Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:36

https://doi.org/10.28919/cmbn/8444

ISSN: 2052-2541

RECATEGORIZATION METHOD BASED ON DEPENDENCE BETWEEN QUALITATIVE VARIABLES USING JOINT CORRESPONDENCE ANALYSIS WITH ELLIPTICAL CONFIDENCE REGIONS

RENATA SYIFA KRISTANTO, IRLANDIA GINANJAR*, TITI PURWANDARI

Department of Statistics, University of Padjadjaran, Bandung, Indonesia

Abstract: Joint Correspondence Analysis (JCA) is a development method of Multiple Correspondence Analysis (MCA) that uses an algorithm to increase the percentage of variance. However, if the analysis used a large number of categories in qualitative data and there is no dependency, the analysis result in two dimensions may not be representative because the data variance is divided into several dimensions. Therefore, a recategorization method based on category dependencies is necessary to get a representative result. Elliptical confidence regions are the technique that can identify the contribution of dependence between two variables. Categories with insignificant contribution of dependencies are combined with other categories based on the shortest Euclidean distance. The novelty of this research is there is a stage of combining categories in correspondence analysis to reach a variance percentage of 70% in two dimensions. The study used data from the Environmental Quality Index (EQI) of Bandung Regency. The EQI consists of the Air Quality Index, the Water Quality Index, and the Land Cover Index. There are 8 characteristics with 37 categories and 31 districts used. According to the Chi-Square Test, 6 significant characteristics have a dependency on the district. Elliptical confidence regions were calculated six times based on simple

*Corresponding author

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

Received January 12, 2024

correspondence analysis. There are differences treatment of characteristics and districts. The recategorization of characteristics is based on elliptical confidence regions, while districts were categorized based on Euclidean distance because the use of elliptical confidence regions generated six different results. The final two-dimensional map of JCA can explain 70.1% of the data variation in the 27th analysis with a total grouping of 5 districts and 17 categories.

Keywords: elliptical confidence regions; EQI; joint correspondence analysis; recategorization.

2020 AMS Subject Classification: 92-10.

1. INTRODUCTION

Correspondence analysis is the method used to create perceptual maps that describe the relationships between variables. The points in perceptual mapping indicate the category of each qualitative variable, where the distance between points has substantive meaning [1]. The visualization graphic aims to aid in interpreting the characteristics of each variable and the relationships between the variables [2].

Simple correspondence analysis is applied to data with two categorical variables [3]. Multiple correspondence analysis (MCA) can be performed simultaneously for datasets with more than two categorical variables and can reveal dependencies between more than two variables and their simultaneous impact on the observed variable [4]. However, MCA has the disadvantage of difficulties in creating a two-dimensional map when the dataset contains numerous categorical variables [5]. Data with numerous categories may result in a lack of dependency between the variables under study, leading to a low percentage of variance in the two dimensions of MCA results [2].

Joint Correspondence Analysis (JCA) is a development method of MCA. JCA can generate greater variance in two dimensions, overcoming the problem of MCA in forming two-dimensional maps [6][7]. JCA has several advantages over MCA. Specifically, JCA can optimize adjustments to all off-diagonal crosses, resulting in a more comprehensive understanding of the relationships between several categorical variables in multivariate cases [8]. JCA also perfectly reproduces simple CA in the two-variable case, since it is also focused exclusively on the single off-diagonal cross-tabulation [9]. The algorithm for joint correspondence analysis is similar to MCA, which uses the Burt matrix derived from the indicator matrix, but there is a stage of Burt matrix reconstruction until convergence [4]. The disadvantage of JCA is an unrepresentative result produced if the data contains many categorical variables [2]. To identify this, recategorization can be performed based on elliptical confidence regions.

Elliptical confidence regions are areas in the shape of an ellipse that indicate the significance of categories [10]. If an origin point falls within the ellipse, it significantly contributes to the structure of dependence. If it falls outside the ellipse, the category is not significant in contributing to the structure of dependence, which can lead to misinterpreting in drawing conclusions [11]. Therefore, it is necessary to merge non-significant categories with other similar categories based on their proximity, using Euclidean distance.

The study of recategorization using JCA with elliptical confidence regions will be conducted to group districts in Bandung Regency based on environmental quality variables. The data used is the Environmental Quality Index (EQI) indicator data from the Bandung Regency in 2022. The EQI values consist of the Air Quality Index, Water Quality Index, and Land Cover Index. The Air Quality Index reflects the condition of air quality, the Air Quality Index depicts air quality in a region, and the Land Cover Index illustrates land cover quality calculated based on forest conditions and non-forest vegetation cover. The data comprises eight qualitative variables with a total of 68 categories. Environmental data indicates the quality of the environment in a region. A value representing the environmental quality of a region is the Environmental Quality Index. Bandung Regency still falls into the moderate category of environmental quality, requiring evaluation for improvement. Therefore, this research aims to create a two-dimensional map that explains the environmental quality characteristics of each district in Bandung Regency. The results of this analysis can help the government in the evaluation of the value of the EQI.

2. MATERIAL AND METHOD

Elliptical confidence regions for JCA in this study are employed to determine dependencies among categorical variables, as well as within each category. Any category that does not significantly contribute to its association structure will be combined with another category based on the closest distance, using Euclidean distance.

2.1. Data Sources

This study aims to assess the environmental quality in Bandung Regency, which comprises 31 districts and 280 sub-districts. The data utilized consists of environmental indicators obtained from the Supporting Area Survey. Eight characteristics variables serve as column categories, including variables related to Toilet Facility Usage (X_1), Final Disposal of Fences (X_2), Liquid Waste Drainage (X_3), Drinking Water Source (X_4), Bathing Water Source (X_5), Household Waste Disposal (X_6), Forest Area Function (X_7), and Waste Processing (X_8). Furthermore, district data are used as row categories. The 8 characteristics were transformed into a contingency table with districts as rows and each category as columns. This resulted in 8 contingency tables that can be used in the chi-square test.

