VOL. 16, NO. 20, OCTOBER 2021 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

IDENTIFYING THE POTENTIAL FOR ONLINE AND MOBILE APPLICATION USE FOR BUS PASSENGERS IN WEST JAVA USING SIMPLIFICATION OF CORRESPONDENCE ANALYSIS

Irlandia Ginanjar[1], Udjianna S. Pasaribu[2], Eric J. Beh[3] and Sapto W. Indratno[2]
[1]Departement of Statistics, Universitas Padjadjaran, Indonesia
[2]Mathematics Study Program, Institut Teknologi Bandung, Indonesia
[3]School of Mathematical and Physical Sciences, University of Newcastle, Australia
E-Mail: irlandia@unpad.ac.id

ABSTRACT

Most bus companies in developed countries have provided web or mobile application-based services for their customers. Unfortunately, such services cannot be offered in many countries, including Indonesia, due to technology infrastructure limitations or a poor accessibility/usability of the internet. This paper aims to identify potential web or mobile application users for bus passengers in Indonesia based on the city from which they reside and their age. This paper analyses real bus passenger data between Bandung and Cirebon on the Indonesian island of Java. Identifying potential users of web or mobile-based applications for bus passengers based on their city of residence and age can be reliably studied using the simplification of correspondence analysis. One can also utilize many of the features for such a study including confidence regions that can be circular or elliptical in shape. This paper shows that there is great potential for Indonesians to use online or mobile-based applications if bus companies, or the government, provide them with such resources freely.

Keywords: simplification of correspondence analysis, elliptical confidence region, potential users.

1. INTRODUCTION

Most bus companies from developed countries have made freely available to their passengers their timetables and routes on the internet or through a mobile application. Research is abundant within the transport literature that has examined the development of web and mobile applications in the transport industry. For example, Zhou *et al.* [1] presented a passenger mobile phone data collection system and used it to model the bus trip route and estimate the arrival time at the various stops in Singapore. Camacho *et al.* [2] provided an overview of the impact of IT-based services currently offered in Australia's public transport sector. Korbel *et al.* [3] presented mobile applications aiding the visually impaired in using public transport. Ferreira *et al.* [4] offered an Android-based mobile application where data is taken from a public transportation information system called XTraN Passenger. Their application is designed to provide public transportation information and advice in real-time to its users.

Aside from the official timetable provided by a bus company or a government's transport department, public service providers, such as Google Maps, help a traveler navigate a local transport system. Such providers enable passengers to know which bus they should choose, when and where it departs from a stop, and when and where the bus will reach the passenger's destination. Google Maps obtain their information from the transit agencies who publish data following the General Transit Feed Specification developed by Google and the Portland TriMet agency. Indonesia's public transport services can be accessed with Google Maps and provide public transport services (such as a train timetable) for the city of Bandung. However, publicly accessible timetables for inter-city bus services in Indonesia are still not accessible; see Figure-1 for a screenshot highlighting the lack of information available through Google Maps for a bus route in the city.

Identifying the potential for online and mobile application use for Bandung and Cirebon bus passengers will be useful to study because the West Java provincial government is currently designing a metropolitan city centered in Cirebon called the Rebana Metropolitan [5]. Based on that, the route will connect two metropolitan cities, namely Bandung Raya and Rebana Metropolitan. The data studied in this paper were obtained from an inter-city bus company between Bandung and Cirebon (Figure-2). Dhiraj [6] writes that Bandung is the 19th most densely populated city in the world. Bandung's high population density causes excessive levels of traffic congestion and increased levels of air pollution. Following community consultation on transportation issues in Bandung, the feedback received showed that its residents felt that public transport is an indispensable asset to a city with a high population density [7]. Indonesia Central Agency on Statistics (BPS) [8] published that Indonesia's population in 2015 was 255.18 million, with a population density of 134 people per km^2. The population in Indonesia is not uniformly distributed, with most residents living in the major metropolitan cities. In particular, Bandung's population density is 14736 people per km^2, and for nearby Cirebon, the population density is 8157 people km^2. For such densely populated regions, the role of an efficient public transport system and the ease with which the population can easily access this transportation system are essential.

