Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:18

https://doi.org/10.28919/cmbn/9034

ISSN: 2052-2541

CHARACTERIZATION METHOD FOR MORE THAN THREE DIMENSIONS BASED ON DEPENDENCE AND CORRELATION USING HYBRID MULTIPLE CORRESPONDENCE ANALYSIS

DHANTI AURILIA PRATIWI, IRLANDIA GINANJAR*, TITI PURWANDARI,

ANINDYA APRILIYANTI PRAVITASARI, GUMGUM DARMAWAN

Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia

Abstract: Characteristics of a district refer to specific attributes that describe its condition and quality. Identifying these characteristics provides more precise insights into districts that require more attention in specific situations. These characteristics are determined based on the interdependence of categories. Dependencies among multiple qualitative and quantitative variables can be analyzed using multiple correspondence analysis (MCA) hybrid with cosine correlation. MCA is particularly suited for analyzing contingency tables with more than two qualitative variables. This study aims to identify the characteristics of objects based on the categories of qualitative variables and the characteristics of objects based on quantitative variables. Several qualitative variables with many categories may result in less representative information, potentially failing to reach 80% cumulative variance in two dimensions. Therefore, the novelty of this study lies in identifying characteristics using Euclidean distance with a variance percentage of 100% in more than three dimensions and used two types of variables. The data used in this study are eight qualitative variables with 49 categories and three quantitative variables from the Supporting Area Survey of Bandung Regency in 2023. The analysis results indicate that the cumulative variance in two dimensions is only 16.9%. Consequently, calculations were performed using an Euclidean distance matrix with 43 dimensions to achieve 100% cumulative variance. The results of the analysis revealed that 28 district groups were identified.

Keywords: multiple correspondence analysis; cosine correlation; Euclidean distance; infrastructure.

2020 AMS Subject Classification: 62H25

*Corresponding author

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

Received November 24, 2024

1. INTRODUCTION

Correspondence analysis is a multivariate analysis method that explores interdependence to analyze objects by depicting patterns of dependency in frequency tables. The dependencies studied are those among qualitative variables not based on causality [1]. Simultaneous data presentation techniques in correspondence analysis simplify data aspects by visually representing data in graphs for easier interpretation [2]. These graphs are often referred to as correspondence maps. Generally, correspondence analysis utilizes contingency tables, which are cross-tabulations of two categorical variables, transforming qualitative data into levels and reducing dimensions for perceptual mapping [3].

Correspondence analysis is classified into Simple Correspondence Analysis (SCA) and Multiple Correspondence Analysis (MCA). SCA is applied to data with two categorical variables [4], whereas MCA is applied to data with more than two categorical variables [1]. Multiple correspondence analysis (MCA) is a method used to analyze contingency tables with more than two qualitative variables [5]. The dependency patterns studied are those between qualitative variables not based on causality [1]. The main objective of correspondence analysis is to estimate the primary coordinates of rows and columns from data with qualitative variables, thus mapping the dependency between qualitative variables [6].

MCA can be performed using several methods, including correspondence analysis based on an indicator matrix and correspondence analysis based on the Burt matrix [7]. The use of correspondence analysis based on the Burt matrix is considered better than the indicator matrix because the elements of the indicator matrix are binary. They only take values of 0 or 1, where the value 0 indicates absence and the value 1 indicates presence [8], so the indicator matrix contains many zeros (sparse matrix) [9]. Additionally, using the Burt matrix is preferable because the mapping obtained in the Burt matrix is based on standard residuals that represent the dependency values between categories. The Burt matrix also results in a smaller primary coordinate scale than the indicator matrix. Therefore, the percentage of inertia in the Burt matrix will be higher compared to using the indicator matrix [10]. MCA will be used in this study because the qualitative variables involved exceed two and encompass many categories. The large number of variables and categories may result in the percentage of variance not reaching the acceptable threshold of 80% in two dimensions [11][12], making it challenging to create a two-dimensional map [13].

In their study, Ginanjar et al. [14] also used the hybrid method of MCA and the cosine association approach. However, that study uses Mahalanobis distance, whereas this study uses Euclidean

distance because the principal coordinate points obtained from MCA already account for different variances. In addition, Kristanto et al. [15] also used a correspondence analysis. However, the correspondence method applied is joint correspondence analysis, not MCA, which produces a two-dimensional map with cumulative variance > 70% in two dimensions. Pinem et al. [16] also employ the MCA method in their study, but they add a step to minimize the number of groups formed by reducing categories. However, even after category reduction, that study still could not produce a two-dimensional map with cumulative variance ≥ 80%. In contrast, this study does not perform category reduction for fear that it would reduce data diversity, resulting in less representative information.

In the study related to infrastructure, Sukarno, D. [17] shows that his study, using qualitative study methods, indicates that currently, the districts in Bandung Regency do not have a planned strategy needed to improve the standards of organizational factors, one of which is infrastructure. Furthermore, the government still needs to understand a more comprehensive approach to what infrastructure is more beneficial and necessary. In addition, Syaiful et al. [18] determine the priority of infrastructure development in the coastal areas of North Sangatta and South Sangatta using the importance-performance analysis technique to produce three priority groups. In contrast, this study will provide information about the infrastructure needed by the districts. It will generate groups of districts without limiting the number of groups, allowing for infrastructure improvements according to their specific needs.

In this study, we utilized qualitative and quantitative variables. One advanced method derived from MCA is hybrid multiple correspondence analysis with cosine correlation, which can accommodate quantitative variables [14]. MCA will produce principal coordinates in a quantitative form [19]. Therefore, the correlation between these two quantitative variables, the principal coordinates from the MCA results and the quantitative variables will be examined. This study aims to identify the characteristics of objects based on the categories of qualitative variables and the characteristics of objects based on quantitative variables. Identifying characteristics will result in groups of objects based on similar characteristics. The novelty of this study lies in the stage of identifying characteristics using Euclidean distance with a variance percentage of 100% in more than three dimensions. In addition, this study also uses two types of variables: qualitative and quantitative variables.

The study of hybrid MCA with cosine correlation will be conducted to group districts in Bandung Regency based on infrastructure condition variables. The data used is secondary data from the

Supporting Area Survey of Bandung Regency in 2023. The data has different measurement scales, including qualitative data with a nominal scale of multiple categories and quantitative data with a ratio scale. The study data consists of eight qualitative variables with 49 categories and three quantitative variables. Each district has different infrastructure condition characteristics, necessitating different handling in each district based on these characteristics. By comparing the infrastructure conditions in each district, there must be a dependency between categories in each qualitative variable of infrastructure conditions and the districts. Suppose there is a dependency between categories in each qualitative variable and the districts and a correlation between quantitative variables. In that case, the districts will be grouped into the same category. This enables the government to create infrastructure-related programs that are accurately targeted according to the varying characteristics of each district.