2.2. Contingency Table

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

Table 1. Contingency Table

District (D)	Characteristic Variables (X)						
	1	2	…	j	…	$q_{\tilde{k}}$	Total
1	n_{11}	n_{12}	…	n_{1j}	…	$n_{1q_{\tilde{k}}}$	$n_{1\bullet}$
2	n_{21}	n_{22}	…	n_{2j}	…	$n_{2q_{\tilde{k}}}$	$n_{2\bullet}$
…	…	…	…	…	…	…	…
i	n_{j1}	n_{j2}	…	n_{ij}	…	$n_{jq_{\tilde{k}}}$	$n_{h\bullet}$
…	…	…	…	…	…	…	…
q_1	n_{q_11}	n_{q_12}	…	n_{q_1j}	…	$n_{q_1q_{\tilde{k}}}$	$n_{q_1\bullet}$
Total	$n_{\bullet1}$	$n_{\bullet2}$	…	$n_{\bullet j}$	…	$n_{\bullet q_{\tilde{k}}}$	n

According to Table 1, can be obtained cross-tabulation matrix $\mathbf{N} = (n_{ij})$ and correspondence matrix in simple correspondence analysis can derived:

$$\widetilde{\mathbf{P}} = \frac{n_{ij}}{n} = (\tilde{p}_{ij}) \tag{1}$$

p_{ij} is joint probability estimator districts and characteristics variable. $\tilde{p}_{i\bullet} = \frac{n_{i\bullet}}{n}$ is marginal probability estimator of characteristics and $\tilde{p}_{\cdot j} = \frac{n_{\bullet j}}{n}$ is marginal probability estimator of districts [12]. The marginal totals of rows and columns of $\widetilde{\mathbf{P}}$ are the vector $\tilde{\boldsymbol{r}}$ with $\tilde{\boldsymbol{r}}_i = \sum_{j=1}^{q_{\tilde{k}}} \tilde{p}_{ij} = \tilde{p}_{i\bullet}$ and $\tilde{\boldsymbol{c}}$ with $\tilde{\boldsymbol{c}}_j = \sum_{i=1}^{q_1} \tilde{p}_{ij} = \tilde{p}_{\bullet j}$, which are row and column mass vectors. The diagonal matrix $\widetilde{\mathbf{D}}_r = diag(\tilde{\boldsymbol{r}})$ and $\widetilde{\mathbf{D}}_c = diag(\tilde{\boldsymbol{c}})$.

2.3. Chi-Square Test

The qualitative data utilized in correspondence analysis must demonstrate interdependence among its variables. In this study, characteristics must depend on the district. The Chi-Square test is one of the hypothesis tests that can be applied to evaluate the dependence among categorical variables in a contingency table [12]. The following is the hypothesis for the chi-square test of the two variables qualitative [10]:

H_0 : $\tilde{p}_{ij} = \tilde{p}_{i\bullet}\tilde{p}_{\bullet j}$; (there is no dependency between the two variables)

H_1 : $\tilde{p}_{ij} \neq \tilde{p}_{i\bullet}\tilde{p}_{\bullet j}$; (there is a dependency between the two variables)

The test statistics of the two variable qualitative chi-square test are as follows

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

Where n is the number of observations, $p_{i\bullet}$ is the marginal probability of the i^{th} district, $p_{\bullet j}$ is the marginal probability of j^{th} characteristics, p_{ij} joint probability of the i and j, q_1 is the number of categories on i^{th}, and $q_{\tilde{k}}$ is the number of categories on j^{th}. The test criterion of the chi-square of two qualitative variables is repulsion, H_0 if $\chi^2 \geq \chi^2_{\alpha(q_1-1)(q_{\tilde{k}}-1)}$ it means that there is a dependence between variables with n $\alpha = 0{,}1$.

2.4. Simple Correspondence Analysis

Simple correspondence analysis (SCA) is a multivariate method based on matrix data [13]. SCA is generally performed on two-way contingency tables, which contain data with two categorical variables [14]. The analysis of simple correspondence uses a correspondence matrix as in formula (1). The row and column profile coordinate relative to the principal axis can be obtained using SVD as follows.

$$\tilde{\mathbf{S}} = \widetilde{\mathbf{D}}_r^{-\frac{1}{2}}\left(\widetilde{\mathbf{P}} - \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^T\right)\widetilde{\mathbf{D}}_c^{-\frac{1}{2}} \tag{3}$$

$$\tilde{\mathbf{S}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}_\alpha\widetilde{\mathbf{V}}^T \tag{4}$$

Equation (3) represents a standard residuals matrix, while equation (4) calculates the Singular Value Decomposition (SVD) with $\widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}} = \mathbf{I}$, $\widetilde{\mathbf{D}}_\alpha = \text{diag}(\widetilde{\boldsymbol{d}})$**,** with $\widetilde{\boldsymbol{d}}$ is vector with elements that are singular values or the root of the eigenvalue $\tilde{\lambda}_{\tilde{\ell}}$ from matrix $\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}$ and $\tilde{\ell} = 1,2,3,\ldots,\tilde{L}$, then $\tilde{d}_{\tilde{\ell}} = \sqrt{\tilde{\lambda}_{\tilde{\ell}}}$. The vector $\widetilde{\boldsymbol{d}}$ contains elements in descending order. The $\tilde{L}$ represents the number of non-zero eigenvalues, with $\tilde{L} = \min(q_1, q_{\tilde{k}}) - 1$ [15]. The graphical representation of the dependency between rows and columns can be depicted based on equations (3) and (4). The i^{th} row profile and j^{th} column profile are shown with the principal coordinates for the row and column with the following equation.

