determined as follows:

$$d(\mathrm{i}, \mathrm{j}) = \sqrt{(\mathrm{x}_{\mathrm{i}1} - x_{j1})^2 + (\mathrm{x}_{\mathrm{i}2} - x_{j2})^2 + ... + (\mathrm{x}_{\mathrm{ip}} - \mathrm{x}_{jp})^2}$$

Another widely known distance metric is the Manhattan distance (or city-block distance). This term originates from the concept of the distance between two points in a city, calculated based on the number of blocks that must be traversed. The Manhattan distance is defined as follows:

$$d(i, j) = | \mathrm{x}_{\mathrm{i}1} - \mathrm{x}_{j1} | + | \mathrm{x}_{\mathrm{i}2} - \mathrm{x}_{j2} | + ... + | \mathrm{x}_{\mathrm{ip}} - \mathrm{x}_{jp} |$$

K-Pototype

K-Prototype analysis is a clustering method designed for mixed-type data specifically, datasets containing both numerical and categorical variables. Developed by Huang, this method extends the K-Means and K-Modes algorithms. It measures dissimilarity by employing squared distance for numerical variables and a matching dissimilarity measure for categorical variables. The weights assigned to categorical variables are user-defined parameters; consequently, inaccuracies in setting these weights can compromise the accuracy of the clustering results (Nisha & Hooda, 2022). The distance measure in k-prototype clustering can be determined by the following equation: (Dinh, Huynh, & Sriboonchitta, 2021).

$$d(\mathrm{x}_i, \mathrm{c}_j) = \sum_{k=1}^{q} (x_{ik} - c_{jk})^2 + \gamma \sum_{h=q+1}^{p} \partial(x_{ih}, \mathrm{c}_{jh})$$

$$\partial(x_{ih}, \mathrm{c}_{jh}) = \begin{cases} 0 & if \;\; x_{ih} = c_{jh} \\ 1 & otherwise \end{cases}$$

where, γ serves as a balancing weight used to prevent one type of attribute from being prioritized, with a factor valued at γ > 0 (Aschenbruck & Szepannek, 2020). x_{ih} is the value of the ith object in the hth variable, dan c_{jh} merupakan nilai pusat pada kelompok ke -j pada variabel ke-h. The stages of k-prototype cluster analysis are as follows: (Aschenbruc, Szepannek, & Wilhelm, 2025):

1. Determine the number of clusters (k) for the initial clustering.
2. Select the object used as the initial centroid
3. Determine the gamma coefficient using the formula $\gamma = \frac{1}{q} \sum_{k=1}^{q} s_k$ where s is the standard deviation of the numerical variable
4. Calculate the distance of each object to the determined centroid.
5. Allocate objects to the nearest cluster.
6. Calculate the new centroid based on the objects within a cluster.
7. The process is carried out iteratively until no objects move between clusters.

Average silhoette width

One analysis method that can be used to evaluate quality and determine the optimal number of clusters in clustering analysis is the Average Silhouette Width. The Average Silhouette Width value ranges from -1 to 1, with values closer to 1 indicating better clustering results (Ikotun, Habyarimana, & Ezugwu, 2025). A value close to 1 indicates that an object is more similar to other objects within the same cluster than to objects in the nearest neighboring cluster. This suggests that the object has been assigned to the appropriate cluster. Conversely, an Average Silhouette Width value close to 0 indicates uncertainty regarding the object's clustering; this implies that the object shares a similar degree of proximity to both its current cluster and its neighboring cluster, making its membership in the specific cluster questionable. Meanwhile, a negative value indicates that the object is potentially closer to a different cluster than to the one it is currently assigned to. Therefore, the Average Silhouette Width can serve as a basis for determining the optimal number of clusters. The equation for calculating the Average Silhouette Width is presented below (Lengyel & Dukát, 2019).

$$S(i) = \frac{1}{n}\sum_{i=1}^{n}\frac{b(x_i) - a(x_i)}{\max(a(x_i); b(x_i))} \tag{3}$$

$$a(x_i) = \frac{1}{n_y - 1}\sum_{jy=1}^{n_y} d(x_{iy}, x_{jy})$$

$$\mathrm{b}(x_i) = \min_{y \neq z}\frac{1}{n_z}\sum_{jz=1}^{n_z} d(x_{iy}, x_{jz})$$

