Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:71

https://doi.org/10.28919/cmbn/9235

ISSN: 2052-2541

DEPENDENCY-BASED CLUSTERING USING CORRESPONDENCE AND WARD'S HIERARCHICAL ANALYSIS: A CASE STUDY ON CLEAN WATER AND SANITATION INDICATORS IN WEST JAVA'S CITIES AND REGENCIES

THERESIA SAMARIA NAULI, IRLANDIA GINANJAR*, DEFI YUSTI FAIDAH

Department of Statistics, University of Padjadjaran, Bandung, Indonesia

Abstract: Correspondence analysis is a graphical technique for depicting relationships between variables in a low-dimensional space, making it ideal for non-metric data and non-linear associations. Multiple Correspondence Analysis (MCA) expands on this by identifying patterns in categorical variables, using the Burt matrix, a multidimensional contingency table. MCA aims for a cumulative variance of at least 70% across two dimensions; however, if this threshold is not met, Euclidean distance can improve object characterization with results extending beyond two dimensions. Though the dependency information from correspondence analysis is objective, it cannot reveal groups with fewer members based on available resources. Hence, cluster analysis is conducted using the principal coordinates from MCA results. This study aims to identify the unique characteristics of each object, allowing more focused evaluations based on specific attributes. By ensuring a cumulative variance of 100%, this method captures all relevant dependency information, offering a deeper understanding of variable relationships. The study stresses the importance of selecting the most suitable clustering model that aligns with the correspondence analysis results. By combining MCA and hierarchical clustering, the study visualizes and groups regencies and cities based on their clean water and sanitation conditions. Initial MCA results showed a cumulative variance of 22.3% in two dimensions, requiring further adjustments for more accurate interpretation. The innovation of this research lies in integrating MCA with hierarchical clustering using Euclidean distance to explore characteristics comprehensively. This method ensures a complete

*Corresponding author

E-mail address: irlandia@unpad.ac.id

Received March 08, 2025

representation of dependency relationships, maintaining a cumulative variance of 100%. A Euclidean distance matrix across 63 dimensions was used to enhance objectivity. The results identify 18 groups of regencies and cities with similar clean water and sanitation characteristics. Among clustering methods, the Ward method was most consistent with MCA findings. Cluster analysis was performed by forming three, four, and five clusters, aligning with government budget constraints.

Keywords: multiple correspondence analysis; cluster analysis; clean water; sanitation; Euclidean distance.

2020 AMS Subject Classification: 62H25.

1. INTRODUCTION

Correspondence analysis is a graphic technique primarily designed to represent associations in a low-dimensional space. Correspondence analysis differs from previously discussed interdependence techniques in its ability to accommodate nonmetric data and nonlinear relationships [1]. Multiple correspondence analysis (MCA) is used to identify relationships and patterns among more than two qualitative variables with categorical characteristics [2]. MCA is a correspondence analysis of a more complex indicator matrix where more than two different variables have been observed for each unit. The Burt matrix is a cross-tabulation that combines all variables from each category in a multidimensional contingency table [2]. This matrix is useful for comparing rows and columns of each variable resulting from the combination in the contingency table.

The Burt matrix yields a smaller primary coordinate scale compared to the indicator matrix. The Burt matrix is the square of the indicator matrix, so the percentage of inertia in the Burt matrix will consistently be higher or optimal compared to the indicator matrix [2]. The minimum criterion for a good cumulative percentage of variance in two dimensions is 70% [3]. Suppose the representation of the two dimensions resulting from the MCA is less than 70%. In that case, the identification of information regarding the characteristics of each object is done using Euclidean distance. In previous research, the identification of information about the similarity between objects and the associations between object characteristics was carried out using the modified Mahalanobis distance [4]. As the Mahalanobis distance is less suitable for this context, where the variance of each dimension was accounted for, as the distance matrix was built from principal coordinates. In correspondence analysis, the distances between principal coordinate points already take into account the variance across dimensions, making Euclidean distance a more appropriate measure.

The Euclidean distance is the straight-line distance between two objects being studied [1]. In the previous study, the cumulative variance of two dimensions from the MCA was only 16%, which is significantly below the acceptable threshold of 70%. Therefore, the analysis utilized Euclidean distance to interpret districts based on their correspondence characteristics. Kristanto *et al.* [5] identified sub-district group characteristics for the Environmental Quality Index in Bandung Regency based on dependency relationships using Joint Correspondence Analysis, and the study produced a two-dimensional map with a cumulative variance of 70.1%. The JCA map is generated from a new Burt matrix (not the original one), so the resulting cumulative variance does not represent the variance of the original data. Rosa *et al.* [6] used a two-dimensional map as the result of correspondence and cluster analysis provided misleading information, as the cumulative variance obtained from the two dimensions only reached 55.9%. Correspondence analysis methods can represent dependencies between cities/regencies and conditions of clean water and sanitation. MCA can transform categorical variables into continuous coordinates, enabling the application of cluster analysis algorithms to both numerical and categorical data to group objects based on similar characteristics [7].

Cluster analysis is used to examine the similarity of observations among the studied subjects with the aim of forming groups of similar subjects, thereby creating partitions or sequences of partitions of these subjects. Cluster analysis is divided into two methods: hierarchical and non-hierarchical methods [8]. Kim *et al.* [9] have compared proposed methods through simulations to evaluate their effectiveness. For instance, one study suggested a method combining Multiple Correspondence Analysis (MCA) and K-means Cluster Analysis (CA) and compared it with existing methods that also combine MCA and CA, methods applying MCA and CA sequentially, and methods using only CA. The simulation results demonstrated that the proposed method outperformed or performed equally well compared to other methods in identifying true clusters. Florensa *et al.* [10] explored associations between risk factors and the likelihood of colorectal cancer using correspondence analysis and non-hierarchical K-means cluster analysis, highlighting the utility of these techniques in uncovering meaningful patterns in complex data. However, non-hierarchical cluster analysis was not used due to the small number of observation objects.

Additionally, this method does not produce a dendrogram graph and results in inconsistent clustering outcomes with each increase in the number of clusters. Meanwhile, the study requires clustering of regencies/cities with varying numbers of clusters and consistent group memberships. Hierarchical cluster analysis is appropriate for relatively small datasets, as it allows for clear

visualization using a dendrogram. This method effectively reveals high similarity among members within the same group and low similarity between different groups. Additionally, hierarchical clustering helps enhance the interpretation of cluster relationships by clearly displaying the hierarchical structure of the data, making it easier to identify distinct clusters and their interrelationships [11]. When the number of objects to be clustered is small, dendrograms can still be effectively used, as the visualization remains clear and easy to interpret without issues like overlapping labels or densely packed branches [12].

MCA was used to analyze relationships between care arrangement variables by transforming categorical data into continuous coordinates, thereby reducing data dimensionality. However, in some cases, the resulting groups remained numerous and less interpretable, leading to the application of cluster analysis as an alternative solution to creating more distinct and manageable groups based on the MCA coordinates [13].

Hierarchical clustering methods proceed with a series of sequential merges or a series of sequential splits. Most hierarchical methods typically start with individual objects. Therefore, initially, there are groups equal to the number of objects present. Objects with the highest similarity are grouped first, and these initial groups are combined based on their similarity. As the similarity decreases, eventually, all subgroups are merged into a single group [1]. Previous studies needed to adequately address the importance of consistency with the results of correspondence analysis grouping. One study employed hierarchical cluster analysis and used clustering coefficients from each output as a criterion to identify the best clustering method. Still, it did not explicitly evaluate alignment with correspondence analysis results [14]. Rosa *et al.* [6] directly used Ward's clustering method without exploring other clustering approaches, similarly overlooking the need to determine the most consistent method with the grouping results from correspondence analysis.