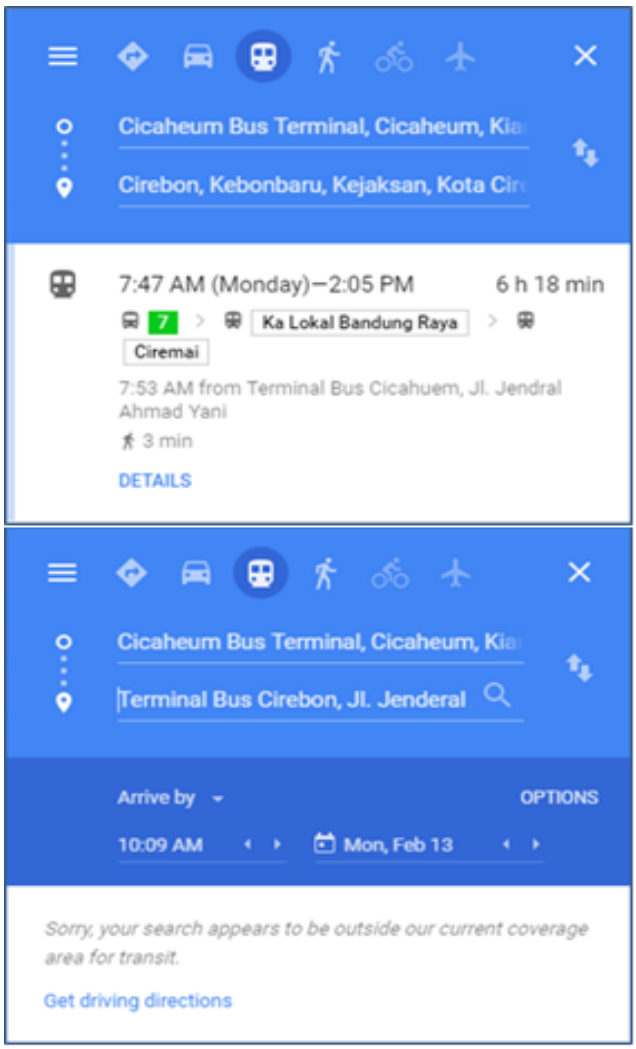

Figure-1. Public transport service information in Indonesia using Google Maps.

Gaining information about the penetration of Internet usage in Indonesia is necessary to understand the potential of utilizing public transport services through the internet or mobile applications. The 2016 survey results of the Indonesia Internet Service Provider Association (APJII) [9] reported some of the critical reasons for the importance of the internet and the need for a mobile application for public transportation in Indonesia. The number of Internet users in Indonesia reached 132.7 million at the end of 2016, with Internet user penetration reaching 51.8% of the population. The APJII also highlighted internet users' penetration by age groups; 75.5% of those aged 10-24 years regularly use the internet. They also found that 75.8% of people aged 25-34 years used the internet, while more than half (54.7%) of those aged 35-44 years used the internet. The report also found that older Indonesians were much less likely to use the internet with 17.2% for those aged 45-54 years, and only 2% aged 55+ years, not accessing it. The report [9] also showed that 85% of internet users access the internet using a mobile phone. More than 98.3% of internet users in Indonesia access the internet at least once a day. Survey data also make it clear that 91.6% of internet users use the internet for public services.

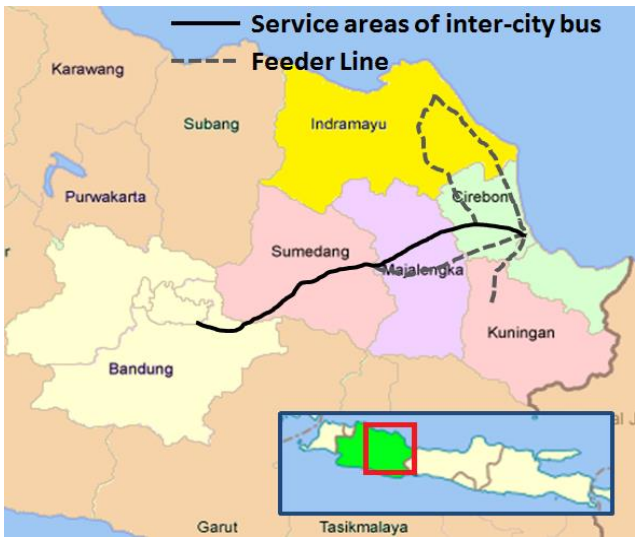

Figure-2. Service areas of the inter-city bus service between Bandung and Cirebon.

There have been extensive studies undertaken to investigate internet usage by bus passengers. For example, Trompet *et al.* [10] described an online data collection and normalization methodology developed by the International Bus Benchmarking Group. Kieu *et al.* [11] proposed a comprehensive method of bus passenger segmentation solely using smart card (SC) data. After reconstructing the travel itineraries from SC transactions, they studied their data using a density-based spatial clustering technique to mine each SC user's travel pattern. Vlassenroot *et al.* [12] collected data from mobile phones and developed new technologies to process and mine the data for behavior analysis. Ho *et al.* [13] proposed a market potential index (MPI) based model and particle swarm optimization for the bus feeder route from Taiwan's Miaoli High-Speed Rail Station.

Data visualization techniques have also been used in research undertaken on public transportation data. In particular, correspondence analysis (CA) has been shown to be highly beneficial for the study of such data. Sarnacchiaro and D'Ambra [14] used the Cumulative CA technique which uses Taguchi's index rather than Pearson's chi-squared statistic to assess the nature of the association between categorical variables. Their study utilized CA to study the satisfaction of potential train passengers between Naples and Rome in Italy. Diana [15] used CA to show if (and how) each attribute is related to the levels of public transport levels and how the relationship is affected by the urban context. Awang [16] applied CA to identify the characteristics of bus passengers across four Malaysian service zones.

