A Hybrid Singly Ordered Correspondence with Correlation Approach to Analyzing the Relationship between Age Groups and Happiness Level in Indonesia

Fajriatus Sholihah, Irlandia Ginanjar, *Member IAENG*, Yuyun Hidayat

***Abstract*— When subpopulation objects are to be analyzed, qualitative data is frequently transformed into a contingency table. The categories of rows and columns in the contingency table are sequentially treated as objects and variables. In addition, the object frequently retrieves additional information with continuous variables. The novelty of this study was that it examined associations between discrete and continuous variables in two stages. The hybrid Singly Ordered Correspondence Analysis (SOCA) method was used to identify associations between nominal-scale objects and ordinal-scale discrete variables in the form of contingency tables. Correlation values were used to identify associations between objects and continuous variables. The SOCA method determined the principal coordinates of discrete variables. Concurrently, the correlation was used to determine the vector coordinates of the continuous variables. Confidence region and p-value calculations were performed to determine the contribution of a category to the structure of the association between two categorical variables with nominal and ordinal scales. The method was applied to population-based data on the level of happiness across age groups. The case study results indicated that all nominal and ordinal categories contributed significantly to the association between age group and province variables in Indonesia. The association results indicated that 27 provinces were positively associated with the age bracket of 17 to 41 years, with these provinces being associated with 13 indicators of the happiness index. The remaining seven provinces were linked to age groups of at least 41 years old, and six happiness index indicators were linked to these provinces.**

***Index Terms*—singly ordered correspondence analysis, confidence regions and approximate p-values, correlation, happiness index**

I. Introduction

In the current era of big data, various types of data exist. Qualitative data is frequently transformed into contingency tables when analyzing objects that belong to subpopulations. The categories of rows and columns in the contingency table are sequentially treated as objects and variables. Therefore, the data consists of discrete object variables [1]. Moreover, the object frequently acquires additional data from continuous variables. Based on this, the researcher can determine the relationship between objects and variables and between discrete and continuous data types. The dependence between two categorical variables is the association of discrete variables [2]. The two variables are independent if the product of their marginal probabilities equals their joint probability. A linear relationship between two continuous variables defines the association of continuous variables.

The selection of statistical methods to identify associations between objects, between variables, and between objects and variables in Rencher [2] depends on the data type employed. Associations in continuous data can be identified using correlation. As with discrete data in contingency tables, the chi-squared test can be used to identify associations. The chi-squared test's association between two categorical variables is their dependence on one another. The chi-square test was unable to identify associations between the categories. The relationship between these categories represents a dependency between two categories of categorical variables. Therefore, additional analysis using correspondence analysis is required to determine this. Correspondence analysis is used to determine relationships between categories of two categorical variables, where row categories represent objects and column categories represent their characteristics. The standard residual matrix, measuring the disparity between observed and expected data, reflects the relationship between these categories. This residual value is obtained by subtracting the observed proportion from the expected proportion. The principal component analysis (PCA) biplot is a statistical method that can identify associations between objects, between variables, and between objects and variables. However, in a PCA biplot, associations are determined using continuous variables.

Based on the described methods, an approach is required to analyze associations between objects, between variables, and between objects and mixed variables (discrete and continuous variables). The Hybrid Correspondence Analysis and Correlation (HCAC) method [3] is used to analyze discrete variables on a nominal scale, and the correlation method is used to analyze continuous variables. However, if

Manuscript received December 6, 2022; revised June 9, 2023.

This work was supported by the Unpad Lecturer Competency Research (RKDU) Universitas Padjadjaran 2023 Number 1549/UN6.3.1/PT.00/2023.

F. Sholihah is a graduate student of Applied Statistics Study Program, Padjadjaran University, West Java, 455363, Indonesia. e-mail: fajriatus16001@mail.unpad.ac.id.

I. Ginanjar is an associate professor at Statistics Department, Padjadjaran University, West Java, 455363, Indonesia. e-mail: irlandia@unpad.ac.id (corresponding author)

Y. Hidayat. is a professor at Statistics Department, Padjadjaran University, West Java, 455363, Indonesia. e-mail: yuyun.hidayat@unpad.ac.id.

the traditional approach to correspondence analysis is applied to ordinal scale variables, there will be missing information. Consequently, it is necessary to employ hybrid decomposition (HD) for standard residual matrix decomposition, a new method that can be applied to nominal and ordinal scale variables. This HD is a combination of Singular Value Decomposition (SVD) and Bivariate Moment Decomposition (BMD), which is utilized in the Singly Ordered Correspondence Analysis (SOCA) method [4]. The novelty of this study was that it was conducted to analyze the associations between discrete and continuous data in stages using the hybrid SOCA with correlation values to obtain the principal coordinates of the discrete variables and the vector coordinates of the continuous variable. Initially, discrete data was analyzed using the SOCA technique. In addition, the results of the SOCA analysis were analyzed using the correlation method with continuous data. Also, confidence regions and approximate p-values are calculated for each category of variables with nominal and ordinal scales to determine how much each category affects the relationship structure between the two categorical variables.

In previous research, a study was carried out on Indonesia, such as COVID-19, which occurred in West Java [5]. The study aimed to predict the final size of COVID-19 in West Java Province. There is also a study on the spatial prediction of malaria risk that occurs in Bandung City [6]. Meanwhile, Ginanjar et al. [7] conducted a study on expenditure per capita at the sub-district level in Jambi Province. In this research, a study was conducted on happiness in Indonesia. The utilized datasets consist of happiness index indicators and age categories based on each province's population. The index tries to evaluate individuals' level of satisfaction with all domains of human life that are deemed essential by considering sentiments and the meaning of life [8]. The findings suggest that the happiness of the population influences the success of development and social development in the community. Other studies have found that age impacts happiness and that each age group has a distinct definition of happiness [9–13]. Similarly, each location interprets happiness [14–17].