2. MATERIAL AND METHOD

A. Data Sources

This study aims to assess the infrastructure characteristics in Bandung Regency, which comprises 31 districts and 280 sub-districts. The data used in this study is secondary data, specifically the Supporting Area Survey of Bandung Regency for the year 2023. The data is qualitative with a nominal measurement scale and quantitative with a ratio measurement scale. This study uses eight qualitative variables with 18 categories and 31 districts, along with three quantitative variables. The qualitative variables used in this study are Village Medium Term Development Plan in Effect (X_1), Cell Phone/phone signal in most Sub-District Areas (X_2), The Presence of Lighting on the Main Roads of The Sub-Districts (X_3), Existence of LPG Bases/Agents/Sellers (X_4), Final Disposal Place for Most Families' Feces (X_5), Place/Channel for Disposal of Liquid Waste from Most Families' Bathing/Washing Water (X_6), and The source of Water for Bathing/Washing in Most Families Comes From (X_7). Furthermore, district data are used as row categories. The quantitative variables used consist of the Construction of Long Village Roads During 2022 (X_8), the number of families using PLN electricity (X_9), and the Number of Livable Houses (X_{10}). Tables 1 and 2 contain the data used in this study, but the complete data can be viewed on the web page: https://bit.ly/Data_Journal.

Table 1. The Qualitative Data

Number	Name of Village/Sub-district	Name of District	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Arjasari Village	Arjasari District	There is RPJMD	Very Strong Signal	Yes, most of them	There is an Existence of LPG Bases/Agents/Sellers	Earth pit	In the pit/ Open land	Well
2	Lebakwangi Village	Arjasari District	There is RPJMD	Very Strong Signal	Yes, most of them	There is an Existence of LPG Bases/Agents/Sellers	Earth pit	Infiltration pit	Well
3	Batukarut Village	Arjasari District	There is RPJMD	Very Strong Signal	Yes, most of them	There is an Existence of LPG Bases/Agents/Sellers	Earth pit	Drainage (Ditch/ Gutter)	Well
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
280	Sekarwangi Village	Soreang District	There is RPJMD	Strong Signal	Yes, most of them	There is an Existence of LPG Bases/Agents/Sellers	Tank/ wastewater management installation	Drainage (Ditch/ Gutter)	Well

Table 2. The Quantitative Data

Number	Name of Village/Sub-district	Name of District	X_8 (m)	X_9 (family)	X_{10} (house)
1	Arjasari Village	Arjasari District	468	3.765	3.001
2	Lebakwangi Village	Arjasari District	0	4.126	8.517
3	Batukarut Village	Arjasari District	1.000	5.282	4.775
⋮	⋮	⋮	⋮	⋮	⋮
280	Sekarwangi Village	Soreang District	989	3.578	3.478

The seven qualitative variables were transformed into a contingency table with districts as rows and each category as columns. This resulted in 7 contingency tables that can be used in the chi-square test.

B. Contingency Table

The study produced a two-way contingency table, with districts as a row and characteristics as a column. q_1 is the number of categories for the row variable (district) with $j = 1,2,\dots,q_1$, $q_{\tilde{k}}$ is the number of categories for the column variable (characteristics) with $\tilde{j} = 1,2,\dots,q_{\tilde{k}}$, $\tilde{k} = 2,3,\dots,v$, individual in the data is n, and $n_{j\tilde{j}}$ is the number of observations (subdistrict). The contingency table that will be formed is as follows.

Table 3. Contingency Table

Districts	Characteristics Variable (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

According to Table 3, q_1 is the number of categories for variable 1, which is the row variable (village/sub-district), and $q_{\tilde{k}}$ is the number of categories for the column variable (infrastructure condition characteristics) with $\tilde{k} = 2,3,\dots,v$. n is the total number of observations, and $n_{j\tilde{j}}$ is the total number of observations for row category j^{th} and column category j^{th}. Based on Table 1, the cross-tabulation matrix $\mathbf{N}$ can be calculated using the following equation:

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

Based on the cross-tabulation matrix $\mathbf{N}$ on equation (1), can be obtained correspondence matrix $\tilde{\mathbf{P}}$ as follows:

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

In equation (2), $\tilde{p}_{j\tilde{j}}$ is the joint probability estimator districts and characteristics variable. $\tilde{p}_{j\bullet} = \frac{n_{j\bullet}}{n}$ is the marginal probability estimator of characteristics (column), and $\tilde{p}_{\bullet\tilde{j}} = \frac{n_{\bullet\tilde{j}}}{n}$ is the marginal probability estimator of districts (row) [20].

C. Chi-Square Test

The qualitative data used in correspondence analysis must show interdependence among its variables. In this study, the characteristics must be dependent on the district. The Chi-Square test is one hypothesis test that can assess the dependence between categorical variables in a contingency table [21]. The hypothesis for the Chi-Square test of the two qualitative variables is as follows [20].

$H_0: \pi_{j\tilde{j}} = \pi_{j.}.\pi_{.\tilde{j}}$; (there is no dependency between the two variables)

$H_1: \pi_{j\tilde{j}} \neq \pi_{j.}.\pi_{.\tilde{j}}$; (there is a dependency between the two variables)

The test statistics for the independence analysis using the Pearson Chi-square test are as follows.

$$\chi^2 = n\sum_{j=1}^{q_1}\sum_{\tilde{j}=1}^{q_{\tilde{k}}}\frac{(\tilde{p}_{j\tilde{j}}-\tilde{p}_{j\bullet}\tilde{p}_{\bullet\tilde{j}})^2}{\tilde{p}_{j\bullet}\tilde{p}_{\bullet\tilde{j}}} \quad (3)$$

Based on equation (3), n is the number of observations, $\tilde{p}_{j\bullet}$ is the marginal probability of the j^{th} district, $\tilde{p}_{\bullet\tilde{j}}$ is the marginal probability of $\tilde{j}^{th}$ characteristics, $\tilde{p}_{j\tilde{j}}$ joint probability of the j and $\tilde{j}$, q_1 is the number of categories on j^{th}, and qk is the number of categories on $\tilde{j}^{th}$. The Pearson Chi-square test criterion with $\chi^2 \sim \chi^2_v$, $v = (q_1 - 1)(q_{\tilde{k}} - 1)$ is to reject H_0 if the *p-value* $< \alpha$, accept in other cases. The rejection of H_0 is identified based on the p-value written in equation (4) as follows:

$$p - value = P\{\chi^2_v > \chi^2\} \quad (4)$$

D. Multiple Correspondence Analysis

Multiple Correspondence Analysis is used to identify relationships and patterns among more than two qualitative variables with categorical characteristics [5]. Multiple Correspondence Analysis is a method used to analyze contingency tables with more than two qualitative variables [9]. The correspondence matrix formed in multiple correspondence analysis through the indicator matrix is subsequently structured using the Burt matrix [5]. The indicator matrix in multiple correspondence analysis can be presented using a raw data table where the columns represent variables with categorical cells, and the rows represent respondents [22]. In an indicator matrix, the value 0 represents that an object doesn't belong to a specific category, while the value 1 represents that an object belongs to that category [10]. Let us consider a data matrix of size $n \times v$ with the notation $\mathbf{Y} = (y_{ik})$, where n denotes the number of objects, n denotes the number of qualitative variables, $i = 1,2,\dots,n$ and $k = 1,2,\dots,v$. If q_k represents the number of categories for the k^{th} variable and $\mathbf{Z}_\text{k} = (z_{ijk})$ represents the indicator matrix for the k^{th} variable, with elements of size $n \times q_k$, where z_{ijk} is the $(i,j)^{th}$ element of Z_k and $j = 1,2,\dots,q_k$. In this context, the indicator matrix can be represented as the following equation [1]:

$$\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \cdots \quad \mathbf{Z}_\text{v}] \quad (5)$$

