e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v8i1.14729

Variable Selection in Clustering for Sanitation Access Analysis in East Java Supporting SDG 6

Mohammad Dian Purnama

Department of Mathematics, Faculty of Mathematics and Natural Sciences,
State University of Surabaya (UNESA),
Surabaya, Indonesia 60231
mohammaddian.20053@mhs.unesa.ac.id

Correspondence: mohammaddian.20053@mhs.unesa.ac.id

Abstract – *To have sanitation we need to think about a few things to make people healthy and help the world be a better place. This study is trying to figure out how people in East Java get to use sanitation. We are looking at an important things that help us understand how people use sanitation. We used a method called clustering to see how different cities and districts in East Java are doing. This study utilized a set of six variables, encompassing the five pillars of community-based total sanitation (STBM). The variables employed following the selection process include awareness of open defecation (SBS), awareness of hand washing with soap (CPTM), and drinking water and food management (PAMMRT). The resulting in this study has three distinct clusters, each reflecting different levels of sanitation across cities and districts in East Java. However, the clustering is important to recognize that the excluded variables maintain considerable value as indicators established by the government. Furthermore, to its capacity to implement the variable selection method in the context of clustering, it is anticipated that this research will serve as a valuable resource for policymakers, providing them with a framework to prioritize specific areas in their efforts to enhance sanitation access for the purpose of achieving sustainable development.*

Keywords: *Clustering; Variable Selection; Sanitation Access; SDG 6*

I. INTRODUCTION

Access to sanitation is very important for achieving the Sustainable Development Goal 6 which is about providing water and sanitation services to everyone in the world. The good sanitation has an impact on many areas of society including health, education, money and the environment (Pereira & Marques, 2021; WHO & UNICEF, 2023). Indonesia has made progress in providing sanitation there are still problems, especially in places like East Java Province.

According to the Central Statistics Agency, 85.56 percent of people in East Java have access to sanitation (BPS, 2024). If we compare with other provinces in Indonesia, such as Bali has the best sanitation access at 96.83 percent and Bangka Belitung Island at 94.16 percent. The data show that East Java has good sanitation access but many other provinces have even better access.

This study wants to look the pattern of access to sanitation in East Java and what factors affect it. This paper use methods to group areas in East Java based on their sanitation variables. Grouping areas like this helps to watch patterns and relationships that we might not see if we just look at one thing at a time (Govender & Sivakumar, 2020). By grouping areas based on variables we can find groups that are facing similar problems with sanitation. However to get results from this kind of analysis we need to choose the right variables. If we choose variables that're not important or that are redundant it can hide the real patterns in the data. (Tosunoglu & Kocak, 2023). These variables are important because they were set by the government and can give us insights. So not choosing a variable does not mean it is not important it just means we can look at it in studies or when making policies.

This study should give us detailed information about access to sanitation in East Java, which can help make better policies at the national and local levels. It can also give information to non-governmental organizations and international agencies so they can design programs that are tailored to the specific needs of the area. While other studies have used grouping methods to look at sanitation there is not much research that combines variable selection with other methods to improve the indicators of Sustainable Development Goal 6 in Indonesia (United Nations, 2024). So the findings of this research can add to what we know about access,

to sanitation and how to use grouping methods to achieve sustainable development.

II. METHODS

The present study utilizes the clustvarsel method to ascertain the most salient variables that influence sanitation access. Consequently, the scope will be narrowed to a select few critical variables for analysis. Clustering analysis is defined as any process that produces distinctive classes containing objects that are similar in several ways to members of their class but different from objects in other classes. This approach has the potential to reveal hidden differences in how access to sanitation varies across different regions, which might not have been immediately obvious (Buitrago-Boret et al., 2019).

The research process will be methodically executed in the following sequence: initial data collection, standardization of the collected data, and culminating in the application of the clustvarsel method for variable selection. Then, we get to work on the clustering, so we can find out how many clusters there are and where the cities and districts are in each one. As illustrated in Figure 1, the research methodology involved a series of steps.