We are unable to determine the relationship between the happiness indicator for each province and age group using the BPS-calculated data. Consequently, this research will uncover relationships between BPS-compiled happiness indices for each province and age group. The findings of the study can be used to calculate the level of happiness in each province by age group. Consequently, the government can apply this knowledge when formulating policies. The data is collected in order to improve the welfare of the people in each province.

II. RESEARCH METHODOLOGY

This SOCA approach was used in this study to determine associations between discrete data arranged in contingency tables and ordinal categorical column variables. The provincial population by age group was used as the data source. The correlation value was applied to continuous data to determine relationships, with pleasure level serving as the data. In addition, the computations were performed using R software using the packages CAvariants, factoextra, and zoom.

A. Data Sources

This study examined happiness levels and the total population by age group using secondary data from 34 provinces in Indonesia. This data was retrieved from bps.go.id. There were both discrete and continuous variables used. The concrete variables consisted of row categories as nominal scale objects and column categories as ordinal scale variables. Indonesia had 34 provinces, which served as the nominal categorical variable. Age groups were used as ordinal categorical variables. There are 13 age groups: 17 to 21 years, 22 to 26 years, 27 to 31 years, 32 to 36 years, 37 to 41 years, 42 to 46 years, 47 to 51 years, 52 to 56 years, 57 to 61 years, 62 to 66 years, 67 to 71 years, 72 to 76 years, and 77 or older.

The continuous variable combined the scores for each of the 19 elements comprising the happiness index. This included three life dimensions: the dimension of life satisfaction, with sub-dimensions of personal life satisfaction and social life satisfaction; the dimension of feelings; and the dimension of life's meaning. The 19 indicators of the happiness index included education and skills (X1), occupation/business/main activity (X2), household income (X3), health (X4), housing conditions and home facilities (X5), family harmony (X6), availability of free time (X7), social relations (X8), environmental conditions (X9), security conditions (X10), feelings of happiness/cheerful/joy (X11), feeling not worried/anxious (X12), feeling not depressed (X13), independence (X14), environmental mastery (X15), self-development (X16), positive relationships with others (X17), life goals (X18), and self-acceptance (X19).

B. Chi-Squared Test

A chi-squared test was conducted on the data in the contingency table to analyze the association between two categorical variables. First, construct a contingency table with a sample size of n. This table was a cross-tabulation of two categorical variables. The first variable consisted of I row categories, while the second variable included J column categories; see Table I. Suppose n_{ij} is the joint frequency of the number of individuals in row i and column j, for $i = 1, 2,..., I$ and $j = 1, 2,..., J$.

The hypotheses in the chi-squared test are [15]:

H_0: $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$

H_1: $\pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$

with test statistics

$$\chi^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}} \quad (1)$$

$\hat{\pi}_{ij} = p_{ij} = n_{ij}/n$; $\hat{\pi}_{i\bullet} = p_{i\bullet} = n_{i\bullet}/n$; $\hat{\pi}_{\bullet j} = p_{\bullet j} = n_{\bullet j}/n$.

Where p_{ij} is the joint probability of the i^{th} row category and the j^{th} column category, $p_{i\bullet}$ is the marginal probability of the i^{th} row category, and $p_{\bullet j}$ is the marginal probability of the j^{th} column category. With significance level α, reject H_0 if $\chi^2 \geq \chi^2_{(df;\alpha)}$, $df = (I-1)(J-1)$ for p-value $\leq \alpha$.

TABLE I
THE GENERIC FORM OF A CONTINGENCY TABLE

Row Category	Column Category				
	Column 1	*Column* 2	⋯	*Column J*	*Total*
Row 1	n_{11}	n_{12}	⋯	n_{1J}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	⋯	n_{2J}	$n_{2\bullet}$
⋮	⋮	⋮	⋱	⋮	⋮
Row I	n_{I1}	n_{I2}	⋯	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	⋯	$n_{\bullet J}$	N

C. Singly Ordered Correspondence Analysis

SOCA is a correspondence analysis method that employs the two-way contingency table with nominal and ordinal categorical variables [18]. The decomposition algorithm used in SOCA is hybrid decomposition (HD), which is a combination of singular value decomposition (SVD) and bivariate moment decomposition (BMD). Because it considers the singular vector of SVD and the orthogonal polynomial of BMD, this decomposition considers the structure of nominal and ordinal categorical variables. The correspondence matrix is calculated as the first step in SOCA using the following:

$$\mathbf{P} = \left(p_{ij}\right)$$

Next, define the vectors $\mathbf{r}^t = (p_{1\bullet} \;\; p_{2\bullet} \;\; \cdots \;\; p_{I\bullet})$ and $\mathbf{c}^t = (p_{\bullet 1} \;\; p_{\bullet 2} \;\; \cdots \;\; p_{\bullet J})$, then calculate $\mathbf{R} = \text{diag}(\mathbf{r})$ and $\mathbf{C} = \text{diag}(\mathbf{c})$ [16]. The next step is calculating a standard residual matrix **S** representing the contingency table's associations. Matrix **S** is the residual, which measures the difference between the observed and expected data. Therefore, this residual value is obtained by reducing the observed proportion to the expected proportion. For example, the standard residual matrix [2] is calculated as follows:

$$\mathbf{S} = \left(s_{ij}\right) = \mathbf{R}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^t)\mathbf{C}^{-1/2} \tag{2}$$

with

$$s_{ij} = \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})}{\sqrt{p_{i\bullet}p_{\bullet j}}} = \frac{n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}}{\sqrt{n_{i\bullet}n_{\bullet j}}}$$