The indicator matrix is of size $n \times Q$ with $Q = \sum_{k=1}^{v} q_k$, and each variable k has q_k categories. The indicator matrix cannot reflect dependencies, whereas the aim of this study is to

identify dependencies. Therefore, the next step is to form the Burt matrix. The Burt matrix is a multi-way contingency table that results from the cross-tabulation of the combined indicator matrices of variables for each of their categories [9]. The combined indicator matrix can be defined as $\mathbf{Z}$, as stated in equation (5). There are v qualitative variables, and $\mathbf{Z}_k$ represents the indicator matrix for each category with $k = 1,2, \dots, v$. In correspondence analysis, the Burt matrix forms principal coordinates whose scale or dimension is reduced compared to the indicator matrix.

The Burt matrix can be denoted as $\mathbf{B}$ and is related to the combined indicator matrix by the following formula in equation (6) and (7) [10]:

$$\mathbf{B} = \mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_{\tilde{k}} & \cdots & \mathbf{Z}_1^T\mathbf{Z}_v \\ \mathbf{Z}_2^T\mathbf{Z}_1 & \mathbf{Z}_1^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_2^T\mathbf{Z}_{\tilde{k}} & \cdots & \mathbf{Z}_2^T\mathbf{Z}_v \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{Z}_k^T\mathbf{Z}_1 & \mathbf{Z}_k^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_k^T\mathbf{Z}_{\tilde{k}} & \cdots & \mathbf{Z}_k^T\mathbf{Z}_v \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{Z}_v^T\mathbf{Z}_1 & \mathbf{Z}_v^T\mathbf{Z}_2 & \cdots & \mathbf{Z}_v^T\mathbf{Z}_{\tilde{k}} & \cdots & \mathbf{Z}_v^T\mathbf{Z}_v \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{N}_{12} & \cdots & \mathbf{N}_{1\tilde{k}} & \cdots & \mathbf{N}_{1v} \\ \mathbf{N}_{21} & \mathbf{D}_2 & \cdots & \mathbf{N}_{2\tilde{k}} & \cdots & \mathbf{N}_{2v} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{N}_{k1} & \mathbf{N}_{k2} & \cdots & \mathbf{D}_k & \cdots & \mathbf{N}_{kv} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{N}_{v1} & \mathbf{N}_{v2} & \cdots & \mathbf{N}_{v\tilde{k}} & \cdots & \mathbf{D}_v \end{bmatrix} = (b_{m\tilde{m}}) \quad (6)$$

Where

$$\mathbf{N}_{k\tilde{k}} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1q_{\tilde{k}}} \\ n_{21} & n_{22} & \cdots & n_{2q_{\tilde{k}}} \\ \vdots & \vdots & \ddots & \vdots \\ n_{q_k1} & n_{q_k2} & \cdots & n_{q_kq_{\tilde{k}}} \end{bmatrix} \quad (7)$$

The diagonal matrix $\boldsymbol{D}_k = diag(d_{jk})$, where d_{jk} in equation (6) is the marginal frequency of the j^{th} category of the k^{th} variable for $j = 1,2, \dots, q_k$, $k = 1,2, \dots, v$, and $\tilde{k} = 2,3, \dots, v$. $\mathbf{N}_{k\tilde{k}}$ in equation (7) is the cross-tabulation matrix between the k^{th} and $\tilde{k}^{\text{th}}$ variables. Meanwhile, $b_{m\tilde{m}}$ is an element of the Burt matrix, where $m, \tilde{m} = 1,2, \dots, Q$, so that $j \in m$. $\mathbf{Z}_v^T\mathbf{Z}_{\tilde{v}}$ is the diagonal matrix of the total frequencies for each $\mathbf{Z}_v$ in equation (5).

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 [10]:

$$\mathbf{P} = \frac{1}{g}\mathbf{B} = (p_{m\tilde{m}}) \quad (8)$$

Where $g = \sum_{m=1}^{Q}\sum_{\tilde{m}=1}^{Q} b_{m\tilde{m}}$ and $p_{m\tilde{m}}$ in equation (8) is the element of the Burt Correspondence Matrix for the m^{th} row and the $\tilde{m}^{\text{th}}$ 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 $(\mathbf{c})$ and the row $(\mathbf{r})$ mass of the Burt matrix have equivalent values, expressed through the following formula [10]:

$$\mathbf{c} = \mathbf{r} = \frac{1}{g}\mathbf{B1} \tag{9}$$

Where 1 in equation (9) is a vector of dimension $Q \times 1$ with each element equal to 1. The main diagonal elements in the Burt Correspondence Matrix reflect the marginal probabilities, while the upper and lower triangular regions reflect the joint probabilities. The calculation of the standard residual matrix can be described as follows [10]:

$$\mathbf{S} = \mathbf{D_c}^{-\frac{1}{2}}(\mathbf{P} - \mathbf{cc}^{\mathrm{T}})\mathbf{D_c}^{-\frac{1}{2}} = (s_{m\tilde{m}}) \tag{10}$$

$\mathbf{P}$ is the Burt correspondence matrix from equation (8), and it is the row diagonal matrix. $\mathbf{D_c}$ is the column diagonal matrix, and $\mathbf{c}$ from equation (9) is the row or column mass of Q diagonal matrix. The Burt matrix has equal row and column mass proportions, consisting of $\mathbf{D_r} = \mathbf{D_c} = \mathrm{diag}(\mathbf{c})$, $\mathbf{P1} = \mathbf{c}$ and $\mathbf{1}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}$. $s_{m\tilde{m}}$ being the element of the standardized residual matrix for row m^{th} and column $\tilde{m}^{th}$.