Various studies have described the importance of the internet or mobile applications to monitor bus passenger satisfaction. Many of these studies generally describe the critical benefit of reducing the level of private vehicle usage. However, studies that focus on the

importance of the internet or mobile application use for public transport across Indonesia have been lacking. Therefore, this paper will examine the potential benefits of using the internet or a mobile application for Indonesia's transport sector. We analyzed data that is in the form of a contingency table which can then be studied to (firstly) identify the statistical significance of the association between its categorical variables. This can be easily done by performing a chi-squared test of independence between the variables. However, where an association is detected, this test does not reveal the nature of the association. That is, it does not show how specific row categories, say, compare. Nor does the test determine those row and column categories that are "associated" with each other. Therefore, CA will provide a deeper investigation of the association than the Pearson chi-squared statistic can provide. CA will identify those row (and column) categories that provide a similar, or different, contribution, to the association. It can also determine those categories from different variables that are closely associated and those that are not. The benefits of the correspondence analysis method are discussed in more detail by Greenacre [17] [18], Lebart *et al*. [19], and Beh and Lombardo [20].

The most utilized feature from performing a CA is the visual summary it provides of the association between the variables of a contingency table; such a summary is referred to as a correspondence plot and is constructed by simultaneously plotting row and column principal coordinates (PCo) in the same low-dimensional space. Typically, these coordinates are obtained via singular value decomposition (SVD) of the transformed contingency table. However, an alternative, and far more simple, approach is to adopt the strategy of Ginanjar [21] who proposed a simplification of correspondence analysis (SoCA) where a closed form solution can be used to determine the PCo's. The analysis can then be explored more deeply by constructing elliptically shaped confidence regions to identify those categories that make a statistically significant contribution to this association and those that do not.

From a more practical perspective, these analytical tools will be used to address this paper's key aim; to identify potential users of the internet and a mobile application for bus passengers in Indonesia based on the *passenger's city of residence* and their *age*. This issue will be examined by: 1) constructing a contingency table of the data where the *age* of the passenger is cross-classified with their *city of residence*, 2) formally testing the association between the two categorical variables using Pearson's chi-squared test of independence, and 3) visually identify the cities of residence (rows) and age group (the columns) that make a statistically significant contribution to the association between the two variables.

2. METHODS

2.1 Data

The data from an inter-city bus company was collected from bus passengers between 21 January 2007 and 13 July 2015, inclusive. The data set consists of 16355 individuals and 8 random variables. The random variables include one quantitative variable (*birth date*) and seven categorical variables. These are *name*, *birthplace*, *address*, *the city of residence*, *occupation*, *fixed-line* phone number and *mobile phone* number. The variables of *name*, *address*, *fixed-line* phone number, and *mobilephone* number are unique and confidential identifiers, so they are not used in our study. The variable *birthplace* is not used since it is not deemed to affect the passenger's behavior, while *occupation* does not match any variable name published by APJII [9]. Therefore, our analysis is confined to the study of only two variables - *passenger age* (which, as described below, is obtained from the *birth date* variable) and the *city of residence*. Thus, the data suggests that our analysis is confined to a *simple* CA rather than performing *multiple* CA (MCA).

In terms of the number of passengers, the dominant bus service links the city of Bandung with the city of Cirebon. Therefore, the variable *city of residence* is divided into three city categories: Bandung, Cirebon, and Others. The box plots in Figure-3 summarize the age distribution of passengers from these three city categories. They show that Bandung passengers are generally older than those from Cirebon and Other cities in the region, although the difference in age does not appear to be significant. When examining the passengers' age distribution from these cities using SoCA, any differences will be more clearly identifiable.

To study the association between the *passenger age* and *city of residence* variables, we first construct a contingency table with a sample size of n. This table is a cross-tabulation of two categorical variables. The first variable consists of I row categories while the second variable consists of J column categories; see Table-1. Suppose n_{ij} is the joint frequency of the number of individuals in row i and column j, for $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$. Let the total number of passengers classified into the i^{th} row be denoted by $n_{i\bullet}$ and the total of the j^{th} column be $n_{\bullet j}$, where $i= 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$.

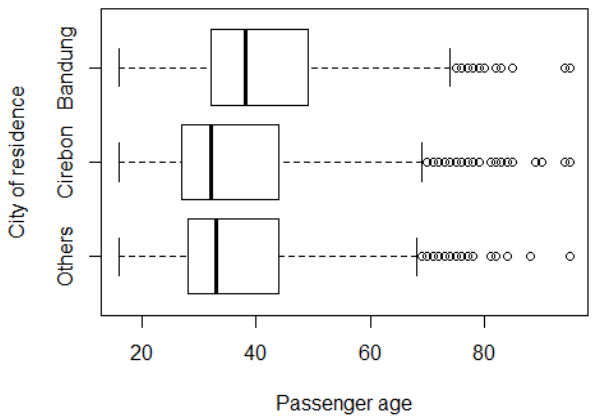

Figure-3. Passenger age boxplot for every city of residence.

Table-1. 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

Before performing a SoCA on our data, some data preparation is required. The age of the passengers collected as part of the study by APJII [9] is recorded as an integer and forms the birth date variable. For the purposes of our study, these integers are intervalized yielding the following five age categories: 16-24, 25-34, 35-44, 45-54, and 55+.

The contingency table is then constructed by cross-classifying the passenger age variable with the city of residence variable which consists of the three city categories; Bandung, Cirebon and Others. This contingency table is given as Table-2. SoCA is ideal for visually examining the association between the row variable (city of residence) and column variable (age) of Table-2.

Table-2. The contingency table of city of residence and age group.