$$\tilde{\mathbf{F}} = \widetilde{\mathbf{D}}_r^{-\frac{1}{2}}\widetilde{\mathbf{U}}\widetilde{\mathbf{D}}_\alpha = \left(\tilde{f}_{i\tilde{\ell}}\right) \tag{5}$$

$$\widetilde{\mathbf{G}} = \widetilde{\mathbf{D}}_c^{-\frac{1}{2}}\widetilde{\mathbf{U}}\widetilde{\mathbf{D}}_\alpha = \left(\tilde{g}_{j\tilde{\ell}}\right) \tag{6}$$

with $\tilde{\mathbf{F}}$ is the principal coordinates of rows and $\widetilde{\mathbf{G}}$ is the principal coordinate of columns. The total variance of matrix data is calculated by inertia, using the following equation.

$$\phi^2 = \sum_{i=1}^{q_1}\sum_{j=1}^{q_{\tilde{k}}}\frac{\left(\tilde{p}_{ij} - \tilde{r}_i\tilde{c}_j\right)^2}{\tilde{r}_i\tilde{c}_j} \tag{7}$$

Inertia can indicate the quality of the resulting map. Two-dimensional maps can be created when the percentage of inertia in two dimensions reaches 70% [16]. Two-dimensional maps are useful because they show information from the third and higher dimensions [17].

2.5. Elliptical Confidence Regions

Recategorization for characteristics used elliptical confidence regions. Elliptical confidence regions can be used to see the contribution of dependence between two variables and can be obtained from the results of a two-dimensional map. The confidence region resulting from the elliptical confidence region considers that in correspondence analysis, the principal inertia of the first axis is always greater than that of the second axis $\tilde{\lambda}_1 > \tilde{\lambda}_2$. The calculation of elliptical confidence regions is very important because each category must contribute to the dependence structure. Otherwise, the conclusions drawn will not be appropriate. Therefore, recategorization is necessary based on the results of elliptical confidence regions using Euclidean distance. To illustrate, the $100(1-\alpha)\%$ confidence ellipses for the i-th nominal category can be formulated by considering the semi-minor and semi-major axis length along the ℓ^{th} principal axis [17].

$$x_{j(\alpha)} = \tilde{d}_1 \sqrt{\frac{\chi^2_\alpha}{X^2}\left(\frac{1}{\tilde{p}_{\bullet j}} - \sum_{\tilde{\ell}=3}^{\tilde{L}} \left(\frac{\tilde{g}_{j\tilde{\ell}}}{\tilde{d}_{\tilde{\ell}}}\right)^2\right)} \text{ and } y_{j(\alpha)} = \tilde{d}_2 \sqrt{\frac{\chi^2_\alpha}{X^2}\left(\frac{1}{\tilde{p}_{\bullet j}} - \sum_{\tilde{\ell}=3}^{\tilde{L}} \left(\frac{\tilde{g}_{j\tilde{\ell}}}{\tilde{d}_{\tilde{\ell}}}\right)^2\right)} \quad (8)$$

Where $\tilde{p}_{\bullet j}$ is j^{th} marginal probability for districts category (row category), χ^2_α is the $(1-\alpha)$ percentile from Chi-Square Statistics in equation (1) with $(q_1 - 1)(q_{\tilde{k}} - 1)$ degree of freedom, χ^2 is Chi-Square values from equation (1), $\tilde{d}_{\tilde{\ell}}$ is $\tilde{\ell}^{th}$ of singular values, $x_{j(\alpha)}$ is the length of the semi-major axis for j^{th} category, $y_{j(\alpha)}$ is the length of the semi-minor axis for j^{th} category, and $\tilde{g}_{j\ell}$ is j^{th} column coordinate.

The categories that significantly contribute to the association structure are represented by Approximate p-values. The hypothesis for Approximate p-values is as follows:

H_0 : $\tilde{g}_j = 0$; (j-th column category does not contribute)

H_1 : $\tilde{g}_j \neq 0$; (j-th column category is contributing)

The formula for the approximate p-value is as follows [18]:

$$(p - value)_{j,D} \approx P\left\{\chi^2_\alpha > X^2 \left(\frac{1}{\tilde{p}_{\bullet j}} - \sum_{\tilde{\ell}=D+1}^{\tilde{L}} \left(\frac{\tilde{g}_{j\tilde{\ell}}}{\tilde{d}_{\tilde{\ell}}}\right)^2\right)^{-1} \sum_{\tilde{\ell}=1}^{\tilde{L}} \left(\frac{\tilde{g}_{j\tilde{\ell}}}{\tilde{d}_{\tilde{\ell}}}\right)^2\right\} \quad (9)$$

with $D = 2$ for two-dimensional map. Categories with a p-value greater than the significance level 0.1 are combined with other categories based on the closest Euclidean distance. This is necessary to ensure that the two dimensions represent 70% of the variance.