Brelle [15] indicates that addressing water and sanitation issues is indeed part of the Sustainable Development Goals (SDGs), which serve as a foundation for realizing Indonesia's national goals. SDGs cover social and economic development issues, including access to water and sanitation, as part of the 17 global goals adopted by the United Nations in 2015 [16]. Gulseven [16] also suggests that achieving these SDGs, including goals related to water and sanitation, requires a collaborative effort among individuals, businesses, and governments. Additionally, Crawford [17] highlights the importance of addressing gendered social norms, particularly around unpaid care work, in order to achieve universal access to water and sanitation (SDG 6).

Research on inequality in access to drinking water and sanitation at the provincial and district/city

levels in Indonesia, based on the 2015 National Socio-Economic Survey (SUSENAS) data, has been conducted by Afifah *et al.* [18]. However, the study was unable to identify specific regions that should be prioritized for improvement. Clustering can be used to evaluate and maintain successful water and sanitation programs. Cluster results can guide the government in policy-making and program development tailored to the characteristics of each cluster.

Additionally, Ortigara *et al.* [19] emphasizes the importance of well-targeted government programs to ensure the achievement of SDG 6 targets in the areas of clean water and sanitation. The report highlights that program integration and precise targeting are essential to address disparities in access to and quality of water and sanitation services, particularly for vulnerable groups. It also underscores the crucial role of governments in integrating effective and sustainable water governance with active participation from the scientific community, international organizations, and various other stakeholders.

This study aims to identify the characteristics of each object by examining the dependency relationships among the categorical variables of these conditions. By applying correspondence analysis and hierarchical cluster analysis, this study will group the cities/regencies based on similarities in clean water and sanitation characteristics. The findings are expected to offer a more comprehensive understanding of the water and sanitation situation across different regions in West Java. This clustering outcome can provide the government with valuable insights for designing targeted policies and programs to improve water quality and sanitation conditions in each area according to their specific needs and priorities.

The novelty of this study lies in its use of MCA combined with hierarchical cluster analysis to provide a more objective and comprehensive understanding of the regional characteristics related to clean water and sanitation. If the correspondence analysis results in a two-dimensional representation with low cumulative variance, it indicates that significant information regarding dependency relationships needs to be accurately captured. To address this limitation, it is essential to achieve a cumulative variance of at least 70%, which allows for a more precise explanation of the dependencies among variables.

However, achieving 100% cumulative variance is optimal, as it ensures that all relevant information from the dataset is accounted for, allowing for a complete understanding of the dependency relationships among the variables. This study lies in its ability to extract dependency information between categories across more than three dimensions using Euclidean distance. This approach facilitates a more thorough examination of the similarities among characteristics,

ensuring an accurate representation of the data.

Additionally, hierarchical cluster analysis is employed to form fewer, more distinct groups based on the MCA results. The best clustering model is selected by comparing which clustering outcome most closely aligns with the output of the correspondence analysis. This combined approach not only enhances the overall accuracy of the analysis but also offers a refined perspective on regional disparities in water and sanitation conditions, enabling the government to implement more precise and targeted interventions tailored to the specific needs of each regency and city.

2. MATERIAL AND METHOD

A. Data Sources

The data used in this research is secondary data obtained from the West Java Diskominfo. This study uses the 2022 National Socio-Economic Survey (SUSENAS) data, which is useful for planning national development programs, monitoring and evaluating national development programs, and providing important strategic indicators to measure the achievement of development goals/SDGs. The data consists of 27 cities/regencies and 25,744 households. The data is qualitative with nominal and ordinal measurement scales. Seven characteristics variables serve as column categories, including variables related to ownership of toilet facilities and who uses them (X_1), types of toilets used (X_2), final disposal sites for feces (X_3), the frequency of septic tanks being emptied or pumped out in the last 5 years (X_4), the main water source used by households for drinking (X_5), the main water source used by households for cooking/bathing/washing/other purposes (X_6), and the distance to the nearest waste/excrement/feces disposal site (X_7). Furthermore, cities/regencies data are utilized as row categories. The 8 characteristics were converted into a contingency table with cities/regencies as the rows and each characteristic as the columns. This process produced 8 contingency tables, which can be employed for the chi-square test.

B. Contingency Table

In this study, the data is presented in the form of a two-way contingency table, with the cities/ regencies categories as row (M) and the indicators of clean water and sanitation as columns (X). q_1 is the number of categories for the row variable (cities/ regencies) with $j = 1,2, \dots, q_1, q_{\tilde{k}}$ is the number of categories for the columns variable (characteristics) with $\tilde{j} = 1,2, \dots, q_{\tilde{k}}, \bar{k} = 2,3, \dots, u$, individual in the data is n, and $n_{j\tilde{j}}$ is the number of observation (cities/regencies). Each cell in this contingency table contains information regarding the frequency of households in West

Java as the unit of observation in the context of the relevant categories in this study. The results, shown in Table 1, form the following contingency table.

Table 1. Contingency Table

Cities/Regencies (M)	Characteristic Variables (X)						
	1	2	...	$\tilde{j}$	...	$q_{\tilde{k}}$	Total
1	n_{11}	n_{12}	...	$n_{1\tilde{j}}$	...	$n_{1q_{\tilde{k}}}$	$n_{1\bullet}$
2	n_{21}	n_{22}	...	$n_{2\tilde{j}}$	...	$n_{2q_{\tilde{k}}}$	$n_{2\bullet}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
j	n_{j1}	n_{j2}	...	$n_{j\tilde{j}}$	...	$n_{jq_{\tilde{k}}}$	$n_{j\bullet}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
q_1	$n_{q_1 1}$	$n_{q_1 2}$	...	$n_{q_1\tilde{j}}$	...	$n_{q_1 q_{\tilde{k}}}$	$n_{q_1\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet\tilde{j}}$	...	$n_{\bullet q_{\tilde{k}}}$	n

Where q_1 represents the number of categories for variable 1, which is the row variable (cities/regencies), $q_{\tilde{k}}$ represents the number of categories for variables 2, 3, 4, …, u which are the column variables (characteristic indicators of clean water and sanitation), and n represents the household frequency. Based on Table 1, the cross-tabulation matrix $\mathbf{N}$ can be calculated using the following equation [5].

$$\mathbf{N} = (n_{j\tilde{j}}) \tag{1}$$

Where $n_{j\tilde{j}}$ represents the elements of the cross-tabulation matrix, j is the category of the city/regency variable with $j = 1,2,\ldots,q_1$, and $\tilde{j}$ is the category of the characteristic variable with $\tilde{j} = 1,2,\ldots,q_{\tilde{k}}$. Based the cross-tabulation matrix $\mathbf{N}$ on equation (1), the correspondence matrix $\tilde{\mathbf{P}}$ is obtained as follows [5].

$$\tilde{\mathbf{P}} = \frac{1}{n}\mathbf{N} = (\tilde{p}_{j\tilde{j}}) = \left(\frac{n_{j\tilde{j}}}{n}\right) \tag{2}$$

Where n represents the number of observations, $\tilde{p}_{j\tilde{j}}$ is join probability estimator cities/regencies and characteristics variable. $\tilde{p}_{j\bullet}$ is marginal probabilty estimator of characteristics and $\tilde{p}_{\bullet\tilde{j}}$ is marginal probability estimator of cities/regencies. The calculation of the marginal probability estimator values of characteristics and cities/regencies are as follows: [3].

$$\tilde{p}_{j\bullet} = \frac{n_{j\bullet}}{n} \quad \text{and} \quad \tilde{p}_{\bullet\tilde{j}} = \frac{n_{\bullet\tilde{j}}}{n} \tag{3}$$

Table 2 presents the data used in the study, and the complete data can be viewed at the following link bit.ly/datawatersanitation.