	16-24	**25-34**	**35-44**	**45-54**	**55+**
Bandung	182	1532	1336	860	720
Cirebon	927	3599	1415	1004	942
Others	327	1742	797	480	456

Table-2 shows that 8.8% of all the passengers were aged 16-24 (and may be referred to as "often" internet users), while 42.12% of passengers were aged 25-34 ("frequent" users of the internet). Meanwhile, 21.74% of passengers were aged 35-44 ("normal" internet users), and 14.36% were 46-54 years of age ("rare" internet users). Table-2 also shows that 12.98% of the passengers were aged 55+ (and age group that very rarely used the internet). Therefore, by considering the mobile phone frequencies for the cities of Bandung and Cirebon, there is great potential for bus passengers to use the internet or a mobile application for accessing timetables and routes. This potential can be visualized using CA. The next section provides a brief overview of CA and a "simplified" version, SoCA, that can be used to calculate closed form PCo's for a contingency table of size $3 \times J$ (such as Table-2) or, equivalently, of the size $I \times 3$.

Pearson's chi-squared test of independence was used on the contingency table shown in Table-2, to identify interdependence between the city of residence and age group. Its calculation gives a statistic of 602.21 and a p-value <0.0001. Based on this value, there exists an association between the city of residence and age group so performing correspondence analysis on Table-2 with provide some meaningful interpretations of this association.

2.2 Correspondence Analysis

Correspondence analysis (CA) is a statistical technique that visualizes the association between two or more categorical variables. Extensive discussions of its theoretical, historical and practical development can be found in, for example, Greenacre [16, 17], Lebart *et al.* [19], and Beh and Lombardo [20]. To perform a CA, consider the contingency table:

$$\boldsymbol{N} = \left(n_{ij}\right),\ i = 1,2,\cdots,I\ ,\ \ j = 1,2,\cdots,J\ . \tag{1}$$

Calculate the correspondence matrix whose elements are the empirical joint distribution of the rows and columns such that:

$$\boldsymbol{P} = \left(p_{ij}\right),\quad p_{ij} = n_{ij}/n\,. \tag{2}$$

From this matrix, one can calculate the vector of row marginal proportions:

$$\boldsymbol{r} = \begin{pmatrix} r_1 & r_2 & \cdots & r_I \end{pmatrix}^t,\quad r_i = \sum_{j=1}^{J} p_{ij}$$

and **R,** which is the diagonal matrix containing these proportions. Similarly, the vector of marginal column proportions is:

$$\boldsymbol{c} = \begin{pmatrix} c_1 & c_2 & \cdots & c_I \end{pmatrix}^t,\quad c_j = \sum_{i=1}^{I} p_{ij}.$$

while the diagonal matrix of these proportions is denoted by ***C***.

Greenacre [16, 17, 21] and Beh and Lombardo [20] described that CA can be performed by considering the matrix of standardized residuals:

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

whose elements reflect the association between the variables by assessing the deviation between the observed cell frequencies and their expected value if the variables

are assumed to be independent. To obtain a visual inspection of the association between the variables, a singular value decomposition (SVD) is applied to the matrix of standardized residuals such that:

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{t},$$

where $\boldsymbol{U}^t\boldsymbol{U} = \boldsymbol{V}^t\boldsymbol{V} = \boldsymbol{I}$, and $\boldsymbol{D} = diag\left(\sqrt{\lambda_1} \quad \sqrt{\lambda_2} \quad \cdots \quad \sqrt{\lambda_L}\right)$. When $I<J$, the eigenvalue λ_ℓ is the ℓ^{th} largest eigenvalue of $SS^t = A$, and the eigenvalues are arranged in descending order, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_L$, where $\ell = 1,2,\cdots,L$ and L is the number of non-zero eigenvalues (= min(I, J) - 1). When $I > J$, the eigenvalue is also the ℓ^{th} largest eigenvalue of $S^tS = A$. Typically, even when I, J = 2, 3, most practitioners of correspondence analysis use the numerically intensive SVD approach to obtain the diagonal matrix of singular values D, the column matrix of left singular vectors $U = (u_{i\ell})$ and the column matrix of right singular vectors $V = (v_{j\ell})$.

The most common way to visualize the association between CA's categorical variables is to use principal coordinates (PCo's). For example, row PCo's are defined by the $I \times L$ matrix where

$$\boldsymbol{Y} = \boldsymbol{R}^{-1/2}\boldsymbol{U}\boldsymbol{D} = \left(y_{i\ell}\right), \quad y_{i\ell} = u_{i\ell}\sqrt{\frac{\lambda_\ell}{r_i}}. \tag{3}$$

Similarly, the column PCo's are defined by the $J \times L$ matrix where

$$\boldsymbol{Z} = \boldsymbol{C}^{-1/2}\boldsymbol{V}\boldsymbol{D} = \left(z_{j\ell}\right), \quad z_{j\ell} = v_{j\ell}\sqrt{\frac{\lambda_\ell}{c_j}}. \tag{4}$$

The first and second columns of the $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are the first and second principal coordinates (PCo's) of the rows and columns. These coordinates are used to construct a correspondence plot that is used to visualize the association between the rows and columns of Table-2. More details about these and other coordinate systems can be found in, for example, Beh and Lombardo [20].