2.6. Joint Correspondence Analysis

Joint correspondence analysis (JCA) is a development method from MCA that can be used on data with more than two qualitative variables simultaneously. MCA uses the Burt matrix from the Indicator matrix. Each row of the indicator matrix contains 1 if the element belongs to the variable category and 0 if the element does not belong to that category [19]. If the number of districts is expressed by n $(m = 1,2,\dots,n)$, the number of characteristics variable is expressed by p $(k = 1,2,\dots,p)$, and q_k is the number of categories in the k^{th}, then the indicator matrix is formulated as follows:

$$\mathbf{Z} = \left[\mathbf{Z}_1\ \mathbf{Z}_2\ \dots\ \mathbf{Z}_p\right] \tag{10}$$

If $\mathbf{Z}_k$ is the indicator matrix of the k^{th} variables, so z_{mj_k} is the (m, j) element of $\mathbf{Z}_k$ with $j = 1,2,\dots,q_k$. Therefore, the indicator matrix and its elements can be expressed as follows.

$$\mathbf{Z}_{\boldsymbol{k}} = \begin{bmatrix} \mathrm{z}_{1k1} & \cdots & \mathrm{z}_{1kq_k} \\ \vdots & \ddots & \vdots \\ \mathrm{z}_{nk1} & \cdots & \mathrm{z}_{nkq_k} \end{bmatrix} \tag{11}$$

The indicator matrix $\mathbf{Z}$ in equation (11) is equal to $n \times Q$ with $Q = \sum_{k=1}^{p} q_k$. From the indicator matrix, the Burt matrix can be obtained by cross-tabulation the indicator matrix.

$$\mathbf{B} = \mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{N}_{12} & \dots & \mathbf{N}_{1p} \\ \mathbf{N}_{21} & \mathbf{D}_2 & \dots & \mathbf{N}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}_{p1} & \mathbf{N}_{p2} & \dots & \mathbf{D}_p \end{bmatrix} = (b_{q\tilde{q}}) \tag{12}$$

$$b = \sum_{q=1}^{Q}\sum_{\tilde{q}=1}^{Q} b_{q\tilde{q}} \tag{13}$$

After obtaining the Burt matrix, the total elements of the Burt matrix can be calculated, which is known as the Burt correspondence matrix. The Burt correspondence matrix is obtained by dividing the Burt matrix by the total number of Burt matrix element values.

$$\mathbf{P} = \frac{1}{b}\mathbf{B} \tag{14}$$

The Burt matrix is symmetric, so the equations for the rows and columns are the same as the following equation [1]:

$$\boldsymbol{r} = \boldsymbol{c} = \frac{1}{b}\mathbf{B1} \tag{15}$$

The row and column proportions of the Burt matrix are equal, then:

$$\mathbf{D}_r = \mathbf{D}_c = diag(\boldsymbol{c}) \tag{16}$$

Then Burt's standard residual matrix is obtained to represent the dependence between categorical variables.

$$\mathbf{S} = \mathbf{D}_c^{-\frac{1}{2}}(\mathbf{P} - \boldsymbol{c}\boldsymbol{c}^T)\mathbf{D}_c^{-\frac{1}{2}} \tag{17}$$

In correspondence analysis, eigenvalues will be obtained using the following formula:

$$\mathbf{S} = \mathbf{V\Lambda V}^T \tag{18}$$

In equation (18), each column $\mathbf{V}$ is an orthogonal matrix $\mathbf{V}^{-1} = \mathbf{V}^T$, then $\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and contains the eigenvector $\boldsymbol{v}_\ell$ of matrix $\mathbf{S}$ corresponding to λ_ℓ, with $\ell = 1,2,\cdots,L$ and L represents the number of non-zero eigenvalues. The notation for each column of matrix $\mathbf{V}$ is $\mathbf{V} = (\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_L)$. The matrix $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues (λ_ℓ) and can be denoted by $\mathbf{\Lambda} = diag(\boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_L)$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_L$. From the eigenvalue, the standard coordinates will be obtained to map points on a two-dimensional map.

$$\mathbf{H} = \mathbf{D}_\mathrm{c}^{-\frac{1}{2}}\mathbf{V} = \left(h_{q\ell}\right) \tag{19}$$

and the principal coordinates are

$$\mathbf{F} = \mathbf{D}_\mathrm{c}^{-\frac{1}{2}}\mathbf{V\Lambda}^{\frac{1}{2}} = (f_{q\ell}) \tag{20}$$

The rows in the $\mathbf{F}$ matrix represent categories, while the columns in the $\mathbf{F}$ matrix represent the coordinates for each dimension. In JCA, the Burt matrix from MCA is reconstructed by updating all the main diagonal values of the Burt matrix with the new main diagonal values of the Burt matrix without changing other values in the Burt matrix. This step aims to increase the variance percentage. The calculations for the reconstruction of the Burt matrix at the t-th iteration is as follows:

$$\widehat{\mathbf{B}}_{(t)} = \left(\hat{b}_{q\tilde{q}(t)}\right), \hat{b}_{q\tilde{q}(t)} = b\ c_{q(t-1)}c_{\tilde{q}(t-1)}\left(1 + \sum_{\ell=1}^{L}\lambda_{\ell(t-1)}^2 h_{q\ell(t-1)}\, h_{\tilde{q}\ell(t-1)}\right) \tag{21}$$

The iteration process is carried out until the state converges. The convergent state is when the absolute of all element matrix E is less than 0.0001 [14]. The convergent state can be obtained using the following equation.

$$\mathrm{E} = \widehat{\mathbf{B}}_{(t+1)} - \widehat{\mathbf{B}}_{(t)} = (\varepsilon_{q\tilde{q}}) \tag{22}$$

$$\theta = \max\left(|\varepsilon_{q\tilde{q}}|\right) \tag{23}$$

The percentage of variance in the two dimensions can be calculated if convergent results are obtained, where $\theta \leq 0.0001$. The quality of the map produced by JCA can be evaluated based on the inertia percentage, which indicates the variance of the data [14].