Table 2. Qualitative Data on Drinking Water and Sanitation Indicators in West Java, 2022

Number	City/Regency	X_1	X_2	X_3	...	X_7
1	Cianjur	Available, used only by household members	Saucer with lid	Ground hole	...	<10 meters
2	Cianjur	Available, used only by household members	Gooseneck	Septic Tank	...	<10 meters
3	Cianjur	Available, shared with specific household members	Falling into a pit	Ground hole	...	<10 meters
4	Cianjur	Available, used only by household members	Gooseneck	Septic Tank	...	<10 meters
5	Cianjur	Available, used only by household members	Gooseneck	Ground hole	...	<10 meters
⋮	⋮	⋮	⋮	⋮	⋱	⋮
25744	Indramayu	No facilities available	Others	Others	...	<10 meters

Data Source: National Socioeconomic Survey 2022

C. Analysis of Independence with the Chi-Square Test

The purpose of the independence test is to determine the relationship between the row categories and the column categories. The Pearson Chi-square test is used for two-way contingency tables. If the results of the independence test indicate an influence or dependence between variables, then correspondence analysis can be applied to examine the contribution of each category. The steps for conducting the Chi-square independence test are as follows [3]:

Hypothesis:

H_0: $\pi_{j\tilde{j}} = \pi_{j\bullet}\pi_{\bullet\tilde{j}}$ (There is no dependence between the city/regency variable and the variable of clean water and sanitation characteristics).

H_1: $\pi_{j\tilde{j}} \neq \pi_{j\bullet}\pi_{\bullet\tilde{j}}$ (There is dependence between the city/regency variable and the variable of clean water and sanitation characteristics).

The test statistic uses the Chi-square with the following equation [20]:

$$\chi^2 = n\sum_{j=1}^{q_1}\sum_{\tilde{j}=1}^{q_{\tilde{k}}}\frac{\left(\tilde{p}_{j\tilde{j}} - \tilde{p}_{j\bullet}\tilde{p}_{\bullet\tilde{j}}\right)^2}{\tilde{p}_{j\bullet}\tilde{p}_{\bullet\tilde{j}}} \tag{4}$$

Where n is the number of observations, $\tilde{p}_{j\bullet}$ is the marginal probability of the j^{th} cities/regencies from equation (3), $\tilde{p}_{\bullet\tilde{j}}$ is the marginal probability of $\tilde{j}^{th}$ characteristics from equation (3), $\tilde{p}_{j\tilde{j}}$ is the joint probability of the j and $\tilde{j}$ from equation (3), q_1 is the number of categories on j^{th}, and $q_{\tilde{k}}$ is the number of categories on $\tilde{j}^{th}$. The test criterion for the Pearson

Chi-square test with $\chi^2 \sim \chi^2_v$, $v = (q_1 - 1)(q_{\tilde{k}} - 1)$is to reject H_0 if the $p - value < \alpha$, otherwise, H_0 is accepted. The rejection of H_0 is identified based on the $p - value$ with the following equation:

$$p - value = P\{\chi^2_v > \chi^2\} \quad (5)$$

D. Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is used to identify relationships and patterns among more than two qualitative variables with categorical characteristics [2]. MCA is an extension of correspondence analysis for more complex indicator matrices, where more than two different variables are observed for each unit. There are two distinct methods for performing MCA, using the indicator matrix or the Burt matrix. In an indicator matrix, the value 0 represents that an object does not belong to a certain category, while the value 1 indicates that an object belongs to that category [4]. The matrix $\mathbf{Y} = \left(y_{ij}\right)$ is a matrix with dimensions $N \times u$, where N represents the number of observation units such as households, and u represents the number of variables. Here, $i = 1,2,..N$ and $k = 1,2,\ldots,u$. If q_k is the number of categories for the k^{th} variable, then $\boldsymbol{Z}_k = \left(z_{ijk}\right)$ is an indicator matrix for the k^{th} variable with dimensions $N \times q_k$, where z_{ijk} represents the element at position (i,j) in $\boldsymbol{Z}_k$, and $j = 1,2,..,q_k$

In this context, the indicator matrix can be represented as the following equation [2]:

$$\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \ldots \quad \mathbf{Z}_\mathrm{u}] \quad (6)$$

The Burt matrix has a symmetric property, reflecting the two-way cross-tabulation of all combinations of categories within qualitative variables. Suppose there are u qualitative variables, and $\mathbf{Z}_u$ represents the indicator matrix for the u^{th} qualitative variable with $k = 1,2,3,\ldots,u$. The combined indicator matrix for the qualitative variables, denoted as $\mathbf{Z}$. The process of cross-tabulating the combined indicator matrix of the qualitative variables ensures that rows and columns of all variables are cross-tabulated. The result of this process is known as the Burt matrix. The form and calculation of the Burt matrix are described as follows [2]:

$$\mathbf{B} = \mathbf{Z}^\mathrm{T}\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \ldots & \mathbf{Z}_1^T\mathbf{Z}_{\tilde{k}} & \ldots & \mathbf{Z}_1^T\mathbf{Z}_u \\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{Z}_2 & \ldots & \mathbf{Z}_2^T\mathbf{Z}_{\tilde{k}} & \ldots & \mathbf{Z}_2^T\mathbf{Z}_u \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{Z}_k^T\mathbf{Z}_1 & \mathbf{Z}_\mathrm{k}^T\mathbf{Z}_2 & \ldots & \mathbf{Z}_k^T\mathbf{Z}_{\tilde{k}} & \ldots & \mathbf{Z}_k^T\mathbf{Z}_u \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{Z}_u^T\mathbf{Z}_1 & \mathbf{Z}_u^T\mathbf{Z}_2 & \ldots & \mathbf{Z}_u^T\mathbf{Z}_{\tilde{k}} & \ldots & \mathbf{Z}_u^T\mathbf{Z}_u \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{N}_{12} & \ldots & \mathbf{N}_{1\tilde{k}} & \ldots & \mathbf{N}_{1u} \\ \mathbf{N}_{21} & \mathbf{D}_2 & \ldots & \mathbf{N}_{2\tilde{k}} & \ldots & \mathbf{N}_{2u} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{N}_{k1} & \mathbf{N}_{k2} & \ldots & \mathbf{D}_k & \ldots & \mathbf{N}_{k\mathrm{u}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{N}_{u1} & \mathbf{N}_{u2} & \cdots & \mathbf{N}_{u\tilde{k}} & \ldots & \mathbf{D}_u \end{bmatrix} = (b_{m\tilde{m}}) \quad (7)$$

Where,

$$\mathbf{N}_{k\tilde{k}} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1q_{\tilde{k}}} \\ n_{21} & n_{22} & \cdots & n_{2q_{\tilde{k}}} \\ \vdots & \vdots & \ddots & \vdots \\ n_{q_k1} & n_{q_k2} & \cdots & n_{q_kq_{\tilde{k}}} \end{bmatrix} \tag{8}$$

Where, $\boldsymbol{D}_k = \mathrm{diag}(d_{\mathrm{jk}})$ with d_{jk} representing the marginal frequency of the j^{th} category of the k^{th} variable for $j = 1,2,\dots,q_k$ and $k,\tilde{k} = 1,2,\dots,u$. $\mathbf{N}_{k\tilde{k}}$ is the cross-tabulation matrix between the k^{th} variable and the $\tilde{k}$ variable, $b_{m\tilde{m}}$ represents the element of the Burt matrix where $m,\tilde{m} = 1,2,\dots,Q$, and $\mathbf{Z}_u^T\mathbf{Z}_{\tilde{u}}$ is the diagonal matrix of the total frequency for $\mathbf{Z}_u$.

