Quantitative Methods for SOCIAL SCIENCES

Edited by: Vinayak Nikam • Abimanyu Jhajhria• Suresh Pal

This reference book is designed keeping in mind the need for the application of advanced quantitative methods in social science research to enhance its accuracy. The chapters are written in such a way that social scientists can easily grasp the methods including their theoretical and practical aspects using statistical software. The book provides comprehensive coverage of multivariate techniques, forecasting methods, structural equations, optimization models, quantitative methods for impact assessment, growth models and other important methods used in social science research.

ABOUT THE EDITORS

Vinayak Nikam is working as Scientist (Senior Scale) at ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He did his PhD from ICAR-Indian Agricultural Research Institute (New Delhi) in agricultural extension. He also served ICAR-Central Soil Salinity Research Institute (Karnal) before joining NIAP in Nov 2015. He also holds faculty membership in the discipline of Agricultural Extension at ICAR-Indian Agriculture Research Institute (New Delhi). Currently, he is working on the performance and impact assessment of Agriculture Extension and Advisory Services as well as associated with technology impact assessment.

Abimanyu Jhajhria is a Scientist (Agricultural Economics) at ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He has received his doctorate in Agricultural Economics from the ICAR-Indian Agricultural Research Institute (New Delhi). His research interest includes markets, trade and institutions. He is involved in the projects on market reforms, agricultural commodity value chains and outlook models for agricultural commodities.

Suresh Pal is Director of ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He has a PhD in agricultural economics from ICAR-Indian Agricultural Research Institute (New Delhi). He has published extensively on different aspects of Indian agriculture and guided doctoral and masters students. He has received awards for his contributions like best journal article awards, Norman E Borlaug International Science Fellowship, Fellow of the Indian Society of Agricultural Economics (Mumbai), and Fellow of the National Academy of Agricultural Sciences.

ISBN:978-81-940080-2-6

ICAR-NATIONAL INSTITUTE OF AGRICULTURAL ECONOMICS AND POLICY RESEARCH (NIAP)
D.P.S.Marg, Pusa, New Delhi - 110012. Phone: +91-11- 25847628, +91-11- 25848731
Fax: +91-11-25842684 . E-mail:director.niap@icar.gov.in; Website: www.http://www.ncap.res.in

भाकृअनुप ICAR

Quantitative Methods for Social Sciences

Vinayak Nikam · Abimanyu Jhajhria · Suresh Pal

Vinayak Nikam ● Abimanyu Jhajhria ● Suresh Pal

Quantitative Methods for SOCIAL SCIENCES

भाकृअनुप ICAR

ICAR-National Institute of Agricultural Economics and Policy Research

Reference book on

Quantitative Methods for SOCIAL SCIENCES

Edited by

Vinayak Nikam
Abimanyu Jhajhria
Suresh Pal

भाकृअनुप
ICAR

ICAR-National Institute of Agricultural Economics and Policy Research
New Delhi

Quantitative Methods for
SOCIAL SCIENCES

Edited by
Vinayak Nikam, Abimanyu Jhajhria and Suresh Pal

© 2019 ICAR-National Institute of Agricultural Economics and Policy Research

Published by
Dr. Suresh Pal
Director, ICAR-National Institute of Agricultural Economics and Policy Research

ISBN: 978-81-940080-2-6

Printed at
Chandu Press: chandupress@gmail.com

CONTENTS

Chapter 5

CORRESPONDENCE ANALYSIS

Deepak Singh, Raju Kumar, Ankur Biswas, R. S. Shekhawat and Abimanyu Jhajhria

INTRODUCTION

Correspondence Analysis (CA) is an exploratory-multivariate-graphical technique representing non-metric information arranged in two-way contingency tables in correspondence maps. A contingency table is a two way cross-classification, which contains the frequency of items (counts) in each cell having the information about the joint distribution of categorical variables. Categorical variables are variables, which are either nominal or ordinal in nature like counts or frequency. Correspondence analysis has unique capability to represent both linear and non-linear relationships. This technique transforms the categorical data into metric data, applies dimensional reduction technique and represents the information into correspondence maps. In multidimensional reduction, the singular value decomposition is used in which the orthogonal components are extracted in decreasing order of importance so that the maximum information can be presented in two or three-dimensional correspondence maps. The correspondence map of correspondence analysis is the geometric approach having ability to represent the interaction of the two categorical variables graphically.