Where,

$d(x_{iy}, x_{jy})$ = the distance between the i-th object in group y and the j-th object in group y

$d(x_{iy}, x_{jz})$ = he distance between the i-th object in group y and the j-th object in group z

Data Source and Variables

This study draws on secondary data from the March 2025 National Socio-Economic Survey (Susenas) conducted by Statistics Indonesia (BPS). The focus is East Kalimantan, covering 6,466 households. Each household is described by fifteen variables: three numeric and twelve categorical. These capture four areas of everyday life: household demographics, housing conditions, access to basic infrastructure, and access to information technology. To make sense of this data, we applied K-Prototype clustering, a method suited to mixed numeric and categorical data. It groups households that share similar characteristics, so that distinct household profiles can emerge. The variables used are detailed below.

Table 1. Variables Used in the Study

Type		Variable	Data Scale	Explanation
Household Demographics	x1	Per Capita Expenditure	Ratio	Rupiah
	x2	Number of Household Members (HHM)	Ratio	Persons

	x3	Ownership of Micro/Small Business	Nominal	1. Yes 5. No
Housing Conditions	x4	Floor Area (m^2)	Ratio	m^2 (rounded to the nearest square meter)
	x5	Urban/Rural Classification	Nominal	1. Urban 2. Rural
	x6	Housing Ownership Status	Nominal	1. Owner-occupied 2. Rented/Leased 3. Rent-free 4. Official/Service housing 5. Other
	x7	Roof Type	Nominal	1. Concrete 2. Roof tiles 3. Corrugated metal 4. Wood/Shingles 5. Asbestos 6. Bamboo/Straw/Palm fiber/Leaves/Thatch 7. Other
	x8	Wall Type	Nominal	1. Brick/Concrete wall 2. Plastered woven bamboo/Wire mesh 3. Wood/Plank/Timber 4. Bamboo/Woven bamboo 5. Other
	x9	Floor Type	Nominal	1. Marble/Granite 2. Ceramic tiles/Tiles/Terrazzo 3. Parquet/Vinyl/Carpet 4. Wood/Plank 5. Cement/Red brick 6. Bamboo/Earth 7. Other
	x10	Land Ownership	Nominal	1. Yes 5. No
Basic Infrastructure Access	x11	Type of Sanitation/Toilet Facility Used	Nominal	1. Water-sealed toilet (leher angsa) 2. Covered pit latrine 3. Uncovered pit latrine 4. Pit latrine/Cesspit
	x12	Main Source of Household Lighting	Nominal	1. PLN electricity with a meter 2. PLN electricity without a meter

				3. Non-PLN electricity 4. No electricity
	x13	Access to Safe Water / Main Source of Drinking Water	Nominal	1. Branded bottled water 2. Refilled drinking water 3. Piped water 4. Borehole/Pump well 5. Protected well 6. Unprotected well 7. Protected spring 8. Unprotected spring 9. Surface water (river/lake/reservoir/pond/irrigation) 10. Rainwater 11. Other
Information and Communication Technology	x14	Access to Mobile Phone	Nominal	1. Yes 5. No
	x15	Internet Access	Nominal	1. Yes 5. No

Step Of Analysis

The steps for the household mapping analysis in East Kalimantan Province are as follows:

1. Conducting a literature review.
2. Collecting data.
3. The collected data must undergo a preprocessing stage to ensure cleanliness and readiness for analysis. This process involves several steps:
   a. Data cleaning: Removing duplicate data, irrelevant text, and unnecessary special characters.
   b. Imputing missing data: If missing data is encountered, imputation is performed using appropriate methods, such as filling in with the mean value or using statistical model-based methods.
4. Describing the socio-economic conditions in East Kalimantan.
5. Mapping community welfare in East Kalimantan using households as the unit of observation.
6. Interpreting results and drawing conclusions.
   The final stage of this research involves interpreting the analysis results and drawing conclusions based on the findings

ANALYSIS AND DISCUSSION

Prior to performing cluster analysis using the K-Prototype method, the collected SUSENAS data was first analyzed using descriptive statistics through data visualization. This stage aimed

to provide an overview of household characteristics in East Kalimantan Province based on the variables used in the study. One of the characteristics examined was the distribution of the number of households across regencies and cities. The distribution of the households included in the study across each regency and city in East Kalimantan Province is presented in the following figure.