2.3 Simplification of Correspondence analysis (SoCA)

Since Table-2 consists of three rows, calculating the PCo's can be calculated in an alternative manner. Ginanjar [21] proposed SoCA which performsCA calculations without the need to use SVD and is appropriate when I and/or J is 3. Here we will provide an overview of this approach to CA.

For the two-way contingency table, $\boldsymbol{N}$, defined by (1), let $\boldsymbol{n}_i = \left(n_{1\bullet} \quad n_{2\bullet} \quad \cdots \quad n_{I\bullet}\right)^t$, $\boldsymbol{n}_j = \left(n_{\bullet 1} \quad n_{\bullet 2} \quad \cdots \quad n_{\bullet J}\right)^t$, $\boldsymbol{D}_r = diag\left(\boldsymbol{n}_i\right)$, and $\boldsymbol{D}_c = diag\left(\boldsymbol{n}_j\right)$ so that the matrix $\boldsymbol{A}$ is defined by:

$$\boldsymbol{A} = \begin{cases} \boldsymbol{D}_r^{-\frac{1}{2}}\left(\boldsymbol{N}\boldsymbol{D}_c^{-1}\boldsymbol{N}^t - \frac{1}{n}\boldsymbol{n}_i\boldsymbol{n}_i^t\right)\boldsymbol{D}_r^{-\frac{1}{2}}, & for\ I \le J \\ \boldsymbol{D}_c^{-\frac{1}{2}}\left(\boldsymbol{N}^t\boldsymbol{D}_r^{-1}\boldsymbol{N} - \frac{1}{n}\boldsymbol{n}_j\boldsymbol{n}_j^t\right)\boldsymbol{D}_c^{-\frac{1}{2}}, & for\ I > J \end{cases}$$

Let $\boldsymbol{N}$be a $3 \times J$ or $I \times 3$ contingency table, with I, $J \ge 3$, so that $\boldsymbol{A} = \left(a_{k\tilde{k}}\right)$ for $k, \tilde{k} = 1,2,3$. Let $f = a_{11} + a_{22} + a_{33}$ $g = a_{11}a_{22} + a_{11}a_{33} + a_{22}a_{33}$, and $h = a_{12}^2 + a_{13}^2 + a_{23}^2$. Ginanjar [23] showed that the first and second eigenvalues of $A$are

$$\lambda_1 = \frac{f + \sqrt{f^2 - 4(g-h)}}{2} \text{ and } \lambda_2 = f - \lambda_1. \tag{5}$$

respectively.

Let $\tilde{\boldsymbol{u}}_\ell = \left(\tilde{u}_{1\ell} \quad \tilde{u}_{2\ell} \quad \tilde{u}_{3\ell}\right)^t$ be an orthogonal vector associated with the ℓ^{th} eigenvalue λ_ℓ for $\ell = 1,2$. The elements of this vector can be calculated by

$\tilde{u}_{1\ell} = a_{12}a_{23} - a_{13}a_{22} + a_{13}\lambda_\ell$,
$\tilde{u}_{2\ell} = a_{12}a_{13} - a_{23}a_{11} + a_{23}\lambda_\ell$,
$\tilde{u}_{3\ell} = \left(a_{11} - \lambda_\ell\right)\left(a_{22} - \lambda_\ell\right) - a_{12}^2$,

so that the matrix of left (row) singular vectors is:

$$\boldsymbol{U} = \left(u_{i\ell}\right), \quad u_{i\ell} = \frac{\tilde{u}_{i\ell}}{\sqrt{\sum_{i=1}^{3}\tilde{u}_{i\ell}^{\ 2}}}. \tag{6}$$

Similarly, the matrix of right (column) singular vectors can be derived from:

$$\boldsymbol{V} = \left(v_{j\ell}\right),$$
$$v_{j\ell} = \frac{1}{n\sqrt{n_{\bullet j}\lambda_\ell}}\left(\sum_{i=1}^{3}\frac{u_{i\ell}}{\sqrt{n_{i\bullet}}}\left(n_{ij}n - n_{i\bullet}n_{\bullet j}\right)\right) \tag{7}$$

From equations (5), (6), and (7), the eigenvalues λ_ℓ and the elements of the matrices $\boldsymbol{U} = (u_{i\ell})$ and $\boldsymbol{V} = (v_{j\ell})$ are obtained analytically. The principal coordinate can therefore be obtained far more simply than SVD allows by substituting these closed-form solutions to λ_ℓ, $u_{i\ell}$ and $v_{j\ell}$ into equations (3) and (4).