$$\text{trace}(\mathbf{\Lambda}) = \text{trace}(\mathbf{S}^2) = \text{trace}(\mathbf{F}^T\mathbf{F}) = \text{trace}(\mathbf{F}\mathbf{F}^T) = \sum_{\ell=1}^{L} \lambda_\ell \tag{24}$$

The variance of each dimension and the cumulative variance can be obtained from the following equation.

$$\phi_d = \left(\frac{\lambda_d}{\sum_{\ell=1}^{L} \lambda_\ell}\right) \tag{25}$$

$$\tau_D = \left(\frac{\sum_{d=1}^{D} \lambda_d}{\sum_{\ell=1}^{L} \lambda_\ell}\right) \tag{26}$$

with ϕ_d is variance coverage for each d^{th} dimension with $d = 1,2,\cdots,L$, τ_D is cumulative variance for $D = 2$ dimensions, λ_d is d^{th} eigen value, and λ_ℓ is ℓ^{th} eigen value. If the percentage variance in two dimensions is less than 70%, there may be categories that do not have significant contribution dependence. Therefore, it is necessary to calculate elliptical confidence regions to determine which categories have significant dependencies.

2.7. Euclidean Distance

Euclidean distance is the geometric distance between two objects. In characteristic recategorization, Euclidean distance is calculated based on the principal coordinate value within elliptical confidence regions, different from object recategorization, where Euclidean distance is based on the principal coordinate value resulting from JCA. If the distance between the objects or characteristics is smaller, then the similarities between them will be closer to each other. Euclidean distance from principal coordinate value within elliptical confidence regions with vectors

$\tilde{\boldsymbol{g}}_{j\tilde{\ell}} = (\tilde{g}_{j1}, \tilde{g}_{j2}, \dots, \tilde{g}_{j\tilde{L}})$ and $\tilde{\boldsymbol{g}}_{\tilde{j}\tilde{\ell}} = (\tilde{g}_{\tilde{j}1}, \tilde{g}_{\tilde{j}2}, \dots, \tilde{g}_{\tilde{j}\tilde{L}})$ can be formulated as follows [19].

$$\mathrm{d}(\tilde{\boldsymbol{g}}_{j\tilde{\ell}}, \tilde{\boldsymbol{g}}_{\tilde{j}\tilde{\ell}}) = \sqrt{\left(\tilde{\boldsymbol{g}}_{j\tilde{\ell}} - \tilde{\boldsymbol{g}}_{\tilde{j}\tilde{\ell}}\right)^T (\tilde{\boldsymbol{g}}_{j\tilde{\ell}} - \tilde{\boldsymbol{g}}_{\tilde{j}\tilde{\ell}})} \tag{27}$$

with $j, \tilde{j} = 1,2, \dots, q_{\tilde{k}}$ dan $j \neq \tilde{j}$

Euclidean distance is also used to group districts without using elliptical confidence regions, but from the principal coordinates of the district derived from JCA. With the vectors $\boldsymbol{f}_{i\ell} = (f_{i1}, f_{i2}, \dots, f_{i_L})$ and $\boldsymbol{f}_{\tilde{\imath}\ell} = (f_{\tilde{\imath}1}, f_{\tilde{\imath}2}, \dots, f_{\tilde{\imath}L})$, the Euclidean distance can be formulated as follows.

$$\mathrm{d}(\boldsymbol{f}_{i\ell}, \boldsymbol{f}_{\tilde{\imath}\ell}) = \sqrt{(\boldsymbol{f}_{i\ell} - \boldsymbol{f}_{\tilde{\imath}\ell})^T (\boldsymbol{f}_{i\ell} - \boldsymbol{f}_{\tilde{\imath}\ell})} \tag{28}$$

with $i, \tilde{\imath} = 1,2, \dots, q_1$ and $i \neq \tilde{\imath}$. Districts with the smallest Euclidean distance are combined by the geographic intersection.

3. MAIN RESULTS

Data with 8 characteristic variables are transformed into a contingency table. There are 8 contingency tables formed according to the number of characteristic variables. In carrying out correspondence analysis, characteristics variables must be dependent on the district variable. Therefore, a Chi-Square test is needed to see the dependence between characteristic variables on the district variable. The results of the Chi-Square Test can be seen in Table 1.

Table 1. Chi-Square Test

Category	X-Square	df	p-value
District vs X_1	53.27	60	0.7128
District vs X_2	130.78	90	0.0032
District vs X_3	151.59	120	0.0270
District vs X_4	341.01	210	2.739E-08
District vs X_5	289.4	150	6.564E-11
District vs X_6	74.673	30	1.119E-05
District vs X_7	267.14	150	1.306E-08
District vs X_8	69.827	60	0.1808

Based on the results of the Chi-Square test in Table 1, variables X_1 (Use of Defecation Facilities) and X_8 (Waste Processing) are not dependent on district variables. This is indicated by a p-value greater than 0.1. Therefore, these two variables do not need to be included in the analysis. After that, JCA was carried out on the 6 variables that were significantly dependent on the district variable. However, the results of the JCA on these 6 characteristic variables still cannot be depicted in a two-dimensional map, because the percentage of variation in two dimensions does not reach 70%

Table 2. Percentage of Variance of Joint Correspondence Analysis

Axis	**2**	**3**	**4**	**5**	**…**	**25**
λ_ℓ	0.05506	0.019486	0.01707	0.01437	…	0.000873
τ_D	44.9%	53.2%	60.4%	66.9%	…	100%

To solve the problem, recategorization using elliptical confidence regions is necessary. Elliptical confidence regions can show the contribution of dependence between variables. This contribution of dependence can be seen from the approximate p-values. If the approximate p-values are less than 0.1, then the category significantly contributes to the dependency structure and vice versa. Categories that are not significant will be combined with other categories that are similar based on the closest Euclidean distance. Euclidean distance is obtained from the principal coordinate results of correspondence analysis. There is an iteration process until the results of the elliptical confidence regions for the significant categories contribute to the dependency structure. The following is an example of the first elliptical confidence regions of variable X_2.