Decomposition of the diagonal matrix of standardized residual matrix $\mathbf{S}$ is performed to obtain the principal coordinate vectors of each orthogonal dimension. Eigenvalue decomposition (EVD) is the decomposition method used in this study because the Burt matrix and the standardized residual matrix are symmetric and semi-positive definite matrices [14]. Eigen Value Decomposition (EVD) can be defined in the matrix $\mathbf{S}$ in equation (11):

$$\mathbf{S} = \mathbf{E\Lambda E}^{\mathrm{T}} \tag{11}$$

$\mathbf{E}$ in equation (11) is an orthogonal matrix (such that $\mathbf{E}^{-1} = \mathbf{E}^{\mathrm{T}}$), $\mathbf{EE}^{\mathrm{T}} = \mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$, where each column of $\mathbf{E}$ represents eigenvectors that are orthogonal to each other. The columns of matrix $\mathbf{E}$ are denoted as $\mathbf{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_L)$. $\mathbf{\Lambda}$ in equation (11) is a diagonal matrix of eigenvalues λ_ℓ, denoted as $\mathbf{\Lambda} = \mathrm{diag}\,(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_\ell, \dots, \lambda_L), \lambda_\ell =$ eigenvalues ℓ^{th} and $\lambda_1 > \lambda_2 > \cdots > \lambda_\ell > \cdots > \lambda_L, l = 1,2, \dots, L$.

The first step in generating a correspondence map is to obtain the principal coordinates for each category, which illustrates the relationship 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 symmetrical. Thus, the formula for the standard coordinates of both the row and column can be seen as follows:

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{c}}^{-\frac{1}{2}}\mathbf{E} = (h_{m\ell}) \tag{12}$$

Based on equation (12), ℓ represents dimension, and m represents a category. Each row represents standard coordinates for each dimension. Standard coordinates $\mathbf{H}$ alone are still

inadequate for use. This occurs because in standard coordinates $\mathbf{H}$, the results obtained may not adequately represent the dependencies in correspondence mapping. Each coordinate also has different variances, hence the need to use principal coordinates represented by eigenvalues. Principal coordinates are linear combinations of eigenvectors derived from the dependence values between row categories and column categories, and the principal coordinates of each category can construct a correspondence map [23]. The formula for principal coordinates is as follows:

$$\mathbf{F} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{E}\boldsymbol{\Lambda}^{\frac{1}{2}} = (f_{m\ell}) \tag{13}$$

$$\tilde{\mathbf{F}} = \left(f_{j\ell}\right) \tag{14}$$

where $\mathbf{D_c}$ is the diagonal matrix of the column in matrix $\mathbf{F}$, which represents a category. The column matrix $\mathbf{F}$ in equation (13) represents the coordinates for each dimension, while $\tilde{\mathbf{F}}$ in equation (14) is the main coordinates of the subdistrict and the ℓ^{th} dimensions.

Total inertia is the percentage of variability used to measure mapping quality. The proportion of inertia for each dimension and the percentage of missing categories or the percentage of lost information can be determined through total inertia calculation, as total inertia can describe mapping quality [24]. Total inertia is the sum of the squared distances of each coordinate point (row/column) from its centre. Total inertia is obtained based on the following equation [10]:

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

Based on equation (15), the inertia value indicates the contribution of the k^{th} row to the total inertia. Once the inertia value is obtained, the variance can be determined. The variance explained by the δ^{th} dimension depends on the percentage contribution ϕ_δ from each eigenvalue. Variance coverage for each dimension, as well as inertia for D dimensions where $D \leq L$ or cumulative variability percentage, can be defined as follows:

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

$$\tau_D = \frac{\sum_{\delta=1}^{D}\lambda_\delta}{\sum_{\ell=1}^{L}\lambda_\ell} \tag{17}$$

Where ϕ_δ in equation (16) represents the variance coverage of each dimension $\delta = 1,2,\ldots, L$. τ_D denotes the percentage of diversity or variance coverage of the D dimensions. λ_δ in equation (17) signifies the eigenvalue obtained from Eigen Value Decomposition (EVD) for dimension δ^{th}, while λ_ℓ in equation (17) denotes the eigenvalue for dimension ℓ^{th}. The representation of dependence generated from principal coordinates is based on the distances between categories.

Information is extracted based on distance measures to represent the obtained dependencies objectively.

E. Preprocessing Quantitative Variable Data

Preparing data with quantitative variable types is the second step after preparing qualitative variable data, which must be carried out in multiple correspondence analysis. The data is quantitative with a ratio measurement scale. In this study, there are three quantitative variables for which the average will be calculated for each sub-district and data standardization will be performed.

The first step is to find the total for each sub-district. Let y_u be the u^{th} quantitative variable with $u = 1,2, \dots, w$, where w is the number of quantitative variables. If the number of villages for j^{th} sub-district in variable u is N_j, then the matrix for the u^{th} quantitative variable in j^{th} sub-district can be written in equation (18):

$$\mathbf{Y}_j = \begin{bmatrix} y_{11j} & y_{12j} & \cdots & y_{1u} & \cdots & y_{1wj} \\ y_{21j} & y_{22j} & \cdots & y_{2u} & \cdots & y_{2wj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{N_j1j} & y_{N_j2j} & \cdots & y_{wu} & \cdots & y_{N_jwj} \end{bmatrix} \tag{18}$$

The total for each sub-district is used because the units of each quantitative variable are summed without resulting in a new unit. The total of the quantitative variable for j^{th} sub-district in variable u can be written in the equation (19):

$$\tau_j = \left[\sum_{i=1}^{Nj} y_{i1j} \ \sum_{i=1}^{Nj} y_{i2j} \dots \sum_{i=1}^{Nj} y_{ij}\right] = \left[\tau_{1j} \ \tau_{2j} \dots \ \tau_{wj}\right] \tag{19}$$

Next is data standardization, which is used when there are differences in measurement units among variables in a study. Standardization is typically performed by converting each variable into a standard value by subtracting the mean and dividing by the standard deviation.

The data consists of n objects, comprising 31 sub-districts in Bandung Regency, with y variables represented by three quantitative variables: the length of rural road construction during 2022, the number of households using PLN electricity, and the number of livable houses, presented in an initial data matrix $\mathbf{Y}$ of size $n \times v$. The formula for standardization for each variable can be written in the following equation:

$$\bar{\tau}_u = \frac{\sum_{j=1}^{q_1} \tau_{ju}}{q_1} \qquad \sigma_{uu} = \frac{\sum_{j=1}^{q_1} (\tau_{ju} - \bar{\tau}_u)^2}{q_1} \qquad \sigma_u = \sqrt{\sigma_{uu}}$$

$$x_{ju} = \frac{\tau_{ju} - \bar{\tau}_u}{\sigma_u}, j = 1,2, \dots, q_1 ; u = 1,2, \dots, w \tag{20}$$

With x_{ju} in equation (20) being the data after standardization for row j^{th} and column u^{th}. τ_{ju} is the total data for row j^{th} and column u^{th}.$\bar{\tau}_u$ is the average of the total data for column u^{th}, and σ_u is the standard deviation of the total data for column u^{th}.