The matrix **S** decomposes using HD to get the principal coordinates orthogonal to each other. The hybrid decomposition of the standard residual matrix can be written as follows [20]:

$$\mathbf{S} = \mathbf{U}\mathbf{Z}\widetilde{\boldsymbol{\mathcal{B}}}^t \tag{3}$$

were

- $\mathbf{U}^t\mathbf{R}\mathbf{U} = \widetilde{\boldsymbol{\mathcal{B}}}^t\mathbf{C}\widetilde{\boldsymbol{\mathcal{B}}} = \mathbf{I}$
- **U** is the left singular matrix $(I \times J)$ from SVD. **U** is calculated using $|\mathbf{S}^t\mathbf{S} - \lambda\mathbf{I}| = 0$ and $(\mathbf{S}^t\mathbf{S} - \lambda\mathbf{I})\mathbf{u} = 0$. **u** is an eigenvector which is an element of the **U**.
- $\widetilde{\boldsymbol{\mathcal{B}}}$ is the weighted of column polynomial orthogonal matrix $(J \times (J-1))$ from BMD, calculated by $\widetilde{\boldsymbol{\mathcal{B}}} = \mathbf{C}^{1/2}\boldsymbol{\mathcal{B}}$. Elements of the $\boldsymbol{\mathcal{B}}$, β_{jv}; $j = 1,2,\ldots,J$; $v = 1,2,\ldots,J-1$. Value of $\beta_{j,-1} = 0$, $\beta_{j0} = 1$, and for $v = 1,2,\ldots,j-1$ calculated using

$$\beta_{jv} = A_v\left((j - B_v)\beta_{j,v-1} - C_v\beta_{j,v-2}\right)$$

were

$$B_v = \sum_{j=1}^{J} jp_{\bullet j}\beta_{j,v-1}^2$$

$$C_v = \sum_{j=1}^{J} jp_{\bullet j}\beta_{j,v-1}^2\beta_{j,v-2}^2$$

$$A_v = \left(\sum_{j=1}^{J} j^2p_{\bullet j}\beta_{j,v-1}^2 - B_v^2 - C_v^2\right)^{-1/2}$$

- **Z** is a general correlation matrix $\left(M \times (J-1)\right)$; $M = \min(I, J)$, calculated using

$$\mathbf{Z} = \mathbf{U}^t\mathbf{S}\widetilde{\boldsymbol{\mathcal{B}}}$$

The components of the decomposed **S** are used to calculate the principal coordinates. The matrix of the principal coordinates of the row categories can be calculated as follows [20]:

$$\mathbf{F} = \mathbf{R}^{-1/2}\mathbf{U}\mathbf{Z} = (f_{id}) \tag{4}$$

The principal coordinates of the column categories are calculated using

$$\mathbf{G} = \mathbf{C}^{-1/2}\widetilde{\boldsymbol{\mathcal{B}}} = \left(g_{jd}\right) \tag{5}$$

The quality of the mapping for every dimension $(d = 1,2,\cdots M)$ of the row $\left(\lambda_d^{(r)}\right)$ and column $\left(\lambda_d^{(c)}\right)$ categorical variables can be seen based on the diagonal of the squared value of the general correlation matrix, expressed by

$$\boldsymbol{\lambda}^{(r)} = diag(\mathbf{Z}\mathbf{Z}^t) = \left(\lambda_d^{(r)}\right) \tag{6}$$

and

$$\boldsymbol{\lambda}^{(c)} = diag(\mathbf{Z}^t\mathbf{Z}) = \left(\lambda_d^{(c)}\right). \tag{7}$$

The following equation can express the variance coverage for a D-dimensional map of the row and column categorical variables:

$$\phi_D^{(r)} = \frac{\sum_{d=1}^{D}\lambda_d^{(r)}}{trace(\mathbf{Z}\mathbf{Z}^t)},\;\; \phi_D^{(c)} = \frac{\sum_{d=1}^{D}\lambda_d^{(c)}}{trace(\mathbf{Z}^t\mathbf{Z})}. \tag{8}$$

The quality of the representation of the D-dimensional map is

$$\phi_D = min\left(\phi_D^{(r)}, \phi_D^{(c)}\right). \tag{9}$$

The correspondence map is good if the cumulative variance reaches 70% or more [18].

D. Confidence Regions and Approximate P-Values

Confidence regions and approximate p-values can be used to examine each category's contribution. This calculation is conducted because the existence of these categories will affect decision-making. Categories that do not contribute to the structure of association between these two categorical variables cannot be considered in decision-making. Therefore, a recategorization is needed [21]. Confidence regions for the i^{th} nominal categorical variable are calculated based on confidence ellipses. For example, the confidence ellipses of the i^{th} nominal category can be constructed with the length of the semi-axis along the m^{th} principal axis as follows [21]:

$$x_{im(\alpha)} = \sqrt{\lambda_m \frac{\chi^2_{(df=(I-1)(J-1);\alpha)}}{\chi^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=D+1}^{M} a_{im}^2\right)}, \tag{10}$$

for $m = 1,2,\cdots D$, $p_{i\bullet}$ is the marginal probability of the i^{th} province category, χ^2_α is the α^{th} quantile of the chi-square with degrees of freedom $(I-1)(J-1)$, and $\chi^2 = n\varphi^2$ where φ^2 is the total inertia, a_{im} is the m^{th} dimension of the row singular vector associated with the i^{th} row profile, and the eigenvalue of the m^{th} λ_m dimension row category variable.