Table 3. Elliptical Confidence Regions Variable X_2

Category	**HL Axis**	**p-value**
$X_{2;1}$	1.0726	0.0757
$X_{2;2}$	0.5702	0.9975
$X_{2;3}$	1.1494	0.0538
$X_{2;4}$	0.7852	0.4276

Categories $X_{2;2}$ and $X_{2;4}$ have approximate p-values greater than 0.1, so the categories combined with similar categories. This similarity can be seen from the smallest Euclidian distance among the other categories.

Table 4. 1st Iteration Euclidean Distance

	$X_{2;1}$	$X_{2;2}$	$X_{2;3}$	$X_{2;4}$
$X_{2;1}$	0	1.644147	1.814068	1.733970
$X_{2;2}$	1.644147	0	1.405668	1.012703
$X_{2;3}$	1.814068	1.405668	0	1.356113
$X_{2;4}$	1.733970	1.012703	1.356113	0

The smallest distance from the first iteration of category X_2 (Final Waste Disposal Site) is between $X_{2;2}$ (Ground Hole) and $X_{2;4}$ (Tank or Waste Water Management Installation). These two categories combine to contribute to the dependency structure significantly. The p-values of elliptical confidence regions after combining categories are following.

Table 5. Elliptical Confidence Regions Variable X_2 After Recategorization

Category	**HL Axis**	**p-value**
$X_{2;1}$	1.7908	0.0837
$X_{2;2;4}$	0.1726	0.0837
$X_{2;3}$	1.2325	0.0837

After combining categories, the p-value for variable X_2 elliptical confidence region is less than 0.1. The iteration process is carried out until all categories have approximate p-values smaller than 0.1. The final category results for each variable from this iteration are as follows.

Table 6. New Categories based on Elliptical Confidence Regions

Variable	**Category**
Final Disposal of Faces (X_2)	Other ($X_{2;1}$)
	Soil Pits and Tanks/ Wastewater Management Plants ($X_{2;2,4}$)
	Rice fields/ ponds/ rivers/ lakes/ fields/ gardens ($X_{2;3}$)
Liquid Waste Drainage (X_3)	In Holes/Open Land, Drainage (Sewers/Gutters), and Rivers/Irrigation Channels ($X_{3;1,2,5}$)
	Other ($X_{3;3}$)
	Infiltration Hole ($X_{3;4}$)
Drinking Water Source (X_4)	Refilled Water, Metered Plumbing, Unmetered Plumbing, Wells, Drilled Wells or Pumps, Water Springs ($X_{4;1,3,4,5,6,7}$)

Variable	Category
	Branded Bottled Water $(X_{4;2})$
	River/ lake/ pond/ reservoir/ dam $(X_{4;8})$
Bathing Water Source (X_5)	Metered Plumbing, Drilled Wells or Pumps, River/ lake/ pond/ reservoir/dam $(X_{5;1,5,6})$
	Unmetered Plumbing $(X_{5;2})$
	Water Springs $(X_{5;3})$
Family Waste Disposal (X_6)	Trash bin, then hauled away $(X_{6;1})$
	In the pit or burned $(X_{6;2})$
Forest Area Function (X_7)	Not Forest Area Function, outside Forest Area, Conservation, Production $(X_{7;1,2,3,6})$
	Protection Forest $(X_{7;4})$
	No Forest Area $(X_{7;5})$

Elliptical confidence regions should not be used to group objects because they will result in 6 different outcomes based on categorical variables. Therefore, the treatment for district variables should be different from that of characteristics variables. District variables should only be grouped based on the shortest Euclidean distance from the JCA principal coordinates. Based on these groupings, 5 district groups were formed.

Table 7. District Group

Group	District
1	Arjasari $(D_{1:1})$, Banjaran $(D_{1:3})$, Ibun $(D_{1:15})$, Pacet $(D_{1:23})$, Kertasari $(D_{1:17})$, Pangalengan $(D_{1:25})$, Cimaung $(D_{1:10})$, Cangkuang $(D_{1:5})$, and Pasirjambu $(D_{1:27})$
2	Katapang $(D_{2:16})$, Kutawaringin $(D_{2:18})$, Soreang $(D_{2:31})$, Margaasih $(D_{2:20})$, Ciwidey $(D_{2:13})$, Baleendah $(D_{2:2})$, Pameungpeuk $(D_{2:24})$, Bojongsoang $(D_{2:4})$, and Dayeuhkolot $(D_{2:14})$
3	Cikancung $(D_{3:7})$, Nagreg $(D_{3:22})$, Cicalengka $(D_{3:6})$, Paseh $(D_{3:26})$, Ciparay $(D_{3:12})$, Majalaya $(D_{3:19})$, Rancaekek $(D_{3:29})$, Solokan Jeruk $(D_{3:30})$, Cileunyi $(D_{3:9})$, Cilengkrang $(D_{3:8})$, and Cimenyan $(D_{3:11})$
4	Margahayu $(D_{4:21})$
5	Rancabali $(D_{5:28})$

The object groups in Table 6 are the result of combined object variables based on proximity. The proximity is obtained by calculating the Euclidean distance from the principal coordinates of the

JCA results. The combined process was performed 27 times until the variance percentage in two dimensions exceeded 70.1%. In the 27th iteration, JCA explains 70.1% variance in two dimensions. Therefore, the analysis results can be represented on a two-dimensional map.