F. Cosine Correlation

Cosine correlation is a correlation whose similar measurement method calculates the angle between two non-zero vectors [25]. The cosine of the angle between two vectors is an element of the principal coordinate matrix in multiple correspondence analysis, which can be approximated by the correlation between the two vectors. The principal coordinate matrix is calculated using the correlation between qualitative variables and the coordinates of the object variables. The correlation value between the principal coordinates of objects and the quantitative variable data can be written in equation (21) [14]:

$$\rho_{\ell u} = Cos(\theta_{\ell u}) = \frac{\sum_{j=1}^{q1}(f_{j\ell})\,(x_{ju})}{\sqrt{\sum_{j=1}^{q1} f_{j\ell}^2}\sqrt{\sum_{j=1}^{q1} x_{ju}^2}} \tag{21}$$

Where $\theta_{\ell u}$ represents the angle between vector ℓ and vector j, the term $\rho_{\ell u}$ in equation (21) denotes the correlation between the principal coordinate values $f_{j\ell}$ from equation (14) and the object characteristics x_{ju} from equation (20). $f_{j\ell}$ refers to the principal coordinates here, while x_{ju} represents the data after standardization. Based on equation (21), the form of the cosine correlation matrix equation and the coordinate vector of quantitative characteristics can be defined as follows:

$$\mathbf{R} = (\rho_{\ell u}) \tag{22}$$

$$\mathbf{\Psi} = \mathbf{\Lambda}^{\gamma}\mathbf{R} \quad ;\ 0 < \gamma < 1 \tag{23}$$

$$\mathbf{\Psi}^{\mathrm{T}} = \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1u} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2u} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{\ell 1} & \psi_{\ell 2} & \cdots & \psi_{\ell u} \end{pmatrix} \tag{24}$$

$\mathbf{R}$ in equation (22) represents the cosine correlation matrix between $f_{j\ell}$ from equation (14) and x_{ju} from equation (20). The matrix $\mathbf{\Psi}$ in equation (23) will be transposed into the matrix $\mathbf{\Psi}^{\mathrm{T}}$ in equation (24) denotes the coordinate matrix of quantitative characteristic vectors, while $\mathbf{\Lambda}$ in equation (23) is the diagonal matrix of eigenvalues. $\rho_{\ell u}$ in equation (22) indicates the correlation between the principal coordinate values $f_{j\ell}$ and the object characteristics x_{ju}. Additionally, $\psi_{\ell u}$ in equation (24) represents an element of the coordinate matrix of quantitative characteristic

vectors. The value of γ in equation (23) is in the range $0 < \gamma < 1$, with extreme values of $\gamma = 0$ and $\gamma = 1$. Determining the value of γ is crucial for considering the visualization of a specific dimensional space on the map, where γ functions as a weighting factor to adjust the size of vectors in the map dimension [26].

If the calculation of cumulative diversity percentage results in more than three dimensions, then cosine correlation matrix distance calculation can be used. The cosine value of the angle between two vectors depicts the correlation between them; a narrower angle indicates a higher correlation [14]. If there are two vectors $\boldsymbol{f}$ and $\boldsymbol{\psi}$ sized $\ell \times 1$, the cosine of the angle between them can be calculated using the following equation:

$$\tilde{\rho}_{ju} = Cos(\theta_{\mathrm{ju}}) = \frac{\sum_{\ell=1}^{L} f_{j\ell}\, \psi_{\ell u}}{\sqrt{\sum_{\ell=1}^{L} f_{j\ell}^{2}}\sqrt{\sum_{\ell=1}^{L} \psi_{\ell u}{}^{2}}} \tag{25}$$

$$\mathbf{K} = (\tilde{\rho}_{\mathrm{ju}}) \tag{26}$$

Where θ_{ju} represents the angle between vector j and vector u, while $\tilde{\rho}_{\mathrm{ju}}$ in equation (25) denotes the correlation between the principal coordinate values $f_{j\ell}$ from equation (14) and the matrix of the quantitative characteristic vector coordinates $\psi_{\ell u}$ from equation (24). Here, $f_{j\ell}$ refers to the principal coordinate values, and $\psi_{\ell u}$ represents the elements obtained from the transpose of the quantitative characteristic vector coordinates. The matrix $\mathbf{K}$ in equation (26) is the cosine correlation matrix between $f_{j\ell}$ and $\psi_{\ell u}$.

G. Euclidean Distance

The next step is calculating Euclidean distance to determine the characteristics of the object variables with column categories because the main vector coordinate points of each dimension used are perpendicular or orthogonal to each other, and the main coordinate points obtained from the multiple correspondence analysis have considered different variances [27]. 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 into the Euclidean distance as matrix $\mathbf{D}$ can be expressed with the following formula [8]:

$$d(\boldsymbol{f}_m, \boldsymbol{f}_{\tilde{m}}) = \sqrt{(\boldsymbol{f}_m - \boldsymbol{f}_{\tilde{m}})^T (\boldsymbol{f}_m - \boldsymbol{f}_{\tilde{m}})} \tag{27}$$

Where $d(\boldsymbol{f}_m, \boldsymbol{f}_{\tilde{m}})$ is the Euclidean distance between vector $\boldsymbol{f}_m$ and vector $\boldsymbol{f}_{\tilde{m}}$, with $\boldsymbol{f}_m$ being the vector for category m and $\boldsymbol{f}_{\tilde{m}}$ being the vector for category $\tilde{m}$. If the value of $d(\boldsymbol{f}_m, \boldsymbol{f}_{\tilde{m}})$ from equation (27) is smaller, the two matched vectors will be more similar. Conversely, if the

value of $d(\boldsymbol{f}_m, \boldsymbol{f}_{\tilde{m}})$ from equation (27) is larger, the two matched vectors will be more different [28][29].

3. MAIN RESULTS

The data is analyzed using RStudio version 2022.07.2+576, and the complete syntax can be viewed on the webpage https://bit.ly/Syntax_Journal. This study enrolled seven characteristic variables, which were transformed into a contingency table, resulting in seven contingency tables. Furthermore, to perform correspondence analysis, the characteristic variables must depend on district variables. Therefore, a Chi-Square test examines the dependency between the characteristic and the district variable. The results of the Chi-Square test from equations (3) and (4) can be seen in Table 4.

Table 4. Chi-Square Test

Category	χ-Square (χ^2)	df (v)	p-value
District vs X_1	101.41	30	1.11E-09
District vs X_2	122.15	90	0.0136
District vs X_3	55.45	60	0.0031
District vs X_4	25.02	30	0.7239
District vs X_5	149.20	90	8.858E-05
District vs X_6	144.96	120	0.0601
District vs X_7	274.50	180	2.392E-09

Based on Table 4, the results of the Chi-Square test show variables Existence of LPG Bases/Agents/Sellers (Stalls, Shops, Supermarkets, Mobile Gas Sellers) (X_4) and Place/Channel for Disposal of Liquid Waste from Most Families' Bathing/Washing Water (X_6) are not dependent on district variables. This is indicated by a p-value greater than 0.05. Therefore, these two variables do not need to be included in the analysis. Furthermore, MCA was carried out on the five variables significantly dependent on the district variable.