The length of confidence regions for ordinal categorical variables has the same value for each category and dimension. This is due to the long calculation of $100(1-\alpha)\%$ confidence regions based on confidence circles. The following equation [22] calculates the length of the confidence regions for each ordinal categorical variable:

$$c_{(\alpha)} = 2\sqrt{J\frac{\chi^2_{(df=J-1;\alpha)}}{n}}, \quad (10)$$

where n is the total population, J is the number of categories in the ordinal categorical variable, and $\chi^2_{(J-1);\alpha}$ is the α^{th} quantile of the chi-square with degrees of freedom $(J-1)$.

Furthermore, to identify the contribution of the i^{th} row and j^{th} column categories, the following hypotheses were made:

H_0: $f_i = 0$ (The i^{th} row category does not contribute.)

H_1: $f_i \neq 0$ (The i^{th} row category contributes.)

and

H_0: $g_j = 0$ (The j^{th} column category does not contribute.)

H_1: $g_j \neq 0$ (The j^{th} column category contributes.)

The approximate p-value for the i^{th} nominal categorical variable was calculated by the following equation:

$$(p-value)_{i,D} = P\left\{\chi^2 > n\varphi^2\left(\frac{1}{p_{i\bullet}} - \sum_{m=D+1}^{M}\left(\frac{f_{im}}{\sqrt{\lambda_m}}\right)^2\right)^{-1}\sum_{m=1}^{D}\left(\frac{f_{im}}{\sqrt{\lambda_m}}\right)^2\right\}. \quad (11)$$

The approximate p-value for the j^{th} ordinal categorical variable was calculated by the following equation:

$$(p-value)_{j,D} = P\{\chi^2 > n\sum_{m=1}^{D} p_{\bullet j} g_{jm}^2\}. \quad (12)$$

Approximate p-value less than the specified significance level provide evidence that the category contributed to the association between the two categorical variables.

E. Correlation Values

A correlation was used to get the vector coordinates of a continuous variable. The method chosen was the Pearson correlation. The correlation was calculated between Q continuous variables and M principal coordinates of the row categories. The following is the correlation formula between two vectors for $q = 1,2,\cdots,Q$ and $d = 1,2,\cdots,M$ [2]:

$$\rho_{qd} = corr(X_q, \boldsymbol{f}_d) = \frac{\sum_{i=1}^{I}(x_{iq}-\bar{x}_q)(f_{id}-\bar{f}_d)}{\sqrt{\sum_{i=1}^{I}(x_{iq}-\bar{x}_q)^2}\sqrt{\sum_{i=1}^{I}(f_{id}-\bar{f}_d)^2}}, \quad (13)$$

where x_{iq} is the value for the i^{th} row category and the q^{th} continuous variables, $\bar{x}_q$ is the average value of the q^{th} continuous variables, and $\bar{f}_d$ is the average value of the d^{th} principal coordinates.

F. Hybrid SOCA with Correlation Values

The correlation value is a vector coordinate for a continuous variable; it can be written as follows [3]:

$$\boldsymbol{\Pi} = (\rho_{qd}).$$

Furthermore, calculations were carried out to obtain the vector coordinates corresponding to the principal coordinates of the SOCA. They were done on vector coordinates with a significant correlation. This was because decision-making based on a non-significant correlation was not representative. λ_d^{α} became the multiplier for the continuous variable where $0 \leq \alpha \leq 1$. Thus, the vector coordinates of the continuous variable, were calculated using the following formula:

$$\boldsymbol{\Psi} = \boldsymbol{\Pi}\, diag\left(\lambda_d^{\frac{1}{2}\alpha}\right) = (\psi_{qd}). \quad (14)$$

G. Steps of Data Analysis

1. Analyzing discrete data
   - Performing the chi-square test using Equation (1)
   - Calculating the matrix **S** using Equation (2)
   - Decomposition of **S** using HD based on Equation (3)
   - Calculating the principal coordinates of row and column categories using Equations (4) and (5)
   - Calculating inertia for row and column categories using Equations (6) and (7)
   - Calculating confidence regions using Equations (9) and (10) and approximate p-values using Equations (11) and (12)
2. Analyzing continuous data
   - Calculating the correlation between continuous variables and the principal coordinates of the row category using Equation (13)
   - Calculating the vector coordinates corresponding to the principal coordinates of the SOCA using Equation (14)
3. Building a correspondence map containing the principal coordinates of the row category, the column category's principal coordinates, and the continuous variable's vector coordinates

III. RESULTS AND DISCUSSION

This section describes the data processing results and discusses the analysis of discrete and continuous data associations using hybrid SOCA with correlation values. The obtained results are the principal coordinates of the discrete variables and the vector coordinates of the continuous variables for analyzing age group associations for each province and happiness indicator associations for each province and age group.

A. Independence Test of Two Category Variables

The data in the contingency table used was the population by age group for each province. The Chi-squared test was used to compare the variables of province and age group. The calculation of Equation (1) using R software obtained the value of $\chi^2 = 2426979$ with $df = 396$. In addition, an approximate p-value less than 2.2×10^{-16} was also obtained. The approximate p-value is smaller than $\alpha = 0.05$. Thus, it can be concluded that H_0 is rejected. It means there is an association between age group variables and provinces in Indonesia. Because the chi-square test results did not reveal any information about the relationship between the categories, SOCA was used for further analysis.