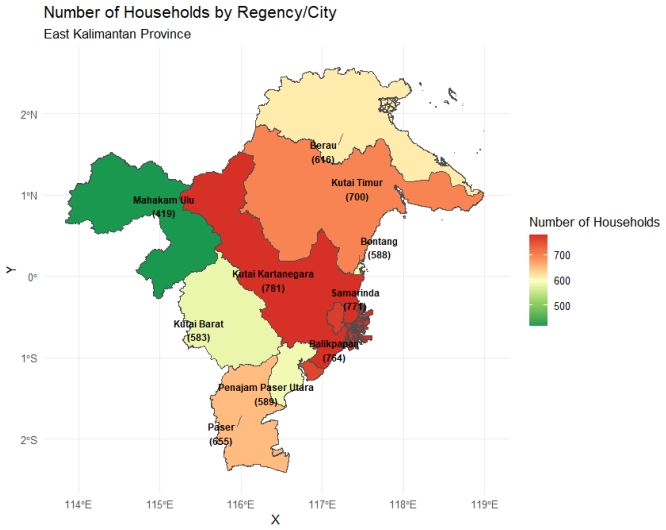


Figure 1. Number of Household Samples per Regency/City

The figure illustrates the number of households in each regency and city within East Kalimantan. Analysis reveals that, among the sampled areas, Kutai Kartanegara Regency has the highest number of households (781), while Mahakam Ulu Regency has the lowest (419), with a total of 6,466 households across East Kalimantan. These households are categorized into urban and rural areas based on the following characteristics:.

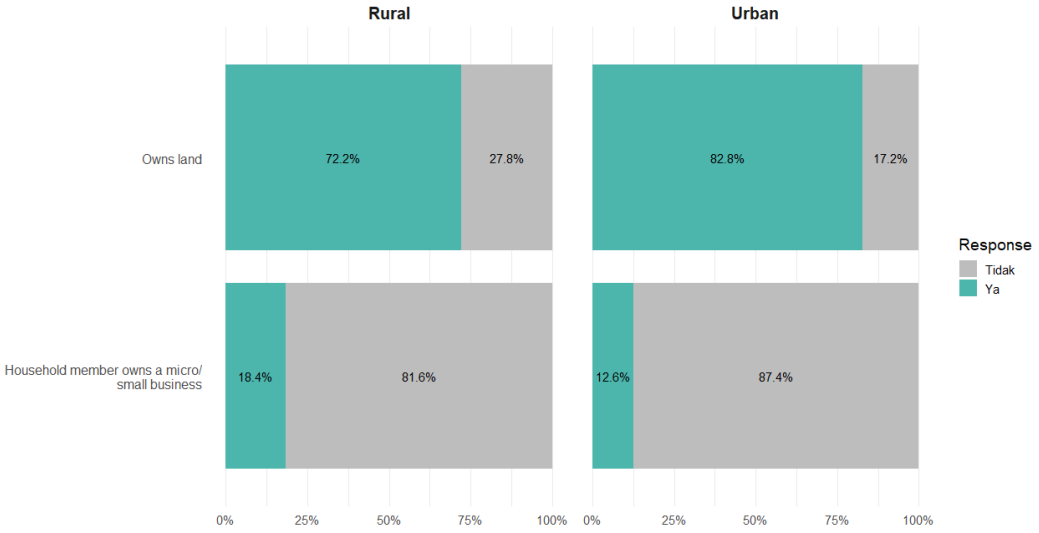


Figure 2. Characteristics of Land Ownership and Micro/Small Enterprise Data

The visualization shows fairly clear differences in household characteristics between urban and rural areas. In urban areas, 82.8% of households own land, while 12.6% of them run a micro

or small enterprise. The picture is different in rural areas. Although only 72.2% of households own land, the share of households running a micro or small enterprise is actually higher, at 18.4%.