After obtaining the Burt matrix, each element in the Burt matrix is divided by the total sum of all the elements in the matrix. This matrix is known as the Burt correspondence matrix. The formula used to calculate the Burt correspondence matrix is as follows [2]:

$$\mathbf{P} = \frac{1}{g}\mathbf{B} = (p_{m\tilde{m}}) \tag{9}$$

Where, $g = \sum_{m=1}^{Q}\sum_{\tilde{m}=1}^{Q} b_{m\tilde{m}}$ and $p_{m\tilde{m}}$ is the element of the Burt correspondence matrix for the m^{th} row and the $\tilde{m}^{th}$ column, involving the proportion of the Burt matrix columns, indicating the ratio between one category and all existing categories. The proportion of the Burt matrix columns (**c**) and the row mass of the Burt matrix (**r**) have equivalent values, expressed through the following formula [2]:

$$\mathbf{r} = \mathbf{c} = \frac{1}{g}\mathbf{B1} \tag{10}$$

Where $\mathbf{1}$ is a vector of dimension $Q \times 1$ with each component having a value of 1. The main diagonal elements in the Burt correspondence matrix reflect the marginal probabilities, while the upper and lower triangular regions of the correspondence matrix indicate joint probabilities. The row total vector $(\mathbf{r})$ is formed by summing each row in the Burt correspondence matrix, while the column total vector $(\mathbf{c})$ is formed by summing each column. The calculation of the standard residual matrix can be described as follows [2]:

$$\mathbf{S} = \mathbf{D_c}^{-\frac{1}{2}}(\mathbf{P} - \mathbf{c}\mathbf{c}^{\mathrm{T}})\mathbf{D_c}^{-\frac{1}{2}} = (s_{m\tilde{m}}) \tag{11}$$

P is the Burt correspondence matrix from equation (9) is the row diagonal matrix, $\mathbf{D_c}$ is the column diagonal matrix, and **c** is the row or column mass consisting of Q vectors. Because in the Burt matrix **B**, the ratio of the proportion of row and column mass has the same value, then the result $\mathbf{D_r} = \mathbf{D_c} = \mathrm{diag}(\mathbf{c})$ and $\mathbf{P1} = \mathbf{c}$ and $\mathbf{1}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}$.

The representation of values and diversity between categories m and $\tilde{m}$ is denoted by $s_{m\tilde{m}}$, which is an element of the residual matrix. The decomposition of the standard residual matrix **S**

is carried out to ensure orthogonal mapping. Eigenvalue decomposition (EVD) is applied because the matrix $\mathbf{S}$ is symmetric and the information from row and column categories is equivalent. The combined eigenvectors yield a set of new variables. The EVD of matrix $\mathbf{S}$ is outlined as follows:

$$\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathbf{T}} \tag{12}$$

Where $\mathbf{V}$ is an orthogonal matrix, meaning $\boldsymbol{V^{-1}} = \boldsymbol{V^{T}}$, and therefore $\mathbf{V}\mathbf{V}^{\mathbf{T}} = \mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$. Each column of $\mathbf{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_L)$ represents an eigenvector that is orthogonal to each other. $\boldsymbol{\Lambda}$ is a diagonal matrix containing eigenvalues (λ_ℓ) in descending order. Mathematically, this can be expressed as $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. Where, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_L)$ and λ_ℓ is the ℓ^{th} eigenvalue of $\mathbf{S}$, arranged such that $\lambda_1, > \lambda_2 > \cdots > \lambda_\ell > \cdots > \lambda_L, \ell = 1,2,3, \dots, L$.

The first step in generating a correspondence map is to obtain the principal coordinates for each category, which illustrate the relationships between categories. The standard coordinates of the row categories and the standard coordinates of the column categories will have the same values. This occurs because the Burt matrix, from which these coordinates are derived, is symmetric. Thus,

the standard coordinates of both the row and column categories can be formulated as follows.

$$\mathbf{H} = \mathbf{D_c}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{V} = (h_{m\ell}) \tag{13}$$

Where ℓ represents the dimension and mmm represents the number of categories. Each row of the matrix represents the standard coordinates for each dimension. The principal coordinates are important because the eigenvalues have been weighted onto these coordinates. Based on the Burt matrix, the principal coordinates for individuals or objects can be formulated as follows.

$$\mathbf{F} = \mathbf{D_c}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{V}\boldsymbol{\Lambda}^{\frac{\mathbf{1}}{\mathbf{2}}} = (f_{m\ell}) \tag{14}$$

Where **Dc** is the diagonal matrix of column masses **c** and each row in matrix **F** represents a category, while the columns in matrix $\mathbf{F}$ represent the coordinates for each dimension. Inertia is a factor that indicates the extent to which the variation is explained by these dimensions. The quality of the mapping can be evaluated from the total inertia, which reflects the percentage of categories or information that is not represented [21]. The total inertia can be formulated as follows.

$$\text{Total Inertia} = \text{trace}\,(\mathbf{F}^{\mathrm{T}}\mathbf{F}) = \text{trace}(\boldsymbol{\Lambda}) \tag{15}$$

The variance coverage for each dimension is denoted as follows:

$$\phi_\delta = \left(\frac{\lambda_\delta}{\sum_{\ell=1}^{L}\lambda_\ell}\right) \tag{16}$$

$$\tau_D = \left(\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_D}{\sum_{\ell=1}^{L} \lambda_\ell} \right) \tag{17}$$

Where τ_D is the percentage of variance from D dimensions or variance coverage, ϕ_δ is the variance coverage for each dimension where $\delta = 1,2, \dots, L, \lambda_\delta$ is the eigenvalue obtained from the EVD, and λ_ℓ is the ℓ^{th} eigenvalue. Inertia can reflect the quality of the generated map. A two-dimensional map can be created when the inertia percentage in two dimensions reaches 70% [22]. Two-dimensional maps are valuable as they provide insights into data from three dimensions and beyond [23].

E. Euclidean Distance

The steps to be taken to cluster cities/regencies using cluster analysis involve utilizing a distance matrix calculated from the principal coordinates obtained from the previous correspondence analysis. If there are q_1 row variable categories (objects) with m, $\tilde{m} = 1,2, \dots, q_1, \dots, Q$, then the vectors $\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_Q$ can be calculated, and the Euclidean distance as the elements of the matrix **D**. Thus, a matrix of size $Q \times Q$, denoted as **D**, is obtained as follows.[24]:

$$\mathbf{D} = \begin{bmatrix} \boldsymbol{D_{11}} & \boldsymbol{D_{12}} \\ \boldsymbol{D_{21}} & \boldsymbol{D_{22}} \end{bmatrix} = \left(d(\mathbf{f}_{\mathrm{m}}, \mathbf{f}_{\tilde{\mathrm{m}}}) \right) \tag{18}$$

Where $d(\mathbf{f}_m, \mathbf{f}_{\tilde{m}})$ is the Euclidean distance between vector $\mathbf{f}_m$ and vector $\mathbf{f}_{\tilde{m}}$, where $\mathbf{f}_m$ is the vector for category-m and $\mathbf{f}_{\tilde{m}}$ is the vector for category-$\tilde{m}$. D_{11} is a matrix of size $q_1 \times q_1$ that

$$d(\mathbf{f}_m, \mathbf{f}_{\tilde{m}}) = \sqrt{(\mathbf{f}_m - \mathbf{f}_{\tilde{m}})^{\mathrm{T}} (\mathbf{f}_m - \mathbf{f}_{\tilde{m}})}$$

represents the distance between objects, while D_{12} is a matrix of size $q_1 \times (Q - q_1)$ that represents the distance between objects and their characteristics.

D. Clustering Using Hierarchical Cluster Analysis

This approach involves applying hierarchical cluster analysis by utilizing the distance matrix calculated from the principal coordinates. By using the principal coordinate values obtained from the multiple correspondence analysis results, and since the focus of the clustering in this study is on cities/regencies, the resulting distance matrix will have dimensions of 27×27. The best clustering method, obtained from the principal coordinates in the correspondence analysis, will be selected based on the consistency of the clustering results for cities/ regencies when compared to the clustering results from the correspondence analysis. Consistent clustering results can be observed through the similarity of the clustering outcomes for the cities/regencies. The similarity in the grouping results of regencies/cities between correspondence analysis and the best cluster

analysis from the principal coordinates obtained from the correspondence analysis indicates that the two methods support each other [25].