In 1960's Jean-Paul Benzecri and his colleagues developed this technique and named the technique as "analyse factorielle des correspondances" but later shortened this to "analyse des correspondances" which is "correspondence analysis" in english translation. It is also referred as reciprocal averaging, dual scaling, optimal scaling or scoring, homogeneity analysis, canonical analysis of contingency tables, categorical discriminant analysis, and multivariate quantification of qualitative data. When the contingency tables are three way or higher order then the multiple correspondence analysis is used which is the extension of correspondence analysis and the analytical procedure is similar to correspondence analysis.

Differences from other multivariate techniques

Conceptually it is similar to Principal Component and Factor analysis dealing with summarizing the data in graphical form through singular value decomposition but being assumption free it is a wonderful approach to work with categorical data where Principal Component and Factor analysis fails. Both multidimensional scaling and correspondence analysis can deal with nominal data but the compositional (Attribute – Based) approach of the correspondence analysis distinguishes it from the multidimensional scaling which is decompositional (Attribute – Free) in nature.

Uniqueness of correspondence analysis

All the multivariate techniques deal with either continuous data or main effects of categorical data but this multivariate technique has the ability to deal with main and interaction effects of the categorical data. Therefore, it has the unique features like

- Capability to use categorical data and therefore assumption free
- Compositional in nature
- Ability to represent the relationship of the rows/objects to each other
- Ability to represent the relationship of the columns/variables to each other
- Ability to represent the interaction of the objects variables to each other
- Interdependence of categorical variables
- Dimension reduction and multivariate in nature
- Capable to represent both linear and non-linear relationships.
- Correspondonce mapping of categorical variables simultaneously which is the explicit objective of correspondence analysis

Limitations

Correspondence analysis performs poor when cell frequencies in the contingency table are very small or zero, in which case it is usually recommended to combine two or more categories to increase the frequencies of the cells. All data should be non-negative under same scale of correspondence analysis to be applicable and the method treats rows and columns equivalently.

Workflow

The workflow for the correspondence analysis is explained in Fig 1.

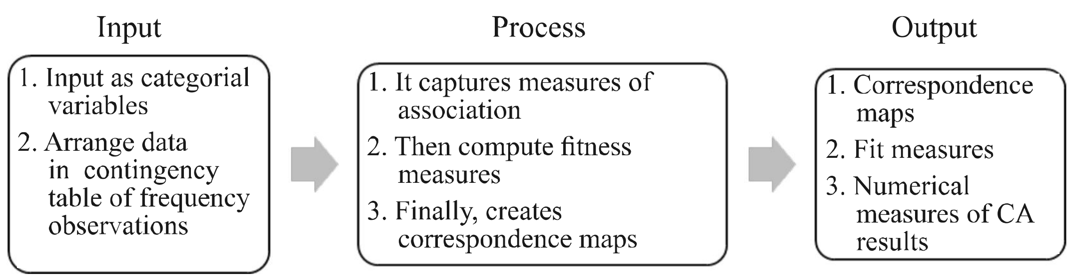

Fig 1: Workflow of the correspondence analysis

Analytical procedure for correspondence analysis

Suppose there are two variables X and Y in nominal scale having levels p and q respectively, which can be expressed in two-way contingency or cross table. The table is bivariate table having i^{th} row and j^{th} column has the cell frequency of $n_{ij} > 0$. If ordinal data is used it is considered as nominal data and the continuous variable like age, plant height, yield, are first converted into nominal data before carrying Correspondence analysis.

This technique examines the correspondence between the observed frequencies and the variables considered either independently or jointly. The contingency table can be expressed in matrix form where n_{ij} are the elements of i^{th} row and j^{th} column of the matrix.

First of all the independence of categories (row by column) is checked by chi-square test statistic and visually the contingency table can be inspected by using graphical matrix and mosaic/ association plots for variables to be interdependent. If the tests are significant, then the further steps are followed for correspondence analysis.