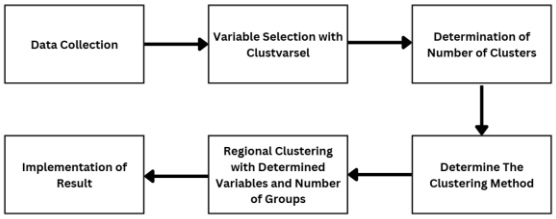

Figure 1. Steps of Analysis

2.1 Data Collection

This study employs secondary data procured from the Central Statistics Agency (BPS) of East Java. The data set encompasses the five pillars of community-based total sanitation (STBM), along with the number of villages implementing the STBM. The five pillars of STBM are as follows: the first pillar is the Stop Open Defecation (SBS) initiative, the second is the Clean Hands with Soap (CPTS) initiative, the third is the Household Drinking Water and Food Management (PAMMRT) initiative, the fourth is the Household Waste Management (PSRT) initiative, and the fifth is the Household Liquid Waste Management (PLCRT) initiative. It should be noted that all data presented herein are sourced from the year 2023. These six indicators function as variables in the clustering process for 38 districts/cities in East Java Province.

2.2 Variable Selection Algorithm

This research employs a model-based clustering method that posits the observed data originates from a combination of G components, with each component symbolizing a probability distribution for distinct groups or clusters (Redivo, 2020). For continuous data, the density of every mixture component is represented by a multivariate Gaussian distribution. Therefore, the overall structure of the Gaussian mixture model can be represented in equation 1.

$$f(x) = \sum_{g=1}^{G} \pi_g \, \varphi \, (x|\mu_g, \Sigma_g) \quad (1)$$

where π_g represents the mixing probabilities with the condition $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$, while $\varphi(\cdot)$ is a multivariate Gaussian density with parameters μ_g, Σ_g for g = 1, ..., G. A simplified parameterization of the covariance matrices is achieved via eigenvalue decomposition in equation 2.

$$\Sigma_g = \lambda_g D_g A_g D_g^T \quad (2)$$

In equation 2, λ_g means a scalar regulating the size of the ellipsoid, A_g is diagonal matrix defining the form of the density contours, and D_g^T means an orthogonal matrix establishing (Fraley et al., 1998).

Sometime, selecting variables for clustering models is addressed by reinterpreting the problem as a procedure for model selection. Their suggestion relies on employing the Bayesian Information Criterion (BIC) to estimate Bayes factors for comparing mixture models applied to nested subsets of variables (Lu & Lou, 2023). Assume that the collection of accessible variables, X, is divided into three non-overlapping sections: the group of already chosen variables (X_{Clust}), the variable being evaluated for either inclusion or exclusion from the active set (X_i) , and the set of leftover variables ($X_{Other} = X \setminus \{X_{Clust} \cup X_I\}$). The addition (or removal) of variables can be evaluated using the subsequent BIC difference in equation 3.

$$BIC_{diff} = BIC_{clust}(X_{clust}, X_i) - BIC_{not}clust(X_{clust}, X_i) \quad (3)$$

where $BIC_{clust}(X_{clust}, X_i)$is the BIC value for the "best" clustering mixture model (i.e., assuming G ≥ 2) fitted using the feature set $\{X_{Clust} \cup X_I\}$, where as $BIC_{not}clust(X_{clust}, X_i)$ is the BIC value for no clustering for the same set of variables.

The BIC value for the no clustering condition can be written in equation 4.

$$BIC_{not}clust(X_{clust}, X_i) = BIC_{clust}(X_{clust}, X_i) + BIC_{reg}(X_i|X_{Clust}) \quad (4)$$

The BIC value for the optimal clustering model obtained from the set X_{Clust}. BIC value for regressing the candidate variable X_i on the variables present in the set X_{Clust} (Maugis et al., 2009). The BIC in Equation (3) serves as an approximation of the logarithm of the Bayes factor that contrasts the model in which the variable of interest, X_i, acts as a clustering variable with the model where this variable is conditionally independent of the clustering. Significant positive values of BICdiff can be interpreted as support for the idea that variable X_i valuable for clustering (Maugis et al., 2009).