The principal coordinates $\mathbf{F}$ can be used to map the dependency between categories. The standard coordinates can be obtained from the calculation of equation (12).

$$\mathbf{F}_{(49\times 43)} = \begin{bmatrix} 0.343768 & 0.66048 & 0.367169 & \cdots & -0.08277 \\ -1.28021 & -1.28291 & 0.254724 & \cdots & 0.296152 \\ 0.297201 & 0.178664 & -0.14608 & \cdots & -0.03054 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.613031 & -0.41413 & -2.02701 & \cdots & 0.141413 \end{bmatrix}$$

The principal coordinates for all categories are obtained by multiplying the standard coordinates by the matrix $\boldsymbol{\Lambda}$, which is an $L \times L$ diagonal matrix of eigenvalues λ_ℓ according to equation

(12).

Table 5. Eigenvalues, Variance Explained, and Cumulative Diversity

ℓ	λ_ℓ	ϕ_δ	τ_D
1	0.146163968	10.7	10.7
2	0.08503761	6.2	16.9
3	0.065125825	4.7	21.7
⋮	⋮	⋮	⋮
43	0.002482092	0.1	100

Based on Table 5 from calculation equations (16) and (17), the results of the MCA on these five characteristic variables used 43 dimensions because they resulted in a cumulative diversity percentage of 100%. The total cumulative diversity percentage obtained when using two dimensions is 16.9%. Persisting with only two dimensions could lead to misleading information. Identifying infrastructure condition characteristics for each district based on qualitative variables is determined by the proximity of coordinate points between the district and the qualitative categories of infrastructure conditions generated by the Euclidean distance matrix. If the distance is closer, the category represents the infrastructure condition characteristics of the district. Conversely, if the distance is farther, it indicates otherwise.

Table 6. Euclidean Distance between Sub-Districts and Categories

	$X_{1,1}$	$X_{1,2}$	$X_{2,1}$	$X_{2,2}$	⋯	$X_{7,5}$	$X_{7,6}$
M_1	**2.1310**	3.3276	2.2260	2.6854	⋯	2.3771	4.8787
M_2	2.8129	**1.8476**	2.7816	3.6890	⋯	**2.5795**	5.1580
M_3	**2.0604**	3.2493	**2.0349**	2.8794	⋯	2.2457	4.7704
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
M_{31}	**2.1601**	3.3175	**2.1859**	2.9377	⋯.	2.3564	4.8649

Based on Table 6 from calculation equation (27), it can be observed that Arjasari district M_1 has the closest Euclidean distance to category $X_{1,1}$ compared to M_1's distances to other categories in variable X_1. Therefore, one of the characteristics of Arjasari district is the existence of the RPJMD (Medium-Term Regional Development Plan) $X_{1,1}$. The exact process is carried out for other districts and categories. Districts with similar characteristics can be grouped into one group based on Euclidean distance in Table 4.

The quantitative data in this study comprises variables that have undergone preprocessing, which consists of two stages: calculating totals for each sub-district based on equation (19) and

standardizing the data based on equation (20). Identification of sub-districts based on infrastructure condition characteristics from 3 quantitative variables is conducted using the cosine correlation method. The cosine angle calculation between two vectors uses the principal coordinate values of qualitative variables with quantitative data.

To obtain the coordinates of quantitative characteristic vectors for three variables, the values of $\boldsymbol{\Lambda}, \mathbf{R}$, dan γ are set to $\frac{1}{2}$. γ is set to $\frac{1}{2}$ to focus on the vectors and the spread of object points. The results of the principal coordinate points and the $\boldsymbol{\Psi}$ vector coordinates from equation (23) are used to identify sub-districts based on qualitative and quantitative characteristics. Qualitative characteristics are identified using the principal coordinate points from 7 qualitative variables, while quantitative characteristics are identified using the $\boldsymbol{\Psi}$ vector coordinates from three quantitative variables. The cosine correlation matrix distance calculation can be used because the cumulative diversity percentage or inertia between the principal qualitative coordinates and the quantitative characteristic vector coordinates is greater than 70%. Therefore, the calculation involves determining the cosine angle matrix between the principal qualitative coordinates and the quantitative characteristic vector coordinates.

The next step is to identify the quantitative characteristics of infrastructure conditions, which consist of three quantitative variables in the districts, using the cosine angle between two vectors. This will result in matrix $\mathbf{K}$ form calculation (26) that contains the correlation coefficients. Here is the interpretation of correlation based on the value of its correlation coefficient [30]:

Table 7. Correlation Value Range and Degree of Relationship

Correlation Value Range	Degree of Relationship
0.00-0.19	Very Weak
0.20-0.39	Weak
0.40-0.59	Moderate
0.60-0.79	Strong
0.80-1.00	Very Strong

Based on Table 7, if the cosine value of the angle is $\geq \pm 0.2$, it indicates that the two variables correlate, although the correlation is weak. In the cosine correlation matrix. Here is the Cosine Correlation Matrix between Principal Qualitative Coordinates and Quantitative Vector Coordinates.

Table 8. Cosine Correlation Matrix

	X_8	X_9	X_{10}		X_8	X_9	X_{10}
M_1	-0.4804 *	-0.2098 *	0.3721 *	M_{17}	-0.1485	0.0837	-0.1691
M_2	-0.1445	0.3768 *	-0.1199	M_{18}	-0.2773 *	0.3597 *	-0.0053
M_3	0.2392 *	-0.2831 *	0.0485	M_{19}	0.2525 *	0.0099	-0.1548
M_4	0.0655	0.2051 *	-0.1348	M_{20}	0.0801	-0.2225 *	-0.1525
M_5	0.1271	-0.0035	0.2806 *	M_{21}	-0.1907	0.0045	-0.0915
M_6	0.0438	-0.2743 *	-0.0485	M_{22}	-0.2568 *	0.0059	0.0194
M_7	-0.2542 *	0.2701 *	-0.3591 *	M_{23}	-0.0962	-0.3140 *	-0.2901
M_8	-0.1523	-0.1064	-0.0743	M_{24}	-0.1783	0.0258	0.0162
M_9	-0.1243	0.1551	-0.0337	M_{25}	0.2012 *	0.1002	0.3589 *
M_{10}	-0.0563	0.4478 *	-0.0987	M_{26}	0.0776	-0.4224 *	0.1774
M_{11}	0.1336	-0.2513 *	0.2673 *	M_{27}	-0.1899	-0.1292	-0.0743
M_{12}	0.3351 *	0.1662	-0.4400 *	M_{28}	0.0524	-0.0692	-0.0324
M_{13}	0.1704	0.0505	-0.3751 *	M_{29}	0.1187	0.3426 *	-0.0078
M_{14}	-0.0197	0.0635	0.0256	M_{30}	-0.1422	-0.1302	-0.0116
M_{15}	-0.0673	0.0651	0.0510	M_{31}	-0.0636	-0.1146	-0.0211
M_{16}	0.3178 *	-0.1386	-0.0116				

Note: The cosine values of the angle marked with (*) are $\geq \pm 0.2$, which indicates that a district has specific infrastructure condition characteristics.