B. Principal Coordinates of Discrete Variables

Furthermore, the analysis was continued with SOCA to identify associations between categories. SOCA was chosen because the categorical variables used were nominal and ordinal. The analysis results were the principal coordinates of the row and column categories. Tables II and III show the inertia calculations for row and column categories using R software and Equations (6), (7), and (8).

TABLE II
MAPPING QUALITY FOR PROVINCE CATEGORIES

d	$\lambda_d^{(r)}$	$\phi_D^{(r)}$	d	$\lambda_d^{(r)}$	$\phi_D^{(r)}$
1	1.15×10^{-2}	89.25	7	5.46×10^{-6}	99.94
2	1.01×10^{-3}	97.15	8	3.86×10^{-6}	99.97
3	2.58×10^{-4}	99.16	9	1.46×10^{-6}	99.98
4	6.55×10^{-5}	99.67	10	1.05×10^{-6}	99.99
5	1.78×10^{-5}	99.81	11	5.24×10^{-7}	100
6	1.14×10^{-5}	99.90	12	3.81×10^{-7}	100

TABLE III
MAPPING QUALITY FOR AGE GROUP CATEGORIES

d	$\lambda_d^{(c)}$	$\phi_D^{(c)}$	d	$\lambda_d^{(c)}$	$\phi_D^{(c)}$
1	1.05×10^{-2}	81.58	7	9.30×10^{-6}	99.78
2	1.94×10^{-3}	96.72	8	3.82×10^{-6}	99.81
3	2.69×10^{-4}	98.81	9	1.37×10^{-5}	99.91
4	6.67×10^{-5}	99.33	10	5.34×10^{-6}	99.95
5	3.03×10^{-5}	99.57	11	2.62×10^{-6}	99.97
6	1.75×10^{-5}	99.70	12	3.23×10^{-6}	100

According to Tables II and III, the age group categories had the lowest variance values. Thus, the map to be built was based on Table III. Table III shows that the first dimension had a variance coverage of 81.58%. Based on this, the one-dimensional map was good at describing the relationship between the provincial category and the age group categories. The term confidence regions is replaced by confidence intervals due to the one-dimensional nature of the representation. Furthermore, the principal coordinates of the row and column categories were also obtained. Then the calculation row coordinates from equation (4) and each category's contribution from equations (9) and (11) are presented in Table IV. While the results of the calculation of the column coordinate from equation (5) and each category's contribution using equations (10) and (12) are presented in Table V.

TABLE IV
ROW COORDINATES, CONFIDENCE INTERVALS (CI), AND P-VALUE OF ROW CATEGORY

Province (i)	f_{i1}	CI	P-value
Aceh	-0.1300	0.0105	<0.0001
North Sumatra	-0.1026	0.0063	<0.0001
West Sumatra	-0.0504	0.0103	<0.0001
Riau	-0.1587	0.0094	<0.0001
Jambi	-0.0425	0.0126	<0.0001
South Sumatra	-0.0851	0.0082	<0.0001
Bengkulu	-0.0428	0.0169	<0.0001
Lampung	-0.0271	0.0081	<0.0001
Bangka Belitung Island	-0.0531	0.0198	<0.0001
Riau Island	-0.2009	0.0166	<0.0001
Jakarta	-0.0900	0.0070	<0.0001
West Java	-0.0323	0.0033	<0.0001
Central Java	0.1260	0.0038	<0.0001
DI Yogyakarta	0.1831	0.0116	<0.0001
East Java	0.1405	0.0035	<0.0001
Banten	-0.0987	0.0067	<0.0001
Bali	0.0733	0.0111	<0.0001
West Nusa Tenggara	-0.0309	0.0106	<0.0001
East Nusa Tenggara	-0.0623	0.0105	<0.0001
West Kalimantan	-0.0904	0.0106	<0.0001
Central Kalimantan	-0.1042	0.0147	<0.0001
South Kalimantan	0.0009	0.0116	<0.0001
East Kalimantan	-0.0795	0.0125	<0.0001
North Kalimantan	-0.0910	0.0293	<0.0001
North Sulawesi	0.0846	0.0147	<0.0001
Central Sulawesi	-0.0527	0.0138	<0.0001
South Sulawesi	0.0142	0.0078	<0.0001
Southeast Sulawesi	-0.1195	0.0151	<0.0001
Gorontalo	-0.0461	0.0218	<0.0001
West Sulawesi	-0.0948	0.0209	<0.0001
Maluku	-0.1144	0.0182	<0.0001
North Maluku	-0.1254	0.0219	<0.0001
West Papua	-0.2121	0.0255	<0.0001
Papua	-0.1954	0.0133	<0.0001

TABLE V
COLUMN COORDINATES, CONFIDENCE INTERVAL, AND P-VALUE OF COLUMN CATEGORY

Age (j) (in years)	g_{j1}	*P-value*	Age (j) (in years)	g_{j1}	*P-value*
17 to 21	-1.3519	<0.0001	52 to 56	0.8380	<0.0001
22 to 26	-1.0391	<0.0001	57 to 61	1.1508	<0.0001
27 to 31	-0.7262	<0.0001	62 to 66	1.4636	<0.0001
32 to 36	-0.4134	<0.0001	67 to 71	1.7765	<0.0001
37 to 41	-0.1005	<0.0001	72 to 76	2.0893	<0.0001
42 to 46	0.2123	<0.0001	77 or more	2.4021	<0.0001
47 to 51	0.5251	<0.0001			

Note: the length of convidence intervals for every age group is 0.0024

All p-values in Tables IV and V are zero. This means that age group categories and Indonesia's 34 provinces all contributed to the association between age group and province variables. This implies that all categories should be considered as the basis for policymaking.