In the linear regression model, X_i may rely on all variables in X_{Clust}, only some of them, or none at all (total independence). Consequently, in accordance with Maugis et al. (2009), the regression involving all previously chosen clustering variables is substituted with regression on a selected subset of these variables, determined using a stepwise approach. This enables a more simplified modeling of the connection between the noise variables and the clustering variables. Lastly, observe that in equations (3) and (4) the group of other variables, X_{other}, has no impact. In every clustering model, the optimal model is determined based on the number of mixture components (assuming G $\geq$ 2) and the parameterization of the model. This research employs different covariance parameterizations offered in the mclust package, with a specific emphasis on 4 models.

Initially, model E (Equal, spherical) represents a model that has the same spherical covariance across all clusters, characterized by Σ_g = I, with λ consistent across all groups. Secondly, model V (Variable, spherical) features spherical covariance but varying volumes for each cluster, expressed as Σ_g = λI. Next, Ellipsoidal, equal volume and shape (EEV) model. The model has configuration with equivalent volume and shape yet varying orientations, but with varying orientations. The covariance structure here is really interesting, because it maintains identical eigenvalues for volume and shape but exhibits distinct eigenvector matrices for orientation. Finally, the VVV model is the most general and adaptable model. This is because it allows each group to have a different volume, shape and direction. (Gogebakan, 2021).

To choose the best model, we look at the BIC for different numbers of clusters (G) and different ways to combine them. The best model for each group of variables is chosen based on the highest BIC value. The BICclust value shows the highest BIC value that is achieved through this process of selecting a model.

2.3 Hierarchical Clustering

Clustering analysis is one method of multivariate method aimed at grouping objects into distinct categories. The objective of cluster analysis is to categorize items with analogous characteristics into a cluster. Studies have shown that things that are closer together are more similar to each other than things that are farther apart (Purnama, 2025).

Hierarchical clustering can be put into two groups: agglomerative nesting (AGNES) and divisive analysis clustering (DIANA) (Purnama & Sofro, 2025). The process of agglomerative clustering starts by putting each object into the right cluster. After this first phase, the smaller groups slowly become one bigger group. Conversely, divisive clustering initiates by assembling all objects into a single cluster. Subsequently, a sequential separation algorithm is employed to disaggregate them into distinct clusters.

A range of agglomerative hierarchical clustering algorithms have been employed to identify clusters, including Single Linkage, Complete Linkage, Average Linkage, and Ward's techniques (Purnama, 2023). The Single Linkage approach is a statistical method used to determine the distance between two clusters. This distance is defined as the minimum separation between two points within the cluster. The equation employed to ascertain this distance is presented in equation 5.

$$D(A,B) = min\{d(y_i, y_j)\}, A \in y_i, B \in y_j \quad (5)$$

In equation 5, $d(y_i, y_j)$ is employed to denote the Euclidean distance, which is the distance between vectors y_i and y_j. At each stage of the Single Linkage method, the distance D(A,B) is calculated for each pair of clusters, and the two clusters with the smallest distance are merged (Purnama, 2023). In contrast to the Single Linkage approach, the Complete Linkage method defines the distance between two clusters as the maximum distance between points in different clusters. The calculation of this distance is performed using the following formula in equation 6.

$$D(A,B) = max\{d(y_i, y_j)\}, A \in y_i, B \in y_j \quad (6)$$

The Complete Linkage method, the distance $D(A,B)$ between two clusters lets say cluster A and cluster B is figured out for every pair of clusters at each step. The two clusters that are the farthest apart are then combined into one cluster. This method measures the distance between clusters by finding the distance between any two points in cluster A and cluster B (Purnama, 2023). Next, Average Linkage methods sorts data based on the average distance between all the points. The distance between cluster A and B is determined by calculating the mean of the distances between each point, n_A in A and each point, n_B in B.

The distance is calculated using the following formula in equation 7.

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j), A \in y_i, B \in y_j \quad (7)$$

The distance $D(A,B)$ in Average Linkage Method is employed to identify and amalgamate the two clusters with the minimal average distance (Purnama, 2023). Next, Ward's method reducing information loss when two object (node) has merged. It does this by looking at the sum of squares error (SSE). With the lowest SSE indicating a higher level of similarity within the cluster (Purnama & Sofro, 2025). The SSE formula is as follows in equation 8.

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})' - (y_i - \bar{y}) \quad (8)$$

In equation 8, y_i is the value of the i-th object, n is the number of variables, and $\bar{y}$ is the mean value of the objects in the cluster.