Table 8 shows that in Baleendah district (M_2), the number of households using PLN electricity is above average. Baleendah district (M_2) does not correlate with the characteristics of village road development length during the year 2022 (X_8) and the number of habitable houses (X_{10}). Therefore, it can be concluded that the infrastructure condition characteristics based on quantitative variables in Baleendah district (M_2) are the number of households using PLN electricity above average. The districts that do not correlate with quantitative characteristics include Cilengkrang (M_8), Cileunyi (M_9), Dayeuhkolot (M_{14}), Ibun (M_{15}), Kertasari (M_{17}), Margahayu (M_{21}), Pameungpeuk (M_{24}), Pasirjambu (M_{27}), Rancabali (M_{28}), Solokanjeruk (M_{30}), and Soreang (M_{31}). Identification of district infrastructure condition characteristics based on dependency and cosine correlation is conducted by grouping qualitative and quantitative variables. The results of this grouping are presented in table 9:

Table 9. Groups of sub-districts based on hybrid multiple correspondence analysis

Group	District	Infrastructure Condition
1	• Arjasari (M_1)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • Construction of the length of village roads during 2022 is below average (X_8) • The number of families using PLN electricity is below average (X_9) • The number of livable houses is above average (X_{10})
2	• Baleendah (M_2)	• There is no RPJMD ($X_{1,2}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$)
3	• Banjaran (M_3)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • Construction of the length of village roads during 2022 is above average(X_8) • The number of families using PLN electricity is below average (X_9)
4	• Bojongsoang (M_4)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • The number of families using PLN electricity is above average (X_9)
5	• Cangkuang (M_5)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • The number of livable houses is above average (X_{10})

Group	District	Infrastructure Condition
6	• Cicalengka (M_6)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • The number of families using PLN electricity is below average (X_9)
7	• Cikancung (M_7)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • Final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • Construction of the length of village roads during 2022 is below average (X_8) • The number of families using PLN electricity is above average (X_9) • The number of livable houses is below average (X_{10})
8	• Cilengkrang (M_8)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is weak ($X_{2,2}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from water springs($X_{7,3}$)
9	• Cileunyi (M_9)	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$)
10	• Cimaung (M_{10})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • The number of families using PLN electricity is above average(X_9)
11	• Cimenyan (M_{11})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • The number of families using PLN electricity is below average(X_9) • The number of livable houses is above average (X_{10})

Group	District	Infrastructure Condition
12	• Ciparay (M_{12})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • Construction of the length of village roads during 2022 is above average (X_8) • The number of livable houses is below average (X_{10})
13	• Ciwidey (M_{13})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • The number of livable houses is below average (X_{10})
14	• Dayeuhkolot (M_{14}) • Margahayu (M_{21})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$)
15	• Ibun (M_{15}) • Kertasari (M_{17})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from water springs ($X_{7,3}$)
16	• Katapang (M_{16}) • Majalaya (M_{19})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • Source of water for bathing/washing from a drilled well or pump ($X_{7,5}$) • Construction of the length of village roads during 2022 is above average (X_8)
17	• Kutawaringin (M_{18})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is very strong ($X_{2,3}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a well ($X_{7,4}$) • Construction of the length of village roads during 2022 is below average (X_8) • The number of families using PLN electricity is above average(X_9)

Group	District	Infrastructure Condition
18	• Margaasih (M_{20})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • The number of families using PLN electricity is below average (X_9)
19	• Nagreg (M_{22})	• There is an RPJMD ($X_{1,1}$) The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • Final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • Construction of the length of village roads during 2022 is below average (X_8)
20	• Pacet (M_{23})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from water springs ($X_{7,3}$) • The number of families using PLN electricity is below average (X_9)
21	• Pameungpeuk (M_{24})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • Final waste disposal site elsewhere ($X_{5,1}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$)
22	• Pangalengan (M_{25})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a metered tap ($X_{7,1}$) • Construction of the length of village roads during 2022 is above average (X_8) • The number of livable houses is above average (X_{10})
23	• Paseh (M_{26})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$) • The number of families using PLN electricity is below average (X_9)

Group	District	Infrastructure Condition
24	• Pasirjambu (M_{27})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$)
25	• Rancabali (M_{28})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in a hole in the ground ($X_{5,2}$) • The source of water for bathing/washing is from water springs($X_{7,3}$)
26	• Rancaekek (M_{29})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is some lighting on the main road ($X_{3,2}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • Source of water for bathing/washing from a drilled well or pump ($X_{7,5}$) • The number of families using PLN electricity is above average (X_9)
27	• Solokanjeruk (M_{30})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site for faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a drilled well or pump ($X_{7,5}$)
28	• Soreang (M_{31})	• There is an RPJMD ($X_{1,1}$) • The cell phone/phone signal is strong ($X_{2,1}$) • There is most of the lighting on the main road ($X_{3,1}$) • A final disposal site faeces in tanks/wastewater management installations ($X_{5,4}$) • The source of water for bathing/washing is from a well ($X_{7,4}$)

Based on Table 9, the districts formed by identifying infrastructure characteristics based on qualitative and quantitative variables resulted in 28 district groups, each with different infrastructure characteristics. Ciwidey District (M_{13}) is in Group 13, and Soreang District (M_{31}) is in Group 28, Initially, they were grouped together when only qualitative characteristics were considered, as they shared the same qualitative characteristics, including There is an RPJMD ($X_{1,1}$),

The cell phone/phone signal is strong ($X_{2,1}$), There is most of the lighting on the main road ($X_{3,1}$), Final disposal site for faeces in tanks/wastewater management installations, and The source of water for bathing/washing is from a well ($X_{7,4}$). However, they ended up in different groups because, when quantitative characteristics were considered, Ciwidey District was found to have a quantitative characteristic that the number of livable houses is below average (X_{10}). In contrast, Soreang District does not have any quantitative characteristics. The grouping of other districts is similarly based on qualitative and quantitative variables. Therefore, it is possible for districts to be grouped together based on qualitative variable identification initially, but they may later be separated when quantitative variables are also considered.