The clustering results based on Euclidean distances from the correspondence analysis can be used to determine the number of clusters to be utilized. To demonstrate the consistency of the city/regency clustering based on the same characteristics, the number of clusters will be determined according to the number of groups formed in the correspondence analysis. The advantage of using cluster analysis is that it allows for the determination of the desired number of groups according to the availability of resources and manpower. Therefore, this study will try several numbers of clusters to compare the number of programs that the government needs to implement based on the issues of clean water and sanitation in the formed city/regency groups.

3. MAIN RESULTS

The data comprising eight characteristic variables is organized into contingency tables, resulting in 8 contingency tables corresponding to each characteristic variable. In conducting correspondence analysis, these characteristic variables must be dependent on the city/regency variable. In this study, $\alpha = 5\%$ is used, meaning a confidence level of 95% for the decision made in hypothesis testing and the degree of freedom. Hence, a Chi-Square test is employed to assess the relationship between the characteristic variables and the city/regency variable. The outcomes of the Chi-Square Test from equations (4) and (5) processed using R software version 1.4.1106 are presented in Table 3. The complete syntax can be found in bit.ly/syntaxjournal.

Table 3. Chi-Square Test

Category	$\chi - square$ (χ^2)	df (v)	p-value
City/Regency vs (X_1)	2171.4	130	< 2.2e-16
City/Regency vs (X_2)	2439.5	104	< 2.2e-16
City/Regency vs (X_3)	11001	156	< 2.2e-16
City/Regency vs (X_4)	4430.5	78	< 2.2e-16
City/Regency vs (X_5)	15527	260	< 2.2e-16
City/Regency vs (X_6)	13355	260	< 2.2e-16
City/Regency vs (X_7)	4368.9	78	< 2.2e-16

Based on the results in Table 3, all characteristic variables related to clean water and sanitation show a significant dependence on the city/regency variable. Therefore, all these characteristic variables will be utilized in this study. The study proceeds by using all characteristic variables in the multiple correspondence analysis. The cumulative percentage of variance reaches 100% at the 63rd dimension. The principal coordinate matrix $\mathbf{F}$ for all categories is obtained by multiplying

the standardized coordinates by the matrix $\mathbf{\Lambda}$, which is a diagonal matrix of the eigenvalues (λ_ℓ) arranged in descending order according to equation (14).

$$\mathbf{F}_{(75\times 63)} = \begin{bmatrix} 0.1467 & 0.1956 & -0.0054 & -0.0053 & \cdots & -0.0014 \\ 0.5159 & 0.1878 & -0.5034 & -0.2913 & \cdots & -0.0008 \\ 0.5285 & 0.1579 & -0.4617 & -0.1958 & \cdots & -0.0031 \\ 0.0417 & 0.0684 & -0.0736 & 0.1924 & \cdots & -0.0045 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.2545 & 0.0900 & 0.0803 & 0.0114 & \cdots & -0.0020 \end{bmatrix} \tag{20}$$

The principal coordinates for all categories are calculated by multiplying the standardized coordinates by the matrix $\mathbf{\Lambda}$, which is a diagonal matrix of the eigenvalues λ_ℓ according to equations (17). The percentage of variance results and variance coverage for each dimension based on equation (16) and (17) are as follows.

Table 4. Percentage of Variance

Dim	λ_ℓ	ϕ_δ	τ_D
1	0.203369	14.3	14.3
2	0.114221	8.0	**22.3**
3	0.084426	5.9	28.3
⋮	⋮	⋮	⋮
63	0.000440	0.0	100

Table 4 shows the total cumulative percentage of variance obtained using two dimensions is 22.3%. If only two dimensions are used, the information obtained might be misleading. Therefore, the percentage of variance results form the basis for determining the relevant coordinate dimensions for the qualitative characteristics.

Next, the clustering process is based on the dependence of clean water and sanitation characteristics on the 27 cities/regencies in West Java, determined by the distances between points. The closer the distance, the more representative the category is of the clean water and sanitation characteristics in that city/regency, while a greater distance indicates the opposite.

Table 5. Euclidean Distance Between Cities/Regencies and Characteristic Variables

M	$X_{1.1}$	...	$X_{3.1}$	$X_{3.2}$	$X_{3.3}$	$X_{3.4}$	$X_{3.5}$	$X_{3.6}$	$X_{3.7}$	...	$X_{7.1}$	$X_{7.2}$	$X_{7.3}$	$X_{7.4}$
M_1	1.59	...	**1.61**	3.42	1.96	1.97	10.4	4.55	2.76	...	2.17	1.66	**1.57**	2.08
M_2	1.85	...	1.96	3.54	2.12	**1.73**	10.7	4.56	2.78	...	2.36	1.89	**1.83**	2.22
M_3	1.82	...	1.92	3.53	1.99	**1.84**	10.6	4.61	2.75		2.33	1.84	**1.78**	2.27
⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
M_{20}	2.81	...	2.34	3.69	2.20	2.60	10.8	4.83	3.40	...	2.71	2.29	2.32	2.59
⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
M_{27}	2.39	...	2.36	3.88	2.74	2.78	10.8	4.25	2.55	...	2.71	2.47	2.41	2.85

Table 5 is the results of the Euclidean matrix for the correspondence analysis based on equation (18). It can be seen that Bogor Regencies (M_1) has the nearest Euclidean distance to category $X_{3.1}$ compared to its distance to other categories within variable X_3. This can be interpreted to mean that one of the characteristics of Bogor Regencies is that most households have septic tanks as their final wastewater disposal$(X_{3.1})$. The same steps are applied to other cities/regencies and categories

Cities/regencies that share similar characteristics can be grouped. Table 6 presents this approach, which simplifies the identification of cities/regencies based on their clean water and sanitation characteristics, with the following grouping:

Table 6. Identification of Clean Water and Sanitation Characteristics in Cities/Regencies

Group	City/Regency	Characteristics of Clean Water and Sanitation
1	• Kuningan (M_8) • Pangandaran (M_{18}) • Banjar city (M_{27})	• Has a toilet facility, used only by household members $(X_{1,1})$. • The type of toilet used is a gooseneck closet $(X_{2,1})$. • The final disposal site for feces is a septic tank $(X_{3,1})$. • *In the past 5 years, the septic tank has never been emptied $(X_{4,3})$. • *The main source of drinking water for the household is bottled water $(X_{5,2})$. • The main water source used for cooking/bathing/washing/others is a protected well $(X_{6,5})$. • Distance to waste/sewage/feces storage site is ≥ 10 m $(X_{7,3})$
2	• Bogor (M_1) • Ciamis (M_7)	• Has a toilet facility, used only by household members $(X_{1,1})$. • The type of toilet used is a gooseneck closet $(X_{2,1})$. • The final disposal site for feces is a septic tank $(X_{3,1})$. • *In the past 5 years, the septic tank has never been emptied $(X_{4,3})$. • The main water source used by the household for drinking is a protected well $(X_{5,5})$. • The main water source used for cooking/bathing/washing/others is a protected well $(X_{6,5})$. • Distance to the waste/sewage/feces storage site is ≥ 10 m $(X_{7,3})$.
3	• Sukabumi (M_2) • Cianjur (M_3)	• Has a toilet facility, used only by household members $(X_{1,1})$. • The type of toilet used is a gooseneck closet $(X_{2,1})$. • *The final disposal site for feces is a ground hole $(X_{3,4})$. • *In the past 5 years, the septic tank has been emptied fewer than 6 times $(X_{4,1})$. • The main water source used by the household for drinking is a protected well $(X_{5,5})$. • The main water source used for cooking/bathing/washing/others is a protected well $(X_{6,5})$. • Distance to the waste/sewage/feces storage site is ≥ 10 m $(X_{7,3})$.