Figure 1. Two-dimensional Map

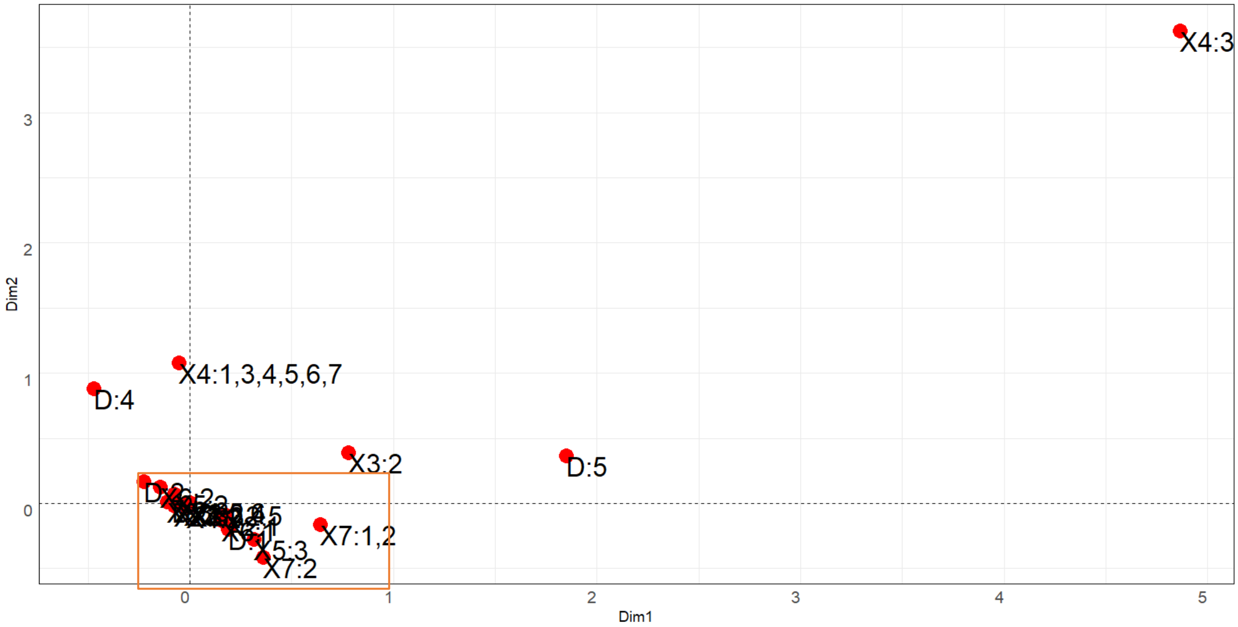
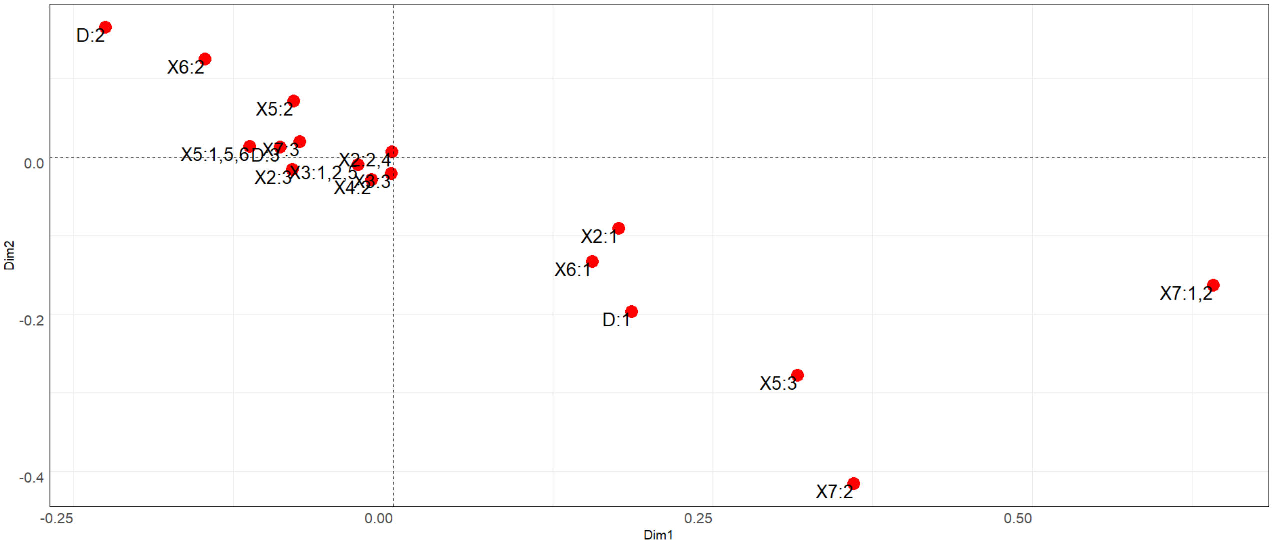


Figure 2. Zoom in on a Two-dimensional Map



Based on the two-dimensional JCA map, information can be obtained that district group 1 has environmental indicators that are interdependent, that is protected forest area function, the source of bathing water is from springs, places where rubbish is thrown in pits or burned, and places where waste is disposed of in addition to the other categories. District group 2 has environmental indicators that are interdependent, namely where rubbish is thrown in the rubbish bin, and then transported. District group 3 consists of districts with environmental characteristics that are related to each other, there are functions of production forest areas, conservation, outside the forest area, and not a function of the forest area. Apart from that, there are sources of bathing water that come from taps without meters, taps with meters, drilled wells or pumps, rivers/ lakes/ ponds/ reservoirs/ situs/ dams, and wells. Places for liquid waste disposal in Drainage (Sewers), Rivers/ Irrigation Channels, Holes/Open Land, and absorption pits. The majority of drinking water sources used come from refilled water, bore wells or pumps, wells, springs, plumbing with meters, and plumbing without meters. District group 3 still throws waste into rice fields/ponds/rivers/lakes/fields/gardens. Margahayu district has different indicators from other districts, the source of drinking water comes from branded bottled water. Rancabali district does not have the same indicators as other districts, indicated by different liquid waste disposal sites (other categories).