Contingency table

	Y_1	Y_2	...	Y_q	$sum(Y_i)$
X_1	n_{11}	n_{12}	...	n_{1q}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2q}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
X_p	n_{p1}	n_{p2}	...	n_{pq}	$n_{p.}$
$sum(X_i)$	$n_{.1}$	$n_{.2}$	...	$n_{.q}$	$n_{..}$

ILLUSTRATION

For the illustration purpose, "Housetasks" data has been taken from factoextra package of R software. In the Housetasks data rows indicates different household tasks and columns indicate the tasks done by different groups of households i.e. wife, husband, jointly and alternatively. The cell values are the frequencies of the tasks done by different groups of a household.

Housetasks	**Wife**	**Alternatively**	**Husband**	**Jointly**	**Total**
Laundry	156	14	2	4	176
Main meal	124	20	5	4	153
Dinner	77	11	7	13	108
Breakfast	82	36	15	7	140
Tidying	53	11	1	57	122
Dishes	32	24	4	53	113
Shopping	33	23	9	55	120
Official	12	46	23	15	96
Driving	10	51	75	3	139
Finances	13	13	21	66	113
Insurance	8	1	53	77	139
Repairs	0	3	160	2	165
Holidays	0	1	6	153	160
Total	**600**	**254**	**381**	**509**	**1744**

Our aim in correspondence analysis is to find out the following points:

1. What is the relationship among household tasks with respect to working group of households?
2. What is the relationship among working group of households with respect to household tasks?
3. What is the relationship of household tasks with working group of households?
4. Can these relationships be shown graphically?

To answer the above four questions in correspondence analysis, the following four solutions can be made in correspondence analysis.

Ans 1. Relationship among household tasks = By Row profile (R)

Ans 2. Relationship among working group of households = By Column profile (C)

Ans 3. Relationship between tasks and working group of households= By weighted Chi square distance

Ans 4. Dimension reduction (Singular value decomposition, SVD) and correspondence maps.

Steps for Analytical Procedure

Step 1: First test the independence of categorical variables with chi square test statistics and if test is significant, then we go for next step.

Step 2: Develop correspondence matrix $m_{ij} = \left(\frac{n_{ij}}{n..}\right)$

Step 3: Develop row profile i.e. $\left(m_{ij} / row\ mass\right)$

Step 4: Develop column profile i.e. $\left(m_{ij} / column\ mass\right)$

Step 5: Analyze weighted χ^2 distance = $D = D_r^{-1/2}\left(M - rc^T\right)D_c^{-1/2}$

Step 6: Carry out singular value decomposition (SVD) or dimension reduction technique

Step 7: Calculate overall fit measures and plot correspondence maps

Step 1: Pearson's Chi-square test for Housetasks data

First of all the chi-square test is applied to the housetasks contingency table to test whether the categorical variables are independent for further analysis.

$$\chi^2 = 1944.456,\ df = 36,\ p\text{-value} = 0$$

The p value is almost zero, which infers that the household tasks and different groups of a households are interdependent. Therefore the correspondence analysis should be applied for inter-relationship of tasks and working group of households.

Step 2: Correspondence matrix M

$$M_{pxq} = m_{ij} = \left(\frac{n_{ij}}{n..}\right) =$$

	Y_1	Y_2	...	Y_q	*row mass*
X_1	m_{11}	m_{12}	...	m_{1q}	$m_{1.}$
X_2	m_{21}	m_{22}	...	m_{2q}	$m_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_p	m_{p1}	m_{p2}	...	m_{pq}	$m_{p.}$
column mass	$m_{.1}$	$n_{.2}$	...	$m_{.q}$	1

=

Correspondence matrix M of Housetasks data:

Task	**Wife**	**Alternatively**	**Husband**	**Jointly**	**Row mass**
Laundry	0.089	0.008	0.001	0.002	**0.025**
Main meal	0.071	0.011	0.003	0.002	**0.022**
Dinner	0.044	0.006	0.004	0.007	**0.015**
Breakfast	0.047	0.021	0.009	0.004	**0.020**
Tidying	0.030	0.006	0.001	0.033	**0.017**
Dishes	0.018	0.014	0.002	0.030	**0.016**
Shopping	0.019	0.013	0.005	0.032	**0.017**
Official	0.007	0.026	0.013	0.009	**0.014**
Driving	0.006	0.029	0.043	0.002	**0.020**
Finances	0.007	0.007	0.012	0.038	**0.016**
Insurance	0.005	0.001	0.030	0.044	**0.020**
Repairs	0.000	0.002	0.092	0.001	**0.024**
Holidays	0.000	0.001	0.003	0.088	**0.023**
Column mass	**0.026**	**0.011**	**0.017**	**0.022**	**1.000**

The vector of row sums of M_{pxq} will be row mass vector i.e.

Row sum vector = r = $(m_{1.} \quad m_{2.} \quad \ldots \quad m_{p.})' =$

$(0.025 \quad 0.022 \quad 0.015 \quad 0.020 \quad 0.017 \quad 0.016 \quad 0.017 \quad 0.014 \quad 0.020 \quad 0.016 \quad 0.020 \quad 0.024 \quad 0.023)'$

The vector of column sums of M_{pxq} will be containing column mass in the following manner.

Column sum vector = c = $(m_{.1} \quad m_{.2} \quad \ldots \quad m_{.q})' = (0.026 \quad 0.011 \quad 0.017 \quad 0.022)'$

Therefore, transpose of c will be

$$c^T = \begin{pmatrix} m_{.1} & m_{.2} & \dots & m_{.q} \end{pmatrix} = \begin{pmatrix} 0.026 & 0.011 & 0.017 & 0.022 \end{pmatrix}$$

The D_r and D_c are the diagonal matrix of row and column profiles respectively i.e.

$$D_r = diag(r) = \begin{bmatrix} m_{1.} & 0 & \dots & 0 \\ 0 & m_{2.} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{p.} \end{bmatrix} = \begin{bmatrix} 0.025 & 0 & \dots & 0 \\ 0 & 0.022 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0.023 \end{bmatrix}$$

and

$$D_c = diag(c) = \begin{bmatrix} m_{.1} & 0 & \dots & 0 \\ 0 & m_{.2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{.q} \end{bmatrix} = \begin{bmatrix} 0.026 & 0 & 0 & 0 \\ 0 & 0.011 & 0 & 0 \\ 0 & 0 & 0.017 & 0 \\ 0 & 0 & 0 & 0.022 \end{bmatrix}$$

Step 3: Row profiles

When the rows of the correspondence matric are divided by row mass, we get row profiles.

Interpretation

The row profiles are used to compare the relationship and asses the proportion of row by column. It can be evaluated which column in each row account for more or less percentage of counts. SVD can also be applied to row profiles to visually compare the row categories i.e tasks in case of household tasks data.

Step 4: Column profiles

When the columns of the correspondence matrix are divided by column mass, we get column profiles.

Interpretation

The column profiles are used to compare the relationship and asses the proportion of column with respect to row. It can be evaluated which rows in each column account for more or less percentage of counts. SVD can also be applied to column profiles to visually compare the column categories i.e working groups in case of household tasks data.

Step 5: The weighted χ^2 distance

$$D = D_r^{-1/2}\left(M - rc^T\right)D_c^{-1/2}$$

Table 1: Weighted χ^2 distance of Housetasks data

Task	Wife	Alternating	Husband	Jointly
Laundry	3.435899	0.460669	0.035104	0.072572
Main meal	2.927131	0.715917	0.130137	0.081171
Dinner	2.161006	0.465753	0.232713	0.381186
Breakfast	2.017169	1.361655	0.449981	0.167867

Task	Wife	Alternating	Husband	Jointly
Tidying	1.391096	0.436609	0.016304	1.629630
Dishes	0.865506	1.008072	0.122516	1.574530
Shopping	0.865512	0.936114	0.286520	1.585126
Official	0.341471	2.11184	0.852017	0.471749
Driving	0.226737	1.942315	2.331838	0.060181
Finances	0.339318	0.539865	0.713327	1.965414
Insurance	0.176797	0.023437	1.642464	2.066348
Repairs	-0.025020	0.089395	4.581751	0.026722
Holidays	-0.024640	0.019741	0.155606	3.843418