Agglomerative Hierarchical Clustering. The method requires a thorough examination of the correlation coefficient values. In the context of cluster trees, this examination is of particular importance. The objective of this evaluation is to ascertain the linear relationship between the cophenetic distances derived from the clustering tree and the original distances (dissimilarities) employed to generate the tree (Saraçlı & Akşit, 2022). The subsequent equation is employed to calculate cophenetic correlation in equation 9.

$$r_{coph} = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{Cik} - \bar{d}_C)}{\sqrt{[\sum_{i<k}(d_{ik} - \bar{d})^2]\,[\sum_{i<k}(d_{Cik} - \bar{d}_C)^2]}} \quad (9)$$

In equation 9, d_{ik} is used to denote the Euclidean distance between two entities, while d_{Cik} signifies the cophenetic distance between the same two entities. The term d_{ik} is used interchangeably with d_{ik} and d_{Cik} is used interchangeably with d_{Cik}. The terms d_{ik} and d_{Cik} represent the average Euclidean distance and cophenetic distance, respectively. A cophenetic correlation value approaching 1 indicates that the clustering solution has effectively captured the underlying structure of the original data configuration.

The decision to combine Gaussian mixture-based variable selection (clustvarsel) with Agglomerative Nesting (AGNES) is theoretically driven by the need to ensure model-based rigor in feature selection while maintaining the interpretability of hierarchical relationships. Clustvarsel identifies the optimal subset of variables that maximize the Bayesian Information Criterion (BIC), which then serves as the robust foundation for the hierarchical grouping process to visualize regional disparities.

III. RESULTS AND DISCUSSION

The selection of variable before clustering was conducted using the clustvarsel method. These six variables included the five parts of community-based total sanitation (STBM), and the number of villages that were using STBM. By using the clustvarsel analysis, a select array of variables was identified as the most salient contributors to cluster differentiation. The following table has been constructed in order to enhance clarity and consistency. Tables 1 and 2 present the selected variables that form the basis for the subsequent clustering process.

Table 1. Selected Variables for Clustering Model

Variable	Type of Step	BICclust	Model	Group
SBS	Add	10.905	E	8
CPTS	Add	-432.372	EEV	6
PAMMRT.	Add	-547.917	EEV	3
PAMMRT	Remove	-432.372	EEV	6
PSRT	Add	-1109.986	EEV	
PAMMRT.	Remove	-432.372	EEV	6

Table 2. Evalaluation of Selected Variables

Variable	BICclust	BICdiff	Group
SBS	10.905	332.084	Accepted
CPTS	-432.372	-97.591	Accepted
PAMMRT.	-547.917	170.595	Accepted
PAMMRT	-432.372	170.595	Rejected
PSRT	-1109.986	-241.83238	Rejected
PAMMRT.	-432.372	170.595	Rejected

After obtaining the results of the variable selection process using a Gaussian-based clustering approach, the proposed variables were identified based on changes in the Bayesian Information Criterion (BIC) values. The proposed variables include awareness of open defecation (SBS), awareness of handwashing with soap (CPTM), and management of drinking water and food (PAMMRT). These variables were incorporated into the clustering process based on BIC efficiency. The SBS variable showed a BIC difference of 332.08372 and was included in the model with configuration E and 8 clusters. Concurrently, the CPTS and PAMMRT variables were also acceptable despite experiencing a decrease in BIC of -97.59108 and 170.59515, respectively; the recommended model is the EEV model with 6 and 3 clusters.

However, removing the PAMMRT variable was not accepted in two separate scenarios, despite a BIC difference of 170.59515. Furthermore, adding the PSRT variable was also not accepted due to its significant impact on the BIC, which showed a

decrease of -241.83238 in the EEV model with three clusters. The selection process resulted in the identification of the most optimal subset of variables, which includes SBS, CPTS, and PAMMRT.

Following the variable selection process, which utilized a stepwise greedy search method for Gaussian-based clustering, three recommended variables were identified: SBS, CPTS, and PAMMRT. After considering these variables, the next step was to perform clustering using the hierarchical agglomerative clustering method. To determine the most effective agglomeration method, a comparative analysis was conducted on four approaches: single linkage, complete linkage, average linkage, and the Ward method. The evaluation was performed by calculating the cophenetic correlation coefficient (cor comp.), which measures the extent to which the dendrogram structure represents the original distances between data points. The comparative analysis of the methods used is presented in Table 3.