C. Vector Coordinates of Continuous Variables

Furthermore, analysis was carried out on continuous data to obtain the vector coordinates of the continuous variable. The vector coordinates of the continuous variable were obtained from the correlation between the principal coordinates of the row categories (F) from Table IV and the continuous variables (X). The following are the results of the calculation of vector coordinates for the happiness index indicator using Equation (14), which correspond to the principal coordinates of the SOCA with a value of $\alpha = 0$.

Based on Table VI, it can be identified that the principal coordinates of the row category and the vector coordinates of the continuous variable had negative values. This means that the continuous variable had an association with the row category. On the other hand, the principal coordinate of a row category with a positive value was associated with a continuous variable with a positive vector coordinate.

TABLE VI
VECTOR COORDINATES OF CONTINUOUS VARIABLE

Variable	ψ_{q1}	Variable	ψ_{q1}
X1	0.0997	X11	−0.0391
X2	−0.0634	X12	0.4595
X3	−0.0245	X13	0.4634
X4	−0.0365	X14	−0.2054
X5	0.1137	X15	−0.1864
X6	−0.1401	X16	−0.1287
X7	−0.3588	X17	−0.2113
X8	−0.0211	X18	−0.0248
X9	0.2078	X19	−0.0945
X10	0.2252		

D. Map of Hybrid Singly Ordered Correspondence Analysis with Correlation Values

Mapping the row coordinates f_{i1} , column coordinates g_{j1}, and vector coordinates ψ_{q1} from Tables IV, V, and VI results in Figure 1. Based on the variance values obtained in Table III, the map is formed in one dimension. Meanwhile, Figures 2 and 3 are enlarged maps of Figure 1.

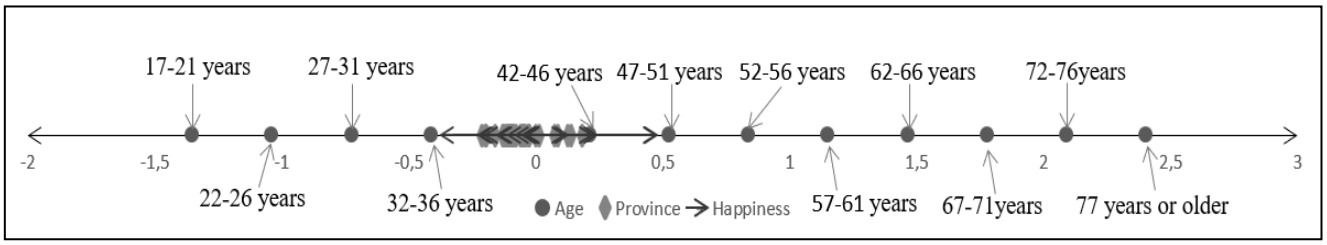

Fig 1. Hybrid SOCA with correlation values map.

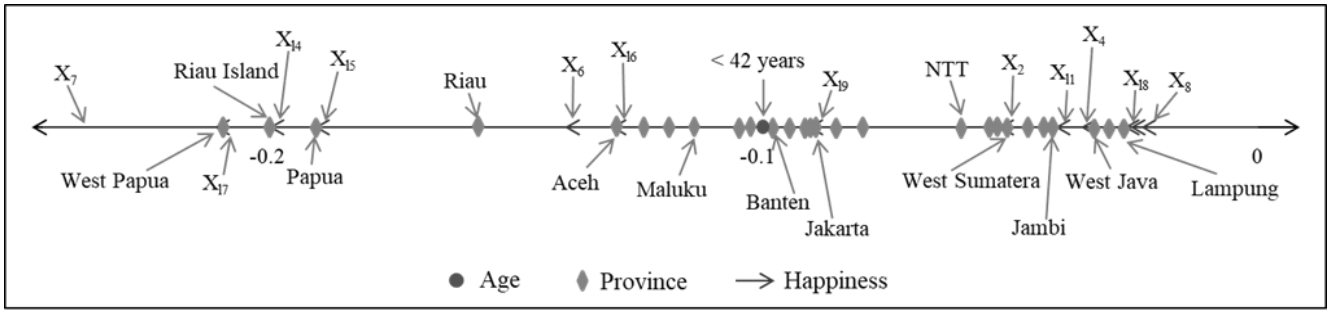

Fig 2. Enlarged map of Hybrid SOCA with correlation values. This is the part of the map with negative coordinates.

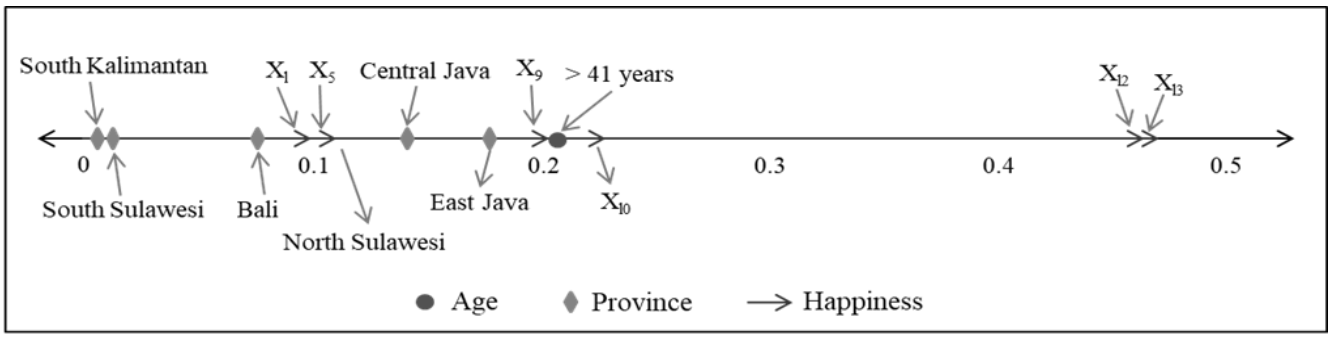

Fig 3. Enlarged map of Hybrid SOCA with correlation values. This is the part of the map with positive coordinates.