2.4 Approximate P-values and Elliptical Confidence Regions

To identify those row and column categories that provide a statistically significant contribution to the association between the row and column variables, Beh and Lombardo [24] derived approximate p-value formulae for each category. These p-values are based on each category's confidence region in a correspondence plot described in Beh [25] which can be circular or elliptical in shape. For example, circular confidence regions for the i^{th} row and j^{th} column categories have a radii length:

$$x_{i(\alpha)} = \sqrt{\frac{\chi^2_{\alpha,2}}{r_i}}\ ,\ x_{j(\alpha)} = \sqrt{\frac{\chi^2_{\alpha,2}}{c_j}}, \quad (8)$$

respectively, where χ^2_α is the chi-squared statistic with $(I-1)(J-1)$ degrees of freedom at the α level of significance; these regions were also described in Lebart *et al.* [19]. Similarly, for the elliptical confidence regions of the i^{th} row and j^{th} column categories, their semi-axis length along the ℓ^{th} principal axis is:

$$x_{i\ell(\alpha)} = \sqrt{\lambda_\ell \frac{\chi^2_\alpha}{X^2 r_i}}\ , x_{j\ell(\alpha)} = \sqrt{\lambda_\ell \frac{\chi^2_\alpha}{X^2 c_j}}, \quad (9)$$

respectively, where $X^2 = n\sum_{\ell=1}^{L}\lambda_\ell$ is the Pearson chi-squared statistic of Table-1.

To formally assess the statistical significance of the contribution of the i^{th} row and j^{th} column to the association between the variables consider the null and alternative hypotheses associated with the position of the i^{th} row and j^{th} column categories:

$$\begin{matrix} H_0 : y_{i\ell} = 0 \\ H_1 : y_{i\ell} \neq 0 \end{matrix} \text{ and } \begin{matrix} H_0 : z_{j\ell} = 0 \\ H_1 : z_{j\ell} \neq 0 \end{matrix}.$$

When constructing circular confidence regions using equation (8), the approximate p-value of the i^{th} row and j^{th} column, in a 2-dimensional correspondence plot is:

$$\begin{matrix} (p-value)_{i,2} = P\left\{\chi^2 > n_{i\bullet}\sum_{\ell=1}^{2} {y_{i\ell}}^2\right\} \\ \text{and} \\ (p-value)_{j,2} = P\left\{\chi^2 > n_{\bullet j}\sum_{\ell=1}^{2} {z_{j\ell}}^2\right\} \end{matrix}, \quad (10)$$

while for elliptical confidence regions, these approximate p-values are:

$$\begin{matrix} (p-value)_{i,2} = P\left\{\chi^2 > \varphi^2 n_{i\bullet}\sum_{\ell=1}^{2}\left(y_{i\ell}\Big/\sqrt{\lambda_\ell}\right)^2\right\} \\ \text{and} \\ (p-value)_{j,2} = P\left\{\chi^2 > \varphi^2 n_{\bullet j}\sum_{\ell=1}^{2}\left(z_{j\ell}\Big/\sqrt{\lambda_\ell}\right)^2\right\} \end{matrix}. \quad (11)$$

Here $\varphi^2 = X^2/n$ is referred to as the total inertia of the contingency table and is a measure of the strength of association between the variables when ignoring the impact of the sample size on the magnitude of Pearson's chi-squared statistic. More details about these methods can be found in, for example, Beh and Lombardo [24].

3. DATA ANALYSIS AND RESULTS

The PCo's for the three city and the five age categories are calculated by performing SoCA and the classical approach to CA on Table-2 and are summarized in Table-3. Table-3 shows that, to at least the 14th decimal place, the PCo's from both CA methods are, as expected, identical.

Table-3. The PCo's for three residence categories and age categories.

Categories	PCo	SoCA	CA	difference PCo
Bandung	1st	-0.29…5853	-0.29…5880	2.8×10^{-16}
	2nd	-0.01…7987	-0.01…7877	1.1×10^{-15}
Cirebon	1st	0.14…8734	0.14…8770	3.6×10^{-16}
	2nd	-0.02…5292	-0.02…5166	1.3×10^{-15}
Others	1st	0.05…4732	0.05…4681	5.1×10^{-16}
	2nd	0.05…9792	0.05…9838	4.6×10^{-16}
16-24	1st	0.36…6297	0.36…6298	1.1×10^{-16}
	2nd	-0.07…3839	-0.07…3803	3.6×10^{-16}
25-34	1st	0.13…7129	0.13…7128	1.1×10^{-16}
	2nd	0.02…2297	0.02…2289	8.0×10^{-17}
35-44	1st	-0.21…0998	-0.21…1001	2.8×10^{-17}
	2nd	0.01…7172	0.01…7170	2.4×10^{-17}
45-54	1st	-0.18…9732	-0.18…9731	5.6×10^{-17}
	2nd	-0.03…7079	-0.03…7135	5.6×10^{-16}
55+	1st	-0.12…9295	-0.12…9292	2.8×10^{-17}
	2nd	-0.02…6122	-0.02…6116	6.2×10^{-17}

The radii length of the circular confidence regions for each category of Table-2 are calculated using equation (8), while the semi-axis lengths for the elliptical confidence regions are found using equation (9). We have also calculated the approximate p-values using equation (10) and (11) for the three city and five age categories. These lengths and p-values are summarised in Table-4.

Table-4. The approximate p-values for three residence categories and age categories.