4. CONCLUSION

Hybrid multiple correspondence analysis with cosine correlation is a serial combination of correspondence analysis with a cosine correlation approach aimed at identifying dependencies and correlations between two or more qualitative variables and several quantitative variables. Before calculating hybrid multiple correspondence analysis, a chi-square test is necessary to determine which characteristic variables depend on the districts. Of the seven characteristics, five are dependent on the districts. In this study, cumulative variance in two dimensions only reaches 16.9%, which may result in a distorted or incomplete representation of the data, potentially leading to incorrect conclusions. Therefore, the dimensions used to group districts in Bandung Regency are based on the number of categories of qualitative variables, totalling 43 dimensions, encompassing 100% inertia, and grouping cannot be obtained based on a two-dimensional correspondence map. Euclidean distance and cosine correlation are solutions for identifying sub-districts based on the characteristics of infrastructure conditions, which consist of qualitative and quantitative variables with nominal and ratio measurement scales. Based on the results of hybrid multiple correspondence analysis with cosine correlation, 28 district groups were obtained based on the characteristics of infrastructure conditions. Thus, the government can address infrastructure issues in Bandung Regency by evaluating each district's infrastructure problems based on their differing characteristics.

Hybrid of multiple correspondence analysis with cosine correlation can be used to obtain object characteristics based on the dependence between categories of qualitative variables with several quantitative variables using the Euclidean distance matrix. Hybrid multiple correspondence analysis can be applied to another study as long as there are object variables and characteristic

variables consisting of qualitative variables with more than two categories and several quantitative variables. Future studies should consider grouping objects using cluster analysis methods to create more flexible groups as desired.

ACKNOWLEDGEMENTS

The authors would like to thank Asep Rochmansyah, S.Si., M.I.P., who has helped get data and support from The Ministry of Education, Culture, Research, and Technology Fundamental Research 2024 Number (3907/UN6.3.1/PT.00/2024).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] M. Greenacre, J. Blasius, eds., Multiple Correspondence Analysis and Related Methods, Chapman and Hall/CRC, 2006. https://doi.org/10.1201/9781420011319.

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

[3] J.F. Hair Jr., W.C. Black, B.J. Babin, et al. Multivariate Data Analysis, Prentice Hall, 2010.

[4] O.N. C, M. Greenacre, Correspondence Analysis in R, with Two- and Three-Dimensional Graphics: The ca Package, J. Stat. Softw. 20 (2007), 1-13. https://doi.org/10.18637/jss.v020.i03.

[5] E.J. Beh, R. Lombardo, eds., Correspondence Analysis, Wiley, 2014. https://doi.org/10.1002/9781118762875.

[6] E.J. Beh, R. Lombardo, Correspondence Analysis Using the Cressie–Read Family of Divergence Statistics, Int. Stat. Rev. 92 (2024), 17–42. https://doi.org/10.1111/insr.12541.

[7] K.K. Kamalja, N.V. Khangar, Multiple Correspondence Analysis and Its Applications, Electron. J. Appl. Stat. Anal. 10 (2017), 432–462.

[8] A.C. Rencher, Methods of Multivariate Analysis, Wiley, 2002. https://doi.org/10.1002/0471271357.

[9] A.I. D'Enza, M. Greenacre, Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets, in: A. Di Ciaccio, M. Coli, J.M. Angulo Ibanez (Eds.), Advanced Statistical Methods for the Analysis of Large Data-Sets, Springer, Berlin, Heidelberg, 2012: pp. 453–463. https://doi.org/10.1007/978-3-642-21037-2_41.

[10] M. Greenacre, Correspondence Analysis in Practice, Chapman and Hall/CRC, 2017.

[11] D. Ayele, T. Zewotir, H. Mwambi, Multiple Correspondence Analysis as a Tool for Analysis of Large Health Surveys in African Settings, Afr. Health Sci. 14 (2015), 1036. https://doi.org/10.4314/ahs.v14i4.35.

[12] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Pearson/Prentice Hall, Upper Saddle

River, NJ, 2007.

[13] N. Sourial, C. Wolfson, B. Zhu, et al. Correspondence Analysis Is a Useful Tool to Uncover the Relationships among Categorical Variables, J. Clin. Epidemiol. 63 (2010), 638–646. https://doi.org/10.1016/j.jclinepi.2009.08.008.

[14] I. Ginanjar, A.I. Nurwahidah, J. Suprijadi, 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.

[15] 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.

[16] F.F. Pinem, I. Ginanjar, T. Purwandari, Identifikasi Karakteristik Kawasan Kumuh Disetiap Kecamatan di Kabupaten Bandung dengan Analisis Korespondensi Multipel, BIAStatistics J. Stat. Theory Appl. 17 (2023), 95–106.

[17] D. Sukarno, Infrastruktur Dan Teknologi Pada Kecamatan-Kecamatan Di Kabupaten Bandung Dalam Mendukung Pelaksanaan Program Paten (Pelayanan Administrasi Terpadu Kecamatan), J. Manaj. Pelayanan Publ. 1 (2017), 109–124. https://doi.org/10.24198/jmpp.v1i1.13565.

[18] F.A. Syaiful, A.Y. Koswara, Penentuan Prioritas Pengembangan Infrastruktur Wilayah Pesisir Kecamatan Sangatta Utara dan Kecamatan Sangatta Selatan Kabupaten Kutai Timur, J. Tek. ITS, 9 (2021), D161–D166.

[19] T. Murakami, Orthonormal Principal Component Analysis for Categorical Data as a Transformation of Multiple Correspondence Analysis, in: T. Imaizumi, A. Nakayama, S. Yokoyama (Eds.), Advanced Studies in Behaviormetrics and Data Science, Springer, Singapore, 2020: pp. 211–231. https://doi.org/10.1007/978-981-15-2700-5_13.

[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] A. Agresti, An Introduction to Categorical Data Analysis Second Edition, Second Edition, John Wiley & Sons, 2007.

[22] B. Le Roux, H. Rouanet, Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis, Springer, 2004.

[23] I. Ginanjar, Hybrid Korespondensi untuk Menganalisis Obyek Berdasarkan Kategori Kolom dan Karakteristik Obyek, Pros. Semin. Nas. Stat. 2 (2011), 303–313.

[24] 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.

[25] Y. Kawada, Cosine Similarity and the Borda Rule, Soc. Choice Welf. 51 (2018), 1–11. https://doi.org/10.1007/s00355-017-1104-2.

[26] A.W. Kusuma, I.G. Srinadi, K. Sari, Aplikasi Analisis Korespondensi untuk Melihat Karakteristik Usaha Pariwisata di Provinsi Bali, E-J. Mat. 5 (2016), 76–81.

[27] N.J. Le Roux, S. Gardner-Lubbe, J.C. Gower, The Analysis of Distance of Grouped Data with Categorical Variables: Categorical Canonical Variate Analysis, J. Multivar. Anal. 132 (2014), 9–24. https://doi.org/10.1016/j.jmva.2014.07.014.

[28] B. Santosa, Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Graha Ilmu, Yogyakarta, 2007.

[29] R. Balaji, R.B. Bapat, S. Goel, Generalized Euclidean Distance Matrices, arXiv:2103.03603 [math.FA] (2021). https://doi.org/10.48550/arXiv.2103.03603.

[30] Sugiyono, Metode Penelitian Kunatitatif Kualitatif dan R&D, Alfabeta, Bandung, 2016.