Step 6: Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is required to reduce the dimensions of weighted Chi-sqaure distance so that the categorical variables can have visual representation. Therefore, SVD is applied to partition D matrix into three matrices i.e. U, V and S. Here U is pxk matix, V is qxk matrix and S is kxk diagonal matrix whose diagonal elements are $s_1 \geq s_2 \geq s_i \ldots \geq s_k$ or $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \sqrt{\lambda_i} \ldots \geq \sqrt{\lambda_k}$, here k is the reduced dimension of D matrix. The s_i and λ_i are the singular and eigen values of i^{th} component (PCi). Here $k = \min\{p-1, q-1\}$, therefore in "housetasks" data case *k=q-1=3*. The SVD of D matrix is given by

$$D_{pxq} = U_{pxk} S_{kxk} V_{qxk}^T$$

Due to positive definite matrix D, the strict positivity of λ_i is guaranteed. The eigen value λ_k represents the weighted variance explained by k^th^ component. Here PC1 is the component corresponding to λ_1, PC2 is the component corresponding to λ_2 and so on. The first component PC1 has the maximum variance, the second largest variance is observed for PC2 and so on because of diagonal values arranged in $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \sqrt{\lambda_i} \ldots \geq \sqrt{\lambda_k}$ order. The percentage of variation by i^{th} component is given as

$$\%\, variance = \left[\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}\right] * 100$$

Step 7: Overall fit measures and correspondence maps

Overall fit Measures

In overall fit measures, the inertia is defined as percent of variance explained to the total variance by each of the categories. Mass is nothing but row mass and column mass defined earlier in correspondence matrix. Contribution in percent of dimensions indicates what is the contribution of individual categories in building the components and coordinate values are the coordinate values for the principal axis. The correlation in dimensions category indicates about how good the components are good to explain the each of the categories (i.e. row and column elements). Table 2 indicates the eigen values, percent of variance and cumulative per cent of variance of new components after SVD.

In Table 3, in case of rows, the maximum inertia or per cent of variance is explained by repairs, laundry, driving and holidays tasks while in working group of households the maximum variance is explained by husband group. In case of dimension 1, repairs task contributes maximum and shopping contributes minimum and it can be seen in correspondence plot that along X axis (Dimension 1) the repairs task is farthest while shopping is nearest to origin. The correlation column of dimension 1 is indicating that dimension 1 is good to explain laundry, main meal, dinner, repairs tasks and wife working group of households. Similarly, in case of dimension 2, holidays task contributes maximum and official contributes minimum and it can be seen in correspondence plot that along Y axis (Dimension 2) the holidays task is farthest while official is nearest to origin. The correlation column of dimension 2 is indicating that dimension 2 is good to explain shopping, finances and holidays tasks and jointly working group of households.

Table 2: Eigen value and per cent of variance explained by new components of SVD

Dimension	Eigen value λ	% variance explained by eigen value	Cumulative % variance
PC1	0.54	48.69	48.69
PC2	0.45	39.91	88.60
PC3	0.13	11.40	100.00

Table 3: Overall fit measures of each point in first two dimensions of SVD

Character			Dimension 1			Dimension 2		
Row	**Mass**	**Inertia**	**Coordinate**	**Contribution in %**	**Correlation**	**Coordinate**	**Contribution in %**	**Correlation**
Laundry	0.025	0.134	-0.99	18.29	0.74	0.50	5.56	0.18
Main meal	0.022	0.091	-0.88	12.39	0.74	0.49	4.74	0.23
Dinner	0.015	0.038	-0.69	5.47	**0.78**	0.31	1.32	0.15
Breakfast	0.020	0.041	-0.51	3.82	0.50	0.45	3.70	0.40
Tidying	0.017	0.025	-0.39	2.00	0.44	-0.43	2.97	0.54
Dishes	0.016	0.020	-0.19	0.43	0.12	-0.44	2.84	0.65
Shopping	0.017	