Figure 2 shows that the age range of 17 to 41 years had a negative sign. Provinces with negative principal coordinates had positive associations with the age category 17 to 41 years, including Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Island, Riau Island, Jakarta, West Java, Banten, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, Papua, and West Papua. On the other hand, provinces positively associated with the age category of 17 to 41 years were positively associated with 13 indicators of the happiness index. These indicators included occupation (X2), household income (X3), health (X4), family harmony (X6), availability of free time (X7), social relations (X8), feelings of happiness (X11), independence (X14), environmental mastery (X15), self-development (X16), positive relationships with others (X17), life goals (X18), and self-acceptance (X19). Based on this, it was found that achieving the 13 indicators would make the people in the 27 provinces happy. Therefore, these 13 indicators must be maintained if they are good and improved if they are not.

Based on Figure 3, the category of age 41 years or more was positive. Provinces with positive principal coordinates were positively associated with the age category of 41 years or more, such as Central Java, DIY, East Java, Bali, South Kalimantan, North Sulawesi, and South Sulawesi. In addition, provinces positively associated with the age category of 41 years or more were positively associated with six indicators of the happiness index. These indicators included education (X_1), housing conditions and home facilities (X_5), environmental conditions (X_9), security conditions (X_{10}), not feeling worried (X_{12}), and not feeling depressed (X_{13}). Based on this, it was found that the population in the seven provinces would feel happy if these six indicators could be achieved. Therefore, these seven indicators must be maintained if they are good and improved if they are not.

Based on the value of the happiness index, it was discovered that a person is in the happy category if their happiness index score reaches 70 [20]. Therefore, the increase in happiness was carried out based on indicators positively associated with the province with a value of less than 70. Thus, the main priority of increasing happiness in Aceh, North Sumatra, and 25 other provinces was based on 13 indicators positively associated with these provinces with an index value of less than 70. After that, improvements were made to six other indicators with an index of less than 70. Therefore, the main priority was to increase happiness in Central Java, DIY, and five other provinces, starting with six indicators positively associated with these provinces

with an index value of less than 70. Then it proceeded to the remaining 13 indicators with an index of less than 70.
Based on this, the five main priority indicators of increasing happiness in Aceh were household income, self-development, occupation, education, and skills, and feeling unconcerned. Meanwhile, in Central Java Province, there are five main priority indicators of increasing happiness, starting with indicators of education and skills, housing conditions and home facilities, feeling unconcerned, household income, and self-development.

IV. CONCLUSION

Based on the results and discussion, it is possible to conclude that the SOCA method was used to obtain the principal coordinates of the row and column categories. These coordinates were used to identify associations in discrete data made up of contingency tables with nominal and ordinal scales. The dimensions of the map were determined using the smallest variance value between the row and column categories. The vector coordinates of the continuous variable were obtained based on the correlation value between the continuous variables and the principal coordinates of the SOCA. These coordinates were used to identify relationships in continuous data. Using confidence regions or intervals and approximate p-values, the statistical significance of a category contributing to the structure of the association between two categorical variables with nominal and ordinal scales could be identified.

According to the case study, all nominal and ordinal scale variables significantly contributed to the association between age group and province variables. The results of the association revealed that 27 provinces were positively associated with the age category of 17 to 41 years, and these provinces were associated with 13 happiness index indicators. The other seven provinces were associated with the age category of 41 years or older, and these were associated with six indicators of the happiness index.

In order to improve the coverage of the information, the correspondence map should be made in two dimensions for future research. Thus, the calculation of confidence regions or intervals must consider different weights for each dimension, such as using confidence ellipses. If the cumulative variance is more than 70% in high dimensions (more than two), then the information cannot be obtained from the map. Therefore, information is obtained using the distance matrix and cluster analysis of principal coordinates and vector coordinates. Based on case studies, the provincial government, or Bappenas, can evaluate indicators focusing on increasing happiness for each age group in order to improve the welfare of the population.

ACKNOWLEDGMENT

The authors would like to thank Dr. Toni Toharudin and Dr. Yusep Suparman who have helped to get a better analysis.

REFERENCES

[1] Irlandia Ginanjar, N. Sunengsih, and Sudartianto, "The Method to Analyse the Association between Objects and Variables in the Form of I×2 Contingency Tables with Continuous Variables Additional Data", *Journal of Physics: Conference Series*, pp. 1–8, 2021.

[2] Alvin C. Rencher, "Methods of Multivariat Analisis. Second Edition", New York: John Wiley and Sons, Inc, 2002.

[3] Irlandia Ginanjar, "Hybrid Correspondence Analysis and Correlation to Analyse the Market Position from Data with Two Qualitative and p-2 Quantitative Variables", *AIP Conference Proceedings*, pp. 1–4, 2015.