Group	City/Regency	Characteristics of Clean Water and Sanitation
4	• Bandung (M_4)	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
5	• Garut (M_5)	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The final disposal site for feces is a pond/field/river/lake/sea ($X_{3,3}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • The main water source used by the household for drinking is a protected well ($X_{5,5}$). • The main water source used for cooking/bathing/washing/others is a protected well ($X_{6,5}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
6	• Tasikmalaya (M_6)	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • The main water source used by the household for drinking is a protected well ($X_{5,5}$). • The main water source used for cooking/bathing/washing/others is a protected well ($X_{6,5}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
7	• Cirebon (M_9) • Indramayu(M_{12}) • Karawang (M_{15}) • Bekasi (M_{16})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,2}$)
8	• Majalengka (M_{10})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).

Group	City/Regency	Characteristics of Clean Water and Sanitation
9	• Sumedang (M_{11})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a protected spring ($X_{6,7}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
10	• Subang (M_{13}) • Depok City (M_{24})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • The main water source used by the household for drinking is a bore well/pump ($X_{5,4}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,2}$).
11	• Purwakarta (M_{14})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The final disposal site for feces is a ground hole ($X_{3,4}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
12	• West Bandung (M_{17})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The final disposal site for feces is a ground hole ($X_{3,4}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
13	• Bogor City (M_{19}) • Cirebon City (M_{22})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • The main water source used by the household for drinking is piped water ($X_{5,3}$). • The main water source used for cooking/bathing/washing/others is piped water ($X_{6,3}$). • The main drinking water source is not a well/pump/spring ($X_{7,1}$).

Group	City/Regency	Characteristics of Clean Water and Sanitation
14	● Sukabumi City (M_{20})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The final disposal site for feces is a pond/field/river/lake/sea ($X_{3,3}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,3}$).
15	● Bandung City (M_{21})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The final disposal site for feces is a pond/field/river/lake/sea ($X_{3,3}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
16	● Bekasi City (M_{23})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • In the past 5 years, the frequency of septic tank emptying is unknown ($X_{4,4}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,2}$).
17	● Cimahi City (M_{25})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • In the past 5 years, the frequency of septic tank emptying is unknown ($X_{4,4}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).
18	● Tasikmalaya City (M_{26})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • *In the past 5 years, the septic tank has been emptied fewer than 6 times ($X_{4,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$). • The main water source used for cooking/bathing/washing/others is a protected well ($X_{6,5}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,2}$).

* The category still requires improvement and is an issue in the regency/city within that cluster

The dependency information derived from the results of correspondence analysis has been objective. However, due to budget and resource limitations, it cannot capture information from a smaller number of groups. Therefore, cluster analysis is conducted using the principal coordinates from the results of multiple correspondence analysis to obtain information from a fewer number of clusters.

Cluster analysis can be conducted to group regencies/cities with similar characteristics into smaller clusters. In this analysis, the input data used is the Euclidean distance matrix of 27 regencies/cities with the calculation of the Euclidean distance. The results of the Euclidean matrix for the cluster analysis are as follows, based on equation (18).

Table 7. Euclidean Distance Matrix of 27 Regencies/Cities

Category	M_1	M_2	M_3	...	M_{27}
M_1	0	2.4153	2.3920	...	2.8991
M_2	2.4153	0	2.4677	...	3.0822
M_3	2.3920	2.4677	0	...	3.0563
⋮	⋮	⋮	⋮	⋱	⋮
M_{27}	2.8991	3.0822	3.0563	...	0

The clustering process was carried out using five hierarchical cluster analysis methods. The clustering method chosen was the one that most consistently matched the groupings of cities/regencies from the correspondence analysis. Consistency of clustering results was determined by the similarity of city/regency groups between the cluster analysis and the correspondence analysis. The comparison results of this clustering can be seen in Table 8 to determine the best hierarchical method.

Based on Table 8, it can be observed that the number of regencies/cities in each cluster matching the results of the correspondence analysis grouping is highest with the Ward method, amounting to 20. Moreover, in cluster 3, the group members are identical. Therefore, it can be concluded that the Ward method is the most consistent clustering method with the results of the correspondence analysis. Thus, the clustering of regencies/cities in West Java based on the dependency among the categories of the seven variables will use the hierarchical cluster analysis method with Ward's method.

Table 8. Results of Correspondence Analysis and Cluster Analysis

No	Correspondence Analysis	Ward's Method	Single Linkage	Average Linkage	Centroid Method	Complete Linkage
1	Kuningan Pangandaran Banjar City	Kuningan [1]	Kuningan [1]	Kuningan [1]	Kuningan [1]	Kuningan [1]
2	Bogor Ciamis	Bogor [2] Ciamis Bandung	Bogor [2] Ciamis. Bandung, Cirebon, Indramayu, Karawang, Bekasi, Cianjur, Garut, Bekasi City	Bogor [1] Bandung, Cirebon, Indramayu, Karawang, Bekasi	Bogor [2] Ciamis Bandung, Cirebon, Indramayu, Subang, Karawang, Bekasi, Bekasi City, Depok City	Bogor [1] Bandung, Cirebon, Indramayu, Karawang, Bekasi
3	**Cirebon Indramayu Karawang Bekasi**	**Cirebon [4] Indramayu Karawang Bekasi**	Pangandaran [0]	Pangandaran [0]	Pangandaran [0]	Pangandaran [0]
4	Sukabumi Cianjur	Sukabumi [2] Cianjur Garut West Bandung	Sukabumi [1]	Sukabumi [2] Cianjur Garut West Bandung	Sukabumi [1]	Sukabumi [2] Cianjur Garut West Bandung
5	Bandung	Banjar City [0]	Banjar City [0]	Banjar City [0]	Banjar City [0]	Banjar City [0]
6	Garut	Pangandaran [0]	Cirebon City [0]	Cirebon City [0]	Garut [1]	Cirebon [0]
7	Tasikmalaya	Tasikmalaya [1]	Tasikmalaya [1]	Tasikmalaya [1]	Tasikmalaya [1]	Tasikmalaya [1]
8	Majalengka	Majalengka [1]	Majalengka [1]	Majalengka [1]	Majalengka [1]	Majalengka [1]
9	Sumedang	Sumedang [1]	Sumedang [1]	Sumedang [1]	Sumedang [1]	Sumedang [1]
10	Subang Depok City	Subang [1]	Subang [1]	Subang [1]	Cianjur [0]	Subang [1]
11	Purwakarta	Purwakarta [1]	Purwakarta [1]	Purwakarta [1]	Purwakarta [1]	Purwakarta [1]
12	Bandung Barat	Cirebon City [0]	Bandung Barat [1]	Ciamis [0]	Bandung Barat [1]	Ciamis [0]
13	Bogor City Cirebon City	Bogor City [1]	Bogor City [1]	Bogor City [1]	Bogor City [1]	Bogor City [1]
14	Sukabumi City	Sukabumi City [1]	Sukabumi City [1]	Sukabumi City [1]	Sukabumi City [1]	Sukabumi City [1]
15	Bandung City	Bandung City [1]	Bandung City [1]	Bandung City [1]	Bandung City [1]	Bandung City [1]
16	Bekasi City	Bekasi City [1] Depok City	Depok City [1]	Bekasi City [1] Depok City	Cirebon [0]	Bekasi City [1] Depok City
17	Cimahi City	Cimahi City [1]	Cimahi City [1]	Cimahi City [1]	Cimahi City [1]	Cimahi City [1]
18	Tasikmalaya City	Tasikmalaya City [1]	Tasikmalaya City [1]	Tasikmalaya City [1]	Tasikmalaya City [1]	Tasikmalaya City [1]

In this study, hierarchical cluster analysis using Ward's method will be conducted with 3 clusters. 4 clusters. and 5 clusters. The clustering with fewer clusters is aimed at minimizing government expenditures in developing programs to address sanitation and clean water issues in West Java. The results of the clustering into 3 clusters are as follows:

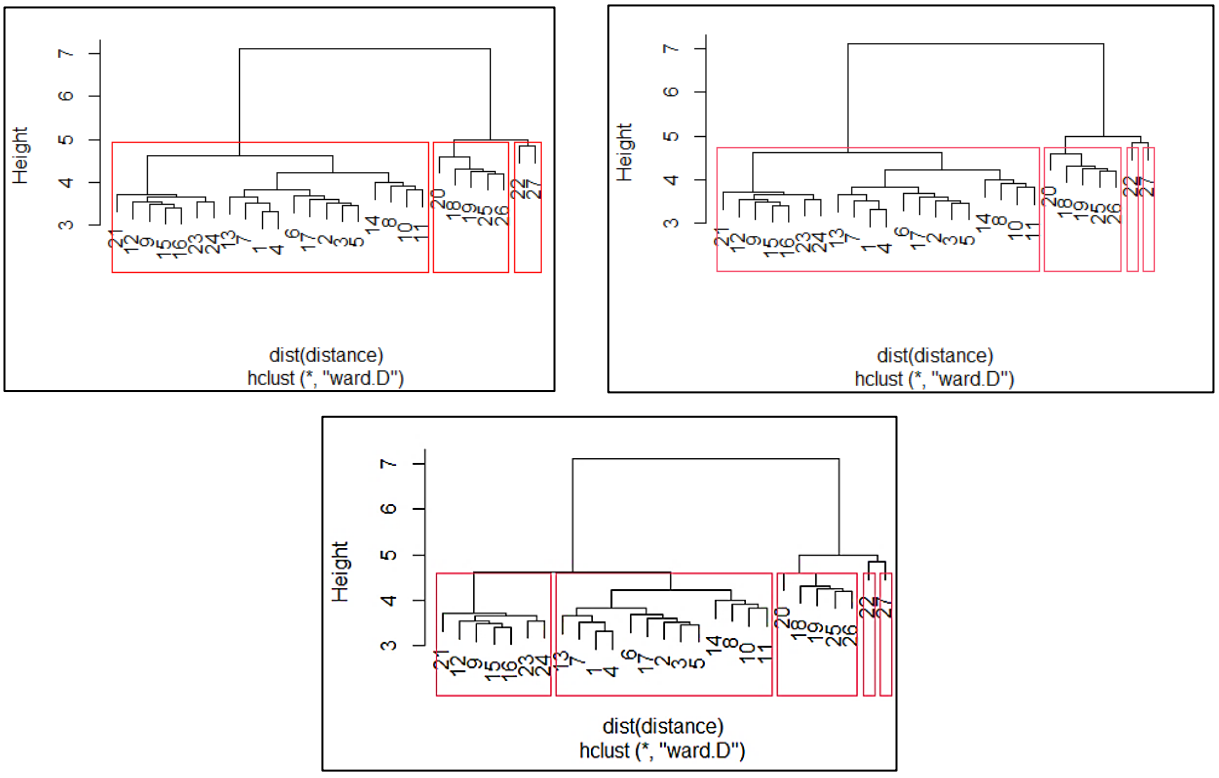

Figure 1. Hierarchical Cluster Analysis Using Ward's Method with 3 clusters, 4 clusters, and 5 clusters

Reducing the number of clusters will decrease government expenditures in addressing clean water and sanitation issues in West Java. The advantage of cluster analysis is that the number of clusters chosen for governmental actions can be adjusted based on the available budget and resources. Next, information regarding the characteristics of the five formed groups will be identified. The characteristics of the five formed clusters will be determined to highlight clearly the issues that need improvement in each group. For clusters 3 and 4, the problems are less apparent, and each group tends to be in a generally good condition. The grouping of cities/regencies can be seen in Figure 1. The members of each cluster, along with their characteristics, can be seen in Table 9 as follows:

Table 9. Characteristics of the 5 Groups of Regencies/Cities

Cluster	City/Regency	Similar Characteristics
1	• Cirebon (M_9), Indramayu (M_{12}), Karawang (M_{15}), Bekasi (M_{16}), Bandung City (M_{21}), Bekasi City (M_{23}), Depok City (M_{24})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The main water source used for cooking/bathing/washing/others is a bore well/pump ($X_{6,4}$). • *Distance to the waste/sewage/feces storage site is < 10 m ($X_{7,2}$) (except in the city of Bandung)
2	• Bogor (M_1), Sukabumi (M_2), Cianjur (M_3), Bandung (M_4), Garut (M_5), Tasikmalaya (M_6), Ciamis (M_7), Kuningan (M_8), Majalengka (M_{10}), Sumedang (M_{11}), Subang (M_{13}), Purwakarta (M_{14}), West Bandung (M_{17})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$).
3	• Pangandaran (M_{18}), Bogor City (M_{19}), Sukabumi City (M_{20}), Cimahi City (M_{25}), Tasikmalaya City (M_{26})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • *The main water source used by the household for drinking is bottled water ($X_{5,2}$) (except in the city of Sukabumi)
4	• Cirebon City (M_{22})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • The main water source used by the household for drinking is piped water ($X_{5,3}$). • The main water source used for cooking/bathing/washing/others is piped water ($X_{6,3}$). • The main drinking water source is not a well/pump/spring ($X_{7,1}$).
5	• Banjar City (M_{27})	• Has a toilet facility, used only by household members ($X_{1,1}$). • The type of toilet used is a gooseneck closet ($X_{2,1}$). • The final disposal site for feces is a septic tank ($X_{3,1}$). • In the past 5 years, the septic tank has never been emptied ($X_{4,3}$). • The main water source used by the household for drinking is bottled water ($X_{5,2}$) • The main water source used for cooking/bathing/washing/others is a protected well ($X_{6,5}$). • Distance to the waste/sewage/feces storage site is ≥ 10 m ($X_{7,3}$).

*The category still requires improvement and is an issue in the regency/city within that cluster

Based on the results in Table 9, there are several issues related to clean water and sanitation that need to be addressed in several regencies/cities in West Java, as follows:

1. In most of Cluster 1, including Cirebon, Indramayu, Karawang, Bekasi, Bekasi City, and Depok City, the distance between residential areas and waste/sewage/feces storage is still less than 10 meters.
2. The current status in this cluster is satisfactory and needs to be maintained.
3. In Banjar City and the majority of Cluster 3 and Cluster 4 areas, including Pangandaran, Bogor City, Tasikmalaya City, and Cimahi City, residents still rely on bottled water as their primary drinking water source $(X_{5,2})$. In contrast, tap water, protected water, bore wells/pumps, or rainwater would be a better source of drinking water.
4. In Cirebon City and Banjar City, there is a characteristic of needing to have emptied septic tanks in the past five years $(X_{4,3})$.

4. CONCLUSION

Multiple correspondence analysis produced a correspondence map in 63 dimensions obtained from principal coordinates. These dimensions involve 100% inertia, making it impractical to use a two-dimensional correspondence map. As a solution, the Euclidean distance method is employed to identify regencies/cities based on their clean water and sanitation conditions. The dependency information from the multiple correspondence analysis results has been objective, but it cannot reveal insights into fewer clusters aligned with budget and resource availability. Therefore, cluster analysis is conducted using the principal coordinates from the multiple correspondence analysis to obtain insights from fewer clusters. If three clusters are formed, then the government only needs to carry out one program to overcome the problem of clean water and sanitation. If four clusters are formed, the government needs to carry out two programs to overcome the issues of clean water and sanitation. If five clusters are formed, the government needs to carry out three programs to overcome the problem of clean water and sanitation. A recommendation for future research is that if the expected outcome is a two-dimensional correspondence map visualization, an alternative method that can be used is Joint Correspondence Analysis (JCA).