This shows that land ownership is indeed more common in urban areas, but small-scale business activity is more widespread among rural households. In other words, owning land does not necessarily go hand in hand with a greater share of households running a business. In rural households, micro and small enterprises are one of the more commonly chosen ways to support daily life.

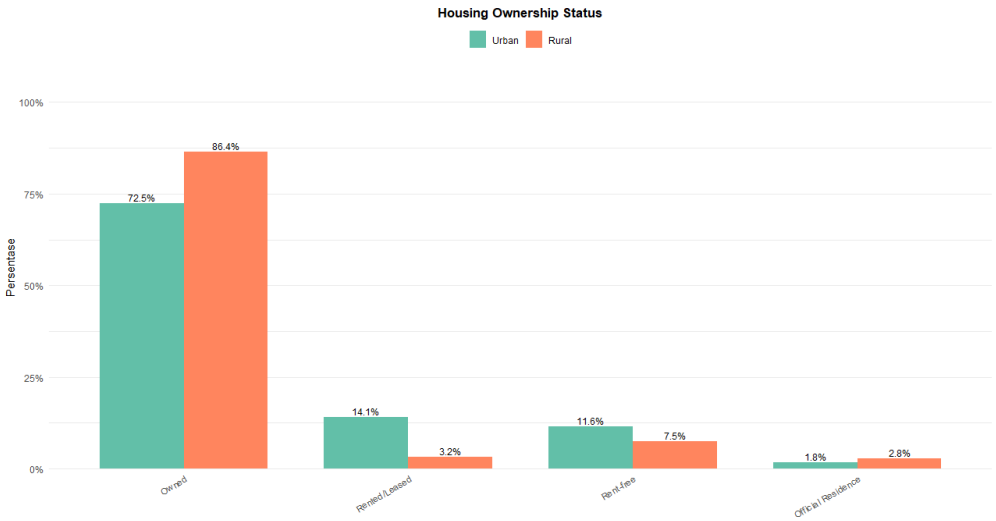


Figure 3. Characteristics of Building Ownership Data

Based on the visualization results, building ownership status differs between urban and rural areas. Owner-occupied buildings account for a higher percentage in rural areas (86.4%) compared to urban areas. Conversely, the rented/leased and rent-free categories are more prevalent in urban areas. Meanwhile, employer-provided housing represents a higher percentage in rural areas (2.8%). These differences in building ownership characteristics reflect variations in housing conditions between urban and rural households. Furthermore, household characteristics based on the primary flooring material are presented in the following figure.

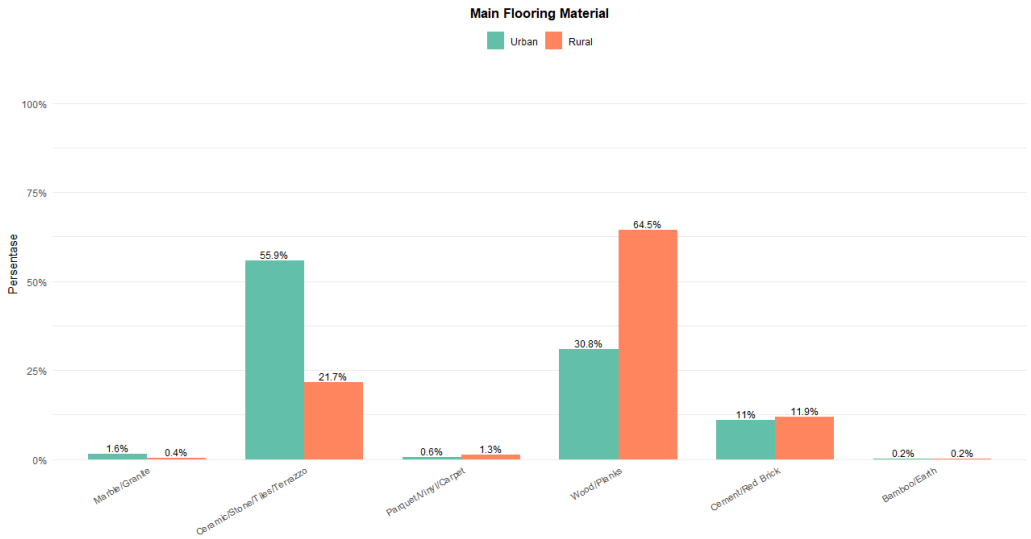


Figure 4. Characteristics of Flooring Material Types

The figure shows that urban and rural households tend to use different flooring materials. In urban areas, ceramic, tile, or terrazzo is the most common choice, used by 55.9% of households. In contrast, in rural areas, wood or plank flooring is more widely used, at 64.5%.