4. CONCLUSION

The method for recategorization with a large number of categories should be based on the results of the elliptical confidence regions. Elliptical confidence regions can see the contribution of the dependence between two qualitative variables. If a category does not contribute significantly to the dependency structure, then the category is combined with another category that has the closest Euclidean distance. Before calculating elliptical confidence regions, a chi-square test is necessary to determine which characteristic variables are dependent on the districts. Of the 8 characteristics, 6 characteristics are dependent on the districts. After recategorization using elliptical confidence regions, the percentage of variance obtained still does not reach the minimum criteria. Therefore, a grouping of district variable categories was carried out based on the closest distance to Euclidean distance based on the principal coordinates from JCA until the percentage of variance in the two dimensions had reached 70%. Based on the results of the analysis obtained, the method using JCA with elliptical confidence regions can be used to obtain a variance percentage of 70.1% in two dimensions. The two-dimensional map shows that several districts in Bandung Regency have not implemented a healthy and clean lifestyle. The government can evaluate environmental indicators so that the EQI value of Bandung Regency increases. The districts of Cikancung, Nagreg, Cicalengka, Paseh, Ciparay, Majalaya, Rancaekek, Solokan Jeruk, Cileunyi, Cilengkrang, and Cimenyan still use rivers/lakes/ponds/reservoirs/dams as a source for bathing, as a place to dispose of feces, and liquid waste disposal site.

Elliptical confidence regions can be used to perform recategorization based on the dependence between qualitative variables in joint correspondence analysis. The method of recategorization can be applied to other research, as long as the object being studied is a region and its characteristics are qualitative variables with more than two categories. Future research should consider computing the relationship between characteristics and multi-way correspondence analysis.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support provided by Unpad Lecturer Competency Research (RKDU) Padjadjaran University 2023 Number 1549/UN6.3.1/PT.00/2023. We also to thank Dr. I Gede Nyoman Mindra Jaya who has helped to get a better analysis.

CONFLICT OF INTERESTS

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

REFERENCES

[1] M. Greenacre, Correspondence analysis in practice, Chapman and Hall/CRC, Boca Raton, 2007. https://doi.org/10.1201/9781420011234.

[2] 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), 1036-1045. https://doi.org/10.4314/ahs.v14i4.35.

[3] J.F. Hair Jr, W.C. Black, B.J. Babin, et al. Multivariate data analysis, 7th ed, Prentice Hall, Upper Saddle River, 2010.

[4] M. Greenacre, J. Blasius, eds., Multiple correspondence analysis and related methods, Chapman and Hall/CRC, 2006. https://doi.org/10.1201/9781420011319.

[5] N. Sourial, C. Wolfson, B. Zhu, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables, J. Clinic. Epidemiol. 63 (2010), 638-646. https://doi.org/10.1016/j.jclinepi.2009.08.008.

[6] K. Adachi, Correspondence analysis, multiple correspondence analysis, and joint correspondence analysis, Japan. Psychol. Rev. 46 (2003), 547-563.

[7] R.J. Boik, An efficient algorithm for joint correspondence analysis, Psychometrika. 61 (1996), 255-269. https://doi.org/10.1007/bf02294338.

[8] A. Giusti, G. Ritter, M. Vichi, eds., Classification and data mining, Springer, Berlin, Heidelberg, 2013. https://doi.org/10.1007/978-3-642-28894-4.

[9] J.K. Vermunt, C.J. Anderson, Joint correspondence analysis (JCA) by maximum likelihood, Methodology. 1 (2005), 18-26. https://doi.org/10.1027/1614-1881.1.1.18.

[10] I. Ginanjar, I. Nurhuda, N. Sunengsih, et al. 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.

[11] F. Sholihah, I. Ginanjar, M. IAENG, et al. A hybrid singly ordered correspondence with correlation approach to analyzing the relationship between age groups and happiness level in Indonesia, IAENG Int. J. Appl. Math. 53 (2023), 19.

[12] A. Agresti, An introduction to categorical data analysis, Wiley, Hoboken, 2007.

[13] J. de Leeuw, P. Mair, Simple and canonical correspondence analysis using the R package anacor, J. Stat. Softw. 31 (2009), 1-18. https://doi.org/10.18637/jss.v031.i05.

[14] O. Nenadic, 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.

[15] I. Ginanjar, U.S. Pasaribu, A. Barra, Simplification of correspondence analysis for more precise calculation which one qualitative variables is two categorical data, ARPN J. Eng. Appl. Sci. 11 (2016), 1983-1991.

[16] G. Li, S. Lu, H. Zhang, et al. 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.

[17] 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.

[18] E.J. Beh, R. Lombardo, Confidence regions and approximatep-values for classical and non symmetric correspondence analysis, Commun. Stat. - Theory Methods 44 (2014), 95-114. https://doi.org/10.1080/03610926.2013.768665.

[19] A.C. Rencher, W.F. Christensen, Methods of multivariate analysis, Wiley, Hoboken, 2012. https://doi.org/10.1002/9781118391686.