For future research, if the desired outcome is a two-dimensional correspondence map visualization, an alternative method to consider is Joint Correspondence Analysis (JCA). However, when the data consists of numerous qualitative variables with several categories within each variable, JCA may yield suboptimal results. In such cases, multiple correspondence analysis with a

categorization approach based on distance matrices is recommended. Additionally, the quality of the correspondence mapping can be assessed through supplementary methods beyond cumulative variance. One approach worth considering is evaluating the s-stress value. A smaller s-stress indicates better mapping quality and results in a more reliable visualization.

The government focuses on evaluating water and sanitation issues in each regency/city based on their specific characteristics. It is advised that the government allocate funds to provide clean drinking water sources, such as piped water, protected water sources, boreholes/pumps, or rainwater harvesting, with priority given to several regencies/cities in West Java. Additionally, it is recommended that education, incentives, and easily accessible septic tank services be provided, along with implementing regulations and monitoring to encourage septic tank maintenance in areas that do not regularly perform this service, with a priority for specific regencies/cities. The government should also prioritize the development of modern sanitation facilities, public education, and enhanced environmental awareness to address the issue of pit latrines as a means of waste disposal, particularly in urban areas of certain regencies/cities. Furthermore, the development of proper sewage management facilities, increased public awareness, enforcement of regulations, and regular monitoring should be undertaken to address the improper disposal of waste into ponds, fields, rivers, lakes, or seas in several areas. The government is advised to relocate waste disposal sites located less than 10 meters from residential areas and provide public education and strict monitoring. Finally, the government can adjust available funding and resources to implement water and sanitation programs, where cluster analysis can help optimize expenses by focusing on fewer groups.

ACKNOWLEDGEMENTS

The authors would like to thank Indah Lesmini, who has helped get data and support. The Padjadjaran University Lecturer Competency Research (RKDU) 2024 Number (1915/UN6.3.1/PT.00/2024) provides the authors with financial support.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Upper Saddle River, 2007.

[2] M. Greenacre, J. Blasius, Multiple Correspondence Analysis and Related Methods, Chapman and Hall/CRC, New York, 2006.

[3] E.J. Beh, Simple Correspondence Analysis: A Bibliographic Review, Int. Stat. Rev. 72 (2004), 257–284. https://doi.org/10.1111/j.1751-5823.2004.tb00236.x.

[4] I. Ginanjar, I. Ade, Nurwahidah, et al. Analysis of Multivariate Associations with Qualitative and Quantitative Variables using Hybrid of Burt Multiple Correspondence Analysis and Cosine Association Matrices (A Case Study: The high school's accreditation in West Java), J. Adv. Res. Dyn. Control Syst. 12 (2020), 826-832.

[5] R.S. Kristanto, I. Ginanjar, T. Purwandari, Recategorization Method Based on Dependence between Qualitative Variables Using Joint Correspondence Analysis with Elliptical Confidence Regions, Commun. Math. Biol. Neurosci. 2024 (2024), 36. https://doi.org/10.28919/cmbn/8444.

[6] P.J. Rosa, D. Morais, P. Gamito, et al. The Immersive Virtual Reality Experience: A Typology of Users Revealed Through Multiple Correspondence Analysis Combined with Cluster Analysis Technique, Cyberpsychol. Behav. Soc. Netw. 19 (2016), 209–216. https://doi.org/10.1089/cyber.2015.0130.

[7] L. Phan, H. Liu, C. Tortora, K-Means Clustering on Multiple Correspondence Analysis Coordinates, Arch. Data Sci. Ser. B. 1 (2019), 1-17. https://doi.org/10.5445/KSP/1000085952/05.

[8] M. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, 1984.

[9] D.H. Kim, G. Lee, New Link of Multiple Correspondence Analysis and K-Means Cluster Analysis, J. Korean Data Inf. Sci. Soc. 33 (2022), 1043–1052. https://doi.org/10.7465/jkdi.2022.33.6.1043.

[10] D. Florensa, J. Mateo-Fornés, F. Solsona, et al. Use of Multiple Correspondence Analysis and K-Means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-Sectional Study, J. Med. Internet Res. 24 (2022), e29056. https://doi.org/10.2196/29056.

[11] R. Kaminskyy, N. Shakhovska, Withdrawal Notice: The Method of Dendrograms Disclosure for Evaluation of Cluster Analysis Results in IoT Domain, Int. J. Sens. Wirel. Commun. Control 10 (2020), 11. https://doi.org/10.2174/2210327910999200821161039.

[12] A. Kaplan, J. Bien, Interactive Exploration of Large Dendrograms with Prototypes, Amer. Statistician 77 (2023), 201–211. https://doi.org/10.1080/00031305.2022.2087734.

[13] J. Dreyer, J.M. Bergmann, K. Koehler, et al. Using Multicorrespondence Analyses and Cluster Analysis to Construct Types of Care Arrangements, Innov. Aging 6 (2022), 175–176. https://doi.org/10.1093/geroni/igac059.703.

[14] S. Kim, R. Sarkar, S. Kumar, et al. Understanding COVID-19 Vaccine Hesitancy in Meghalaya, India: Multiple Correspondence and Agglomerative Hierarchical Cluster Analyses, PLOS Glob. Public Health 4 (2024), e0002250. https://doi.org/10.1371/journal.pgph.0002250.

[15] F. Brelle, How Do Irrigation and Drainage Interventions Secure Food Production and Livelihood for Rural Communities?, Irrig. Drain. 65 (2016), 210–213. https://doi.org/10.1002/ird.1970.

[16] O. Gulseven, How to Achieve Sustainable Development Goals by 2030?, SSRN (2020). https://doi.org/10.2139/ssrn.3592921.

[17] E. Crawford, Achieving Sustainable Development Goals 5 and 6: The Case for Gender-Transformative Water Programmes, Oxfarm International, 2020.

[18] T. Afifah, M.T. Nuryetty, Cahyorini, et al. Subnational Regional Inequality in Access to Improved Drinking Water and Sanitation in Indonesia: Results from the 2015 Indonesian National Socioeconomic Survey (SUSENAS), Glob. Health Act. 11 (2018), 31–40. https://doi.org/10.1080/16549716.2018.1496972.

[19] A. Ortigara, M. Kay, S. Uhlenbrook, A Review of the SDG 6 Synthesis Report 2018 from an Education, Training, and Research Perspective, Water 10 (2018), 1353. https://doi.org/10.3390/w10101353.

[20] I. Ginanjar, I. Nurhuda, N. Sunengsih, Sudartianto, Contribution of a Categorical Statistical Test in Examining Dependencies among Qualitative Variables by Means Simplification of Correspondence Analysis, J. Phys.: Conf. Ser. 1265 (2019), 012022. https://doi.org/10.1088/1742-6596/1265/1/012022.

[21] T. Purwandari, I. Ginanjar, D.D. Dewi, Multiple Correspondence Analysis for Identifying the Contribution of Infant Mortality Indicator Categories, J. Phys.: Conf. Ser. 1776 (2021), 012064. https://doi.org/10.1088/1742-6596/1776/1/012064.

[22] G. Li, S. Lu, H. Zhang, S. Lo, Correspondence Analysis on Exploring the Association between Fire Causes and Influence Factors, Procedia Eng. 62 (2013), 581–591. https://doi.org/10.1016/j.proeng.2013.08.103.

[23] E.J. Beh, Elliptical Confidence Regions for Simple Correspondence Analysis, J. Stat. Plan. Inference 140 (2010), 2582–2588. https://doi.org/10.1016/j.jspi.2010.03.018.

[24] A.C. Rencher, Method of Multivariate Analysis, Wiley, 2002.

[25] N. Maulida, S.P. Wulandari, Analisis Cluster dan Korespondensi terhadap Indikator Pertumbuhan Penduduk Kota Surabaya Tahun 2020, J. Sains Seni ITS, 11 (2022), D43-D49.