Categories	Circle	Ellipse		P-value	
	Radii	Axis 1	Axis 2	Circle	Ellipse
Bandung	0.0360	0.0570	0.0100	<0.0001	<0.0001
Cirebon	0.0276	0.0437	0.0077	<0.0001	<0.0001
Others	0.0397	0.0629	0.0111	<0.0001	0.0019
16-24	0.0646	0.1023	0.0180	<0.0001	<0.0001
25-34	0.0295	0.0468	0.0082	<0.0001	<0.0001
35-44	0.0411	0.0651	0.0114	<0.0001	<0.0001
45-54	0.0506	0.0801	0.0141	<0.0001	<0.0001
55+	0.0532	0.0843	0.0148	<0.0001	0.0001

The p-values associated with the circular and elliptical confidence regions for all of the age categories shows that they all significantly contribute to the association structure between the two variables. This is evident by observing that these regions in the SoCA maps of Figure-4 (with the confidence circles) and Figure-5 (showing the elliptical regions) do not overlap the origin (the point where all of the categories would lie if there was complete independence between the two variables). Figure-4 and Figure-5 also show that the passengers from Cirebon are strongly associated with the younger age groups (16-24 and 25-34) which are frequent internet users. Thus, passengers from Cirebon are likely to benefit greatly from online or mobile app, access to bus transport information. The passengers from other cities are strongly associated with the age category 25-34 thereby showing that they are also likely to benefit from such a service. The passengers from Bandung are strongly associated with age groups for those passengers aged 35 years or older. Therefore, passengers from this city are likely to have benefit from such services although they are deemed to be more in line with normal internet users.

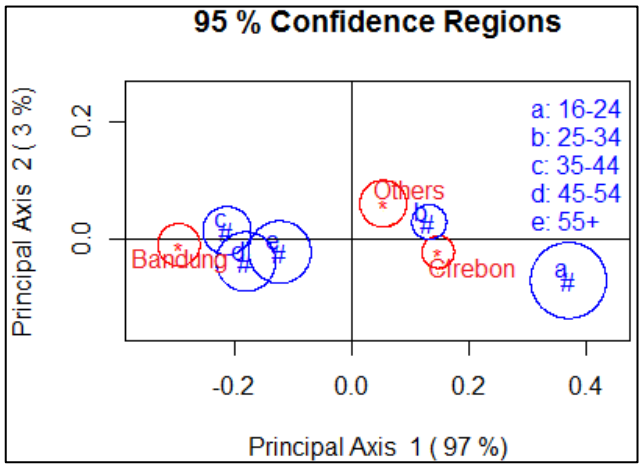

Figure-4. Circular confidence region from the soca of Table-2.

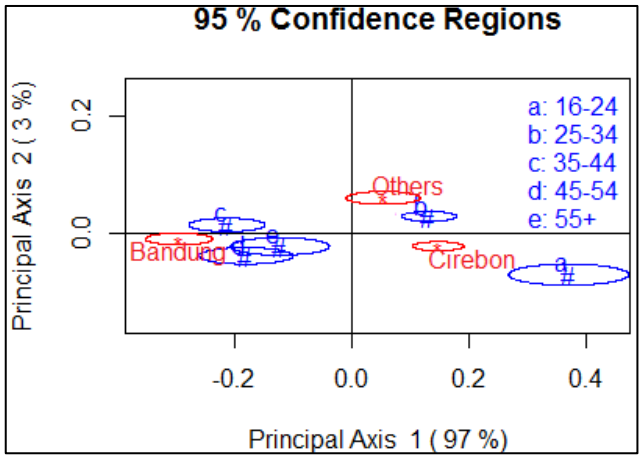

Figure-5. Elliptical confidence region from the SoCA of Table-2.

4. CONCLUSIONS

Using SoCA has identified that there is a potential benefit in adopting online technology in the form of the internet or mobile applications for bus passengers between Bandung and Cirebon. These findings are based the analysis of the passenger's city of residence and age and come from the 2016 survey results of the Indonesian Internet Service Provider Association (APJII). Bus passengers with the highest potential of online/application usage are of an age where they are deemed to be frequent internet users (especially among the young adults of the population) and reside in Cirebon. The circular and elliptical regions confirm that each of the cities and age groups studied plays a statistically significant role in the association that exists in the data.

Those bus passengers who are potentially needing internet or mobile application resources and live in Bandung and Cirebon are demanding that bus companies or the government provide such services and do so free of charge. To do so will reduce the use of private vehicles and help reduce the level of air pollution and traffic congestion between the two cities. Future research into this issue can be used to perform CA on data that incorporate additional factors. Doing so will provide deeper insight into the needs of the bus passengers, and better understand the impact on the public transport system, its impact on the environment and cultural development in Java, Indonesia.

ACKNOWLEDGEMENT

The authors would like to thank Aldo FantinusWiyana, M.Sc., MBA, who has helped get data and support. The authors acknowledge the financial support provided by the acceleration of associate professor research (RPLK) Universitas Padjadjaran 2021, Number 1959/UN6.3.1/PT.00/2021. We also thank for all reviewers.

REFERENCES

[1] P. Zhou, Y. Zheng and M. Li. 2012. How Long to Wait ? Predicting Bus Arrival Time with Mobile Phone based Participatory Sensing Categories and Subject Descriptors. IEEE Trans. Mob. Comput. 13(6): 379-392, doi: 10.1109/TMC.2013.136.

[2] T. D. Camacho, M. Foth and A. Rakotonirainy. 2012. Pervasive Technology and Public Transport: Opportunities beyond Telematics. IEEE Pervasive Comput. 12(1): 18-25, doi: 10.1109/MPRV.2012.61.