Table 3. Cophenetic Correlation Coefficient

Method	Cor Comp
Single	0.8384
Average	0.8797
Complete.	0.7599
Ward.	0.825

From the Table 3, the average linkage method has the highest cophenetic correlation coefficient value, reaching 0.8797. The average linkage approach was employed to form three clusters based on the previously selected variables, namely SBS, CPTS, and PAMMRT.

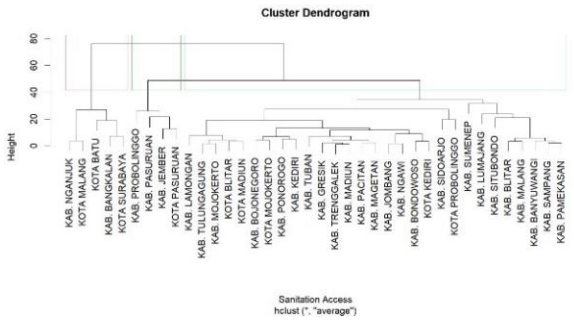

Figure 2. Dendogram of Clustering

Based on Figure 2, the regions in East Java have been divided into three clusters based on their similarities in terms of access to sanitation. In the context of a dendrogram, the proximity of two objects indicates the degree of similarity between them. Specifically, objects with similar characteristics are placed close together because they share a higher degree of similarity. Conversely, objects that are far apart indicate more significant differences between them. The dendrogram also serves as a visual spatial representation correlated with the degree of similarity or difference in their characteristics. Points that are spatially closer indicate more striking similarities, while points that are farther apart indicate more significant differences. From Figure 2, the first cluster includes regions with superior sanitation access and is dominated by major cities, such as Surabaya, Malang, and Batu. The second cluster includes areas with moderate sanitation access, some of which have relatively low levels of urbanization, such as Bojonegoro, Mojokerto, and Kediri. The third cluster includes areas with relatively limited sanitation access, the majority of which are in coastal areas or regions facing geographical constraints in sanitation management, such as Sumenep, Sampang, and Pamekasan.

Table 4. The Average Variable of Each Cluster

Cluster	SBS	CPTS	PAMMRT
1	100	29.74	27.682
2	54	75.	81.54
3	97.01	81.142	81.267

Based on Table 4, the results of the descriptive statistical analysis for all formed clusters are presented, including the mean values of the three variables for each cluster. Cluster 1 showed the highest mean value for awareness of the prohibition against defecating in public places (100), but showed low mean values for awareness of handwashing with soap and the management of drinking water and food (29.74 and 27.68, respectively). This finding suggests that the group in this cluster exhibits a high level of awareness regarding the practice of defecating in public spaces. However, the group demonstrates a lower level of proficiency in the other two variables. Conversely, Cluster 2 exhibited the lowest mean score for awareness of not defecating in public, with an average of 54.00. However, this cluster demonstrated relatively high mean scores for awareness of hand washing with soap and drinking water, as well as for food management, with averages of 75.22 and 81.54, respectively. These findings suggest that Cluster 2 exhibits a deficiency in the domain of public defecation awareness, while demonstrating proficiency in the domains of hand washing with soap and drinking water, as well as food management. Cluster 3 demonstrated high scores in all variables, namely awareness of open defecation, awareness of hand washing with soap, and drinking water and food management, with scores of 97.09, 81.14, and 81.27, respectively. These results indicate that this group exhibited consistent and high performance in all aspects measured.

An examination of this average value pattern reveals that Cluster 1 can be categorized as a low-level cluster. This is because, although Cluster 1 has the highest value in terms of awareness of not defecating indiscriminately, the other two variables show much lower numbers. Cluster 2 can be categorized as medium level, given that, although the value of awareness of not defecating indiscriminately is lower than the other clusters, the other variables are quite high, indicating equilibrium in several aspects. Conversely, Cluster 3 has been designated as a high-level cluster due to its superior performance in all variables, as indicated by its elevated scores in comparison to other clusters.