This difference gives a glimpse of how housing conditions vary between the two settings. The use of ceramic, tile, or terrazzo in urban areas suggests that most households there live in buildings made with more permanent materials. Meanwhile, the dominance of wood or plank flooring in rural areas reflects housing characteristics that remain closely tied to materials available in their surroundings. Next, the following figure shows household characteristics based on the main source of lighting in each area.

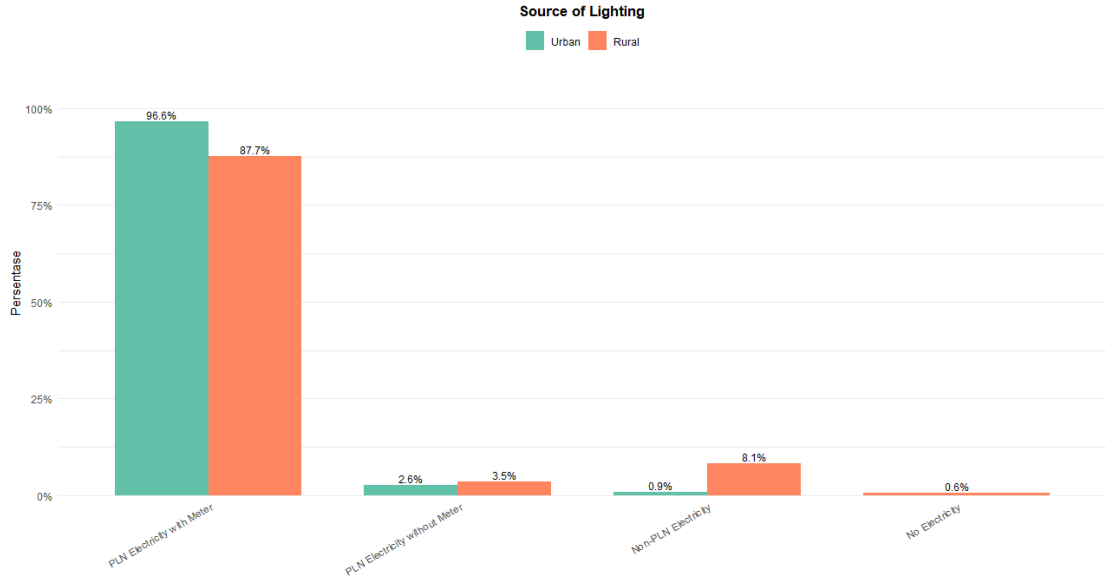


Figure 5. Characteristics of Light Sources

Visualization results indicate that metered PLN electricity is the dominant source of household lighting in both urban and rural areas. While 96.6% of urban households use metered PLN electricity compared to 87.7% in rural areas alternative lighting sources remain more prevalent in rural settings. This is evidenced by the relatively higher percentages of rural households using unmetered PLN electricity, non-PLN electricity, or non-electric lighting sources compared to their urban counterparts. These findings suggest that access to electric lighting remains relatively lower in rural areas.

Cluster Analysis of Household Data

After visualizing the data, the next step is to standardize the numerical data before performing cluster analysis. Cluster analysis is conducted to map households that share similar characteristics. The K-Prototype method is used for this mapping, and the optimal value of k is determined using the silhouette method. The results regarding the optimal k for the K-Prototype method are presented below..