[4] Eric J. Beh, "Simple Correspondence Analysis of Nominal-Ordinal Contingency Tables", *Journal of Applied Mathematics and Decision Sciences*, pp. 1–17, 2008.

[5] Subiyanto, Yuyun Hidayat, Eddy Afrianto, and Sudradjat Supian, "Numerical Estimation of the Final Size of the Spread of COVID-19 in West Java Province, Indonesia", *Engineering Letters*, vol. 29, issue 3, pp. 1015–1019, 2021.

[6] IGNM. Jaya, Y. Andriyana, and B. Tantular, "Spatial Prediction of Malaria Risk with Application to Bandung City, Indonesia", *IAENG International Journal of Applied Mathematics*, vol. 51, no. 1, pp 199–206, 2021.

[7] Irlandia Ginanjar, Septie Wulandary, and Toni Toharudin, "Empirical Best Linear Unbiased Prediction Method with K-Medoids Cluster for Estimate Per Capita Expenditure of Sub-District Level", *IAENG International Journal of Applied Mathematics*, vol. 52, issue 3, pp. 610–616, 2022.

[8] BPS, "Indeks Kebahagiaan", Jakarta: BPS, 2017.

[9] Sandhyarani Moirangthem and Satyananda Panda, "Happiness accross Age Groups: Findings Based on Three Measure", *International Journal of Health Science and Research*, vol. 8, no. 10, pp. 15–25, 2018.

[10] Priya Ratti and Daisy Sharma, "Happiness: does it transform with age?", *The International Journal of Indian Psychology*, vol. 9, no.1, pp. 960–979, 2021.

[11] Hui-Chuan Hsu1, Wen-Chiung Chang, Young-Sook Chong, and Jeong Shin An, "Happiness and Social Deeterminants across Age Cohort in Taiwan", *Journal of Health Psychology*, pp. 1–12, 2015.

[12] Theresia Puji Rahayu, "Determinan Kebahagiaan di Indonesia", *Jurnal Ekonomi dan Bisnis*, vol. 19, no. 1, pp. 139–147, 2016.

[13] Aisah Indati, "Konsep Kearifan pada Dewasa Awal, Tengah, dan Akhir", *Prosiding Temilnas XI IPPI*, pp. 26–35, 2019.

[14] Yukiko Uchida, Vinai Norasakkunkit, and Shinobu Kitayama, "Cultural Constructions of Happiness: Theory and Empirical Evidence", *Journal of Happiness Studies*, vol. 5, no. 3, pp. 223–239, 2004.

[15] Antonella Delle Fave, et al, "Lay Definitions of Happiness across Nations: The Primacy of Inner Harmony and Relational Connectedness", *Frontiers in Psychology*, vol. 7, no. 30, pp. 1–23, 2016.

[16] Herlani Wijayanti and Fivi Nurwianti, "Kekuatan Karakter dan Kebahagiaan pada Suku Jawa", *Jurnal Psikologi*, vol. 3, no. 2, pp. 114–122, 2010.

[17] Sari Zakiah Akmal1 and Fivi Nurwianti, "Kekuatan Karakter dan Kebahagiaan pada Suku Minang", *Jurnal Psikologi*, vol. 3, no. 1, pp. 16–24, 2009.

[18] Eric J. Beh and Rosaria Lombardo, "Confidence Regions and Approximate pvalues for Classical and Non Symmetric Correspondence Analysis", *Communications in Statistics - Theory and Methods*, vol. 4, no. 1, pp. 95–114, 2014.

[19] R. A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analisis. Sixth Edition*, Upper Saddle River, New Jersey: Pearson Prentice Hall, 2007.

[20] Rosaria Lombardo, Eric J. Beh, and Pieter M. Kroonenberg, "Modelling Trends in Ordered Correspondence Analysis", *The Psychometric Society*, vol. 81, no. 2, pp. 325 – 49, 2015.

[21] Irlandia Ginanjar, Ade Irma Nurwahidah, Jadi Suprijadi, Toni Toharudin, Sukono, "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)", *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 6, pp. 826–832, 2020.

[22] Eric J. Beh, "Confidence Circles For Correspondence Analysis Using Orthogonal Polynomials", *Journal of Applied Mathematics and Decision Sciences*, vol. 5, no. 1, pp. 35–45, 2001.

[23] OECD, "OECD Guidelines on Measuring Subjective Well-being", OECD Publishing, 2013.

Fajriatus Sholihah received her Magister in Applied Statistics from the University of Padjadjaran's Department of Statistics in 2022. She works as a contract employee at the BPS-Statistics of the Cirebon Regency. Her research interests are related to multivariate statistical analysis, data

mining, and regression. She has published research papers in various national journals.

Irlandia Ginanjar (M'20) graduated from the Institut Teknologi Bandung in 2017 with a Ph.D. in mathematics. He is an associate professor at Universitas Padjadjaran's Department of Statistics. His research interests include multivariate statistical analysis, unsupervised machine learning, small area estimation, sectoral data analysis, and marketing. He has published academic articles in national and international publications and conference proceedings, some of which Scopus has indexed. In addition, he serves as chairman of the department of statistics and the research group for big data analysis.

Yuyun Hidayat is currently a professor at the Department of Statistics, Universitas Padjadjaran, Indonesia. His research interests are related to quality control, time series, and management. He has published research papers in national and international journals and conference proceedings, some of which are indexed in Scopus. He received his PhD in the sciences of mathematics at Universiti Malaysia Terengganu (2018). Moreover, he acts as head of the research group in quality control.