IV. CONCLUSION

The selection of variables was executed through the utilization of the clustvarsel algorithm, a methodology designed to identify the most pertinent variables for the purpose of clustering. After identifying the five variables representing the pillars of community-led total sanitation (CLTS), as well as the number of villages implementing CLTS, it was found that not all variables to be tested were entirely irrelevant. The selection of variables in the clustering process serves to focus on the most influential factors. Future research efforts may include a comprehensive examination of sanitation conditions by incorporating all five pillars or by introducing additional variables. Following variable selection, hierarchical clustering was performed using the Agglomerative Nesting (AGNES) method with four agglomeration approaches. The average agglomeration method was identified as the optimal approach based on the cophenetic correlation coefficient. The clustering results yielded three distinct clusters, each exhibiting unique sanitation characteristics. Cluster 1 can be categorized as a low-level cluster, Cluster 2 as a medium-level cluster, and Cluster 3 was designated as a high-level cluster due to its superior performance across all variables, as evidenced by its higher scores compared to other clusters.

Further research is needed to explore the influence of additional factors on sanitation behavior and to evaluate the long-term impact of sanitation programs. The selection of variables in this study serves as a methodological framework. However, future studies may re-evaluate or expand the selection criteria to ensure a more holistic understanding of sanitation issues. Additionally, the use of other clustering methods could be employed to improve the accuracy of the clustering. Based on this study, recommendations for improvement include enhancing public education, improving access to sanitation facilities, and integrating sanitation programs with broader public health initiatives to ensure the achievement of sustainable improvements in hygiene practices.

AVAILABILITY DATA AND MATERIALS

The dataset used can be accessed on BPS Provinsi Jawa Timur (Statistics of Jawa Timur Province) via the link at *https://jatim.bps.go.id/.*

REFERENCES

Buitrago-Boret, S. E., Martínez-Rivas, R., Florez-Diaz, J., Mijares-Seminario, R., & Rincón, E. (2023). Using cluster analysis on municipal statistical data to configure public policies about Water, Sanitation and Hygiene in Venezuela. arXiv preprint arXiv:2301.12604.

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing, 20(1), 270-281.

Gogebakan, M. (2021). A novel approach for Gaussian mixture model clustering based on soft computing method. IEEE Access, 9, 159987-160003.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric pollution research, 11(1), 40-56.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2023). Handbook of Cluster Analysis. CRC Press.

Lu, Z., & Lou, W. (2023). Bayesian approaches to variable selection in mixture models with application to disease clustering. Journal of Applied Statistics, 50(2), 387-407.

Maugis, C., Celeux, G., & Martin-Magniette, M. L. (2009). Variable selection for clustering with Gaussian mixture models. Biometrics, 65(3), 701-709.

Pereira, M. A., & Marques, R. C. (2021). Sustainable water and sanitation for all: are we there yet? Water Research, 207, 117765.

Purnama, M. D. (2023). Average Linkage-based Agglomerative Hierarchical Clustering terhadap Indikator Pembangunan Ekonomi Jawa Timur 2022. Jurnal Sains dan Seni ITS, 12(6), D477-D482.

Purnama, M. D. (2025). Cluster Analysis of Highest Education Completed in East Java Province with Spherical K-Means Method. Parameter: Journal of Statistics, 5(1), 61-67.

Purnama, M. D., & Sofro, A. Y. (2025). Implementation of agglomerative nesting and divisive analysis in East Java criminality rate hierarchical clustering. In AIP Conference Proceedings (Vol. 3316, No. 1, p. 040001). AIP Publishing LLC.

Redivo, E., Nguyen, H. D., & Gupta, M. (2020). Bayesian clustering of skewed and multimodal data using geometric skewed normal distributions. Computational Statistics & Data Analysis, 152, 107040.

Saraçlı, S., & Akşit, M. (2022). Comparison of hierarchic clustering methods with cophenetic correlation coefficient in big data. Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, 22(3), 552-559.

Tosunoglu, B. A., & Kocak, C. (2023). Feature selection for clustering and classification based attack detection systems in vehicular ad-hoc networks. Microprocessors and Microsystems, 104808.

United Nations. (2024). The Sustainable Development Goals Report 2024. United Nations Publications.