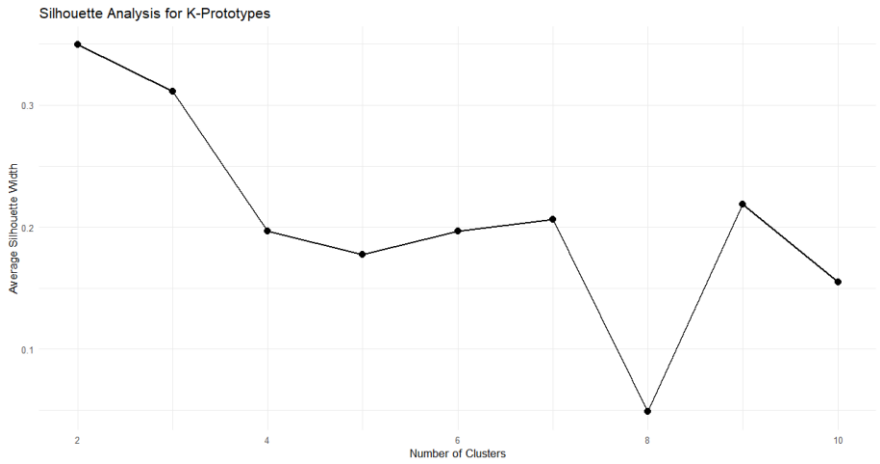


Figure 6. Selection of the Best K

Based on the silhouette value, it was determined that the optimal number of clusters for grouping households in East Kalimantan is two, with a value of $\gamma = 3.462631$; the distribution of households across these clusters is presented in a map and a table, showing 3,374 members in Cluster 1 and 3,092 in Cluster 2. The number of clusters in each regency/city is as follows:

Table 2. Number of Members in Clusters 1 and 2

Regency/City	Cluster 1	Cluster 2	Total
Balikpapan	139 (18,2 %)	625 (81,8 %)	764 (100%)
Berau	386 (62,4 %)	230 (37,3 %)	616 (100%)
Bontang	130 (22,1 %)	458 (77,9 %)	588 (100%)
West Kutai	418 (71,7 %)	165 (28,3 %)	583 (100%)
Kutai Kartanegara	516 (66,1 %)	265 (33,9%)	781 (100%)
East Kutai	493 (70,4 %)	207 (29,6 %)	700 (100%)
Mahakam Ulu	411 (98,1 %)	8 (1,9 %)	419 (100%)

Paser	399 (60,9 %)	256 (39,1 %)	655 (100%)
North Penajam Paser	264 (44,8 %)	325 (55,2 %)	589 (100%)
Samarinda	218 (28,3 %)	553 (71,7%)	771 (100%)
Total	**3.374**	**3.092**	**6.466**

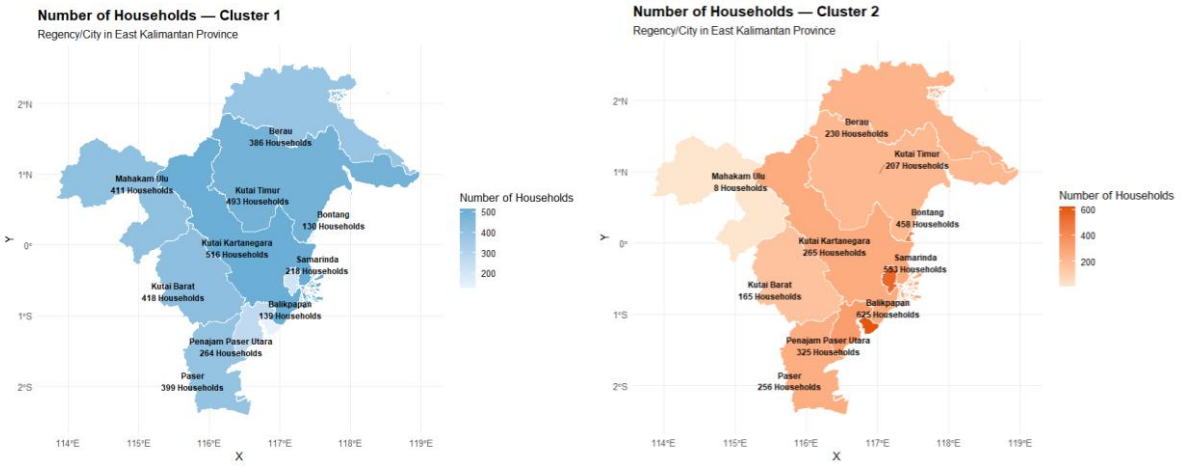


Figure 7. Number of Regencies/Cities in Each Cluster

The analysis results indicate that the four regions with the highest number of households in Cluster 1 are Mahakam Ulu, West Kutai, East Kutai, and Kutai Kartanegara. Meanwhile, the four regions with the highest figures in Cluster 2 are Balikpapan, Bontang, Samarinda, and North Penajam Paser. Specifically, the highest number of households in Cluster 1 is found in East Kutai (493), while in Cluster 2, it is in Balikpapan (625). To identify characteristic differences between Cluster 1 and Cluster 2, a descriptive statistical analysis will be conducted on numerical variables, specifically average per capita expenditure and average floor area of the residential building.

Table 3. Characteristics of Clusters 1 and 2 Based on Numerical Variables

Cluster	Average per capita expenditure (Rp)	Average floor area of residential buildings (m^2)
1	2.023.746	72,1
2	2.938.215	94,3

Based on the table, Cluster 1 is characterized by a lower per capita expenditure compared to Cluster 2 (which stands at 2,023.746). The average floor area of residential buildings in Cluster 1 is also lower than that of Cluster 2 (which is 72.1 m^2). These analysis results indicate that households in Cluster 1 belong to an area with a lower economic level compared to Cluster 2.

The characteristics used to compare Cluster 1 and Cluster 2 based on categorical variables are as follows:

Table 4. Characteristics of Clusters 1 and 2 Based on Categorical Variables

Variable	Cluster 1	Cluster 2
Micro/Small Business Ownership	The majority of households do not own a micro/small business (85.9%), while only 14.1% own a micro/small business.	The majority of households do not own a micro/small business (81.6%), while 18.4% own a micro/small business.
Area Classification	The majority are located in rural areas (59.3%), while 40.7% are located in urban areas.	The majority are located in urban areas (84.1%), while only 15.9% are located in rural areas.
Building Ownership Status	Most households occupy owner-occupied buildings (78.0%). A smaller proportion live in rent-free buildings (11.1%) or rented/leased buildings (9.8%).	Most households occupy owner-occupied buildings (77.7%), followed by rented/leased buildings (approximately 10.0%) and rent-free buildings (8.8%).
Main Roof Material	The majority of households use corrugated metal roofs (91.5%).	The majority of households use corrugated metal roofs (86.6%), followed by roof tiles (approximately 7.5%).
Main Wall Material	The majority of households have walls made of wood, boards, or timber (95.2%).	The majority of households have brick walls (92.3%), while approximately 7.4% have walls made of wood, boards, or timber.
Main Floor Material	The dominant floor material is wood/boards (83.1%), followed by cement/red brick (approximately 12.3%).	The dominant floor material is ceramic tiles/stone/terrazzo (86.1%), followed by cement/red brick (approximately 10.3%).
Land Ownership	Most households own land (74.0%), while 26.0% do not own land.	The proportion of households owning land is higher, at 78.8%, while 21.2% do not own land.
Toilet Type	The majority of households use water-sealed toilets (91.3%).	Toilet facilities are overwhelmingly dominated by water-sealed toilets (98.3%).

Electricity Source	The majority of households use PLN electricity with a meter (89.6%).	Almost all households use PLN electricity with a meter (97.0%).
Drinking Water Source	The majority of households use refilled drinking water (79.6%), followed by piped water (approximately 6.5%).	Refilled drinking water is overwhelmingly dominant (80.4%), followed by piped water (approximately 5.6%).

Based on the analysis, Cluster 1 is characterized as a group of households predominantly located in rural areas, featuring relatively simple housing characteristics. This is clearly seen in the materials used to build their homes. Wood, planks, or logs make up the walls in 95.2% of households, and wood or planks cover the floors in 83.1%. Most homes in this cluster also have zinc roofing (91.5%), gooseneck (water-seal) toilets (91.3%), and metered PLN electricity (89.6%). In terms of assets, most households own their homes (78.0%) and the land they live on (74.0%). Business activity, however, is limited: 85.9% of households do not run a micro or small enterprise.

Cluster 2, by contrast, is made up mostly of urban households living in relatively permanent, modern homes. Brick walls are found in 92.3% of these homes, and 86.1% have ceramic, stone, or terrazzo flooring. Basic utilities are also widely available, with 97.0% of households using metered PLN electricity and 98.3% having gooseneck toilets. Land ownership stands at 78.8%, a slightly higher share than in Cluster 1. Business activity is similarly low: 81.6% of households do not operate a micro or small enterprise, and only 18.4% do. The analysis results show that several variables provide quite different information between Cluster 1 and Cluster 2, such as the urban/rural classification, main wall material, and main floor material.

CONCLUSION

The K-Prototype clustering analysis identified an optimal solution of two clusters, with a value of $\gamma = 3.462631$. Cluster 1 contains 3,374 households and Cluster 2 contains 3,092. Cluster 1 represents rural households living in simpler dwellings, typically with wooden walls and floors. Their housing conditions are relatively modest, and their per capita income is lower. Most of these homes have zinc roofing (91.5%), gooseneck (water-seal) toilets (91.3%), and metered PLN electricity (89.6%). In terms of assets, most households own their homes (78.0%) and the land they live on (74.0%). Business activity is limited, however: 85.9% of households do not run a micro or small enterprise. In contrast, Cluster 2 represents urban households living in more permanent structures, with brick walls and ceramic flooring. Their housing is more modern, and their per capita income is higher than that of Cluster 1.

Based on this analysis, Cluster 1 should be a priority for welfare improvement, given its rural household profile, simpler housing quality, and lower per capita income compared to Cluster 2.

ACKNOWLEDGMENTS

Gratitude is extended to LPPM ITK for facilitating this research and enabling its successful completion, thanks are also due to those who assisted with the study on household segmentation based on housing conditions and productive economic activities.

REFERENCES

Aschenbruc, R., Szepannek, G., & Wilhelm, A. (2025). Initialization strategies for clustering mixed-type data. Advances in Data Analysis and Classification.

Aschenbruck, R., & Szepannek, G. (2020). Cluster Validation for Mixed-Type Data. KIT Scientific, 6(1), 1-12.

Badan Pusat Statistika. (2025). Indikator Kesejahteraan Rakyat. Samarinda: BPS Provinsi Kalimantan Timur.

Badan Pusat Statistika. (2025). Laporan Perekonomian Provinsi Kalimantan Timur 2025. Samarinda: BPS Provinsi Kalimantan Timur.

Cartone, A., Panzera, D., & Postiglione, P. (2021). Regional economic disparities, spatial dependence and proximity structures. Regional Science Association International., 14, 1034–1050.

Dinh, D.-T., Huynh, V.-N., & Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. Information Sciences(571), 418–442.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining. USA: Elsevier Inc.

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. Heliyon, 1-36.

Korir, E. K. (2024). Comparative clustering and visualization of socioeconomic and health indicators: A case of Kenya. Socio-EconomicPlanning Sciences(95), 1-13.

Kuo, T., & Wang, K.-J. (2022). A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification. Computers & Industrial Engineering , 169, 1-12.

Lengyel, A., & Dukát, Z. B. (2019). Silhouette width using generalized mean A flexible method for assessing clustering efficiency. Ecology and Evolution, 9(23), 13231–13243.

Nisha, & Hooda, B. (2022). A Survey on Clustering Methods for Numberic, Categorical, and Mixed Variables Data. Int. J. Agricult. Stat , 18(2), 675-679.

Rezaei, H., & Daneshpour, N. (2023). Mixed data clustering based on a number of similar features . Pattern Recognition , 143.

Wang, H., & Mi, J. (2025). Intuitive-K-prototypes: A mixed data clustering algorithm with intuitionistic distribution centroid. Pattern Recognition, 158, 1-14.

Yusran, M., Nuradilla, S., Putri, M. R., Fitrianto, A., Erfiani, & Yudhianto, R. B. (2025). Household Clustering in West Java Based on Stunting Risk Factors Using K-Modes and K-Prototypes Algorithms. Journal of Applied Informatics and Computing (JAIC), 9(6), 2986~2999.