[3] P. Korbel, P. Skulimowski, P. Wasilewski and P. Wawrzyniak. 2013. Mobile Applications Aiding the Visually Impaired in Travelling with Public Transport. in Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, pp. 825-828, [Online]. Available: https://fedcsis.org/proceedings/2013/.

[4] J. C. Ferreira, H. Silva, J. A. Afonso and J. L. Afonso. 2018. Context aware advisor for public transportation. IAENG Int. J. Comput. Sci. 45(1): 74-81.

[5] Hermansyah. 2020. West Java launches Rebana Metropolitan, Subang Smartpolitan in 2020 WJIS. Humas Provinsi Jawa Barat, http://humas.jabarprov.go.id/west-java-launches-rebana-metropolitan-subang-smartpolitan-in-2020-wjis/3961 (accessed Nov. 16, 2020).

[6] A. B. Dhiraj. 2017. These are the top 20 most densely populated cities in the world. CEOWORLD magazine, http://ceoworld.biz/2017/07/25/these-are-the-top-20-most-densely-populated-cities-in-the-world/ (accessed Jul. 25, 2017).

[7] P. R. Santoso. 2016. Revisited Composition: Implementation Strategy for Mobility Based Development in Bandung City, Indonesia. Delft University of Technology.

[8] Badan Pusat Statistik (BPS). 2015. Profil Penduduk Indonesia Hasil Supas 2015. Jakarta: Badan Pusat Statistik.

[9] APJII. 2016. Penetrasi dan Perilaku Pengguna Internet Indonesia 2016. Jakarta: Indonesia Internet Service Provider Association.

[10] M. Trompet, R. Parasram and R. J. Anderson. 2013. Benchmarking Disaggregate Customer Satisfaction Scores of Bus Operators in Different Cities and Countries. Transp. Res. Rec. J. Transp. Res. Board, 2351: 14-22, doi: 10.3141/2351-02.

[11] L. M. Kieu, A. Bhaskar and E. Chung. 2015. Passenger Segmentation Using Smart Card Data. IEEE Trans. Intell. Transp. Syst. 16(3): 1537-1548, doi: 10.1109/TITS.2014.2368998.

[12] S. Vlassenroot, D. Gillis and R. Bellens. 2015. The Use of Smartphone Applications in the Collection of Travel Behaviour Data. Int. J. Intell. Transp. Syst. Res. 13(1): 17-27, doi: 10.1007/s13177-013-0076-6.

[13] W. C. Ho, J. M. Su and C. Y. Chang. 2019. Maximal market potential of feeder bus route design using particle swarm optimization. IAENG Int. J. Appl. Math. 49(4): 1-8.

[14] P. Sarnacchiaro and A. D'Ambra. 2011. Cumulative Correspondence Analysis to improve the public train transport. Electron. J. Appl. Stat. Anal. Decis. Support

Syst. Serv. Eval. 2(1): 15-24, doi: 10.1285/i2037-3627v2n1p15.

[15] M. Diana. 2012. Measuring the satisfaction of multimodal travelers for local transit services in different urban contexts. Transp. Res. Part A. 46(1): 1-11, doi: 10.1016/j.tra.2011.09.018.

[16] S. B. Awang. 2015. An Investigation on Leading Characteristics of Rapidkuantan Bus Passengers using Correspondence Analysis. Thesis, Universiti Malaysia Pahang.

[17] M. J. Greenacre. 1984. Theory and Applications of Correspondence Analysis. Orlando: Academic Press Inc.

[18] M. Greenacre. 2017. Correspondence Analysis in Practice, Third Edit. Boca Raton: Taylor & Francis Group, LLC.

[19] L. Lebart, A. Morineau and K. M. Warwick. 1984. Multivariate Descriptive Statistical Analysis-Correspondence Analysis and Related Techniques for Large Matrices. New York: John Wiley and Sons, Ltd.

[20] E. J. Beh and R. Lombardo. 2014. Correspondence Analysis Theory, Practice and New Strategies. West Sussex: John Wiley & Sons, Ltd.

[21] I. Ginanjar, U. S. Pasaribu and A. Barra. 2016. Simplification of correspondence analysis for more precise calculation which one qualitative variable is two categorical data. ARPN J. Eng. Appl. Sci. 11(3): 1983-1991.

[22] M. Greenacre. 2013. The contributions of rare objects in correspondence analysis. Ecology. 94(1): 241-249, doi: 10.1890/11-1730.1.

[23] I. Ginanjar. 2017. Penyederhanaan analisis-korespondensi untuk meningkatkan akurasi koordinat utama. Institut Teknologi Bandung.

[24] E. J. Beh and R. Lombardo. 2015. Confidence Regions and Approximate p- values for Classical and Non Symmetric Correspondence Analysis. Commun. Stat. - Theory Methods. 44: 95-114, doi: 10.1080/03610926.2013.768665.

[25] E. J. Beh. 2010. Elliptical confidence regions for simple correspondence analysis. J. Stat. Plan. Inference. 140(9): 2582-2588, doi: 10.1016/j.jspi.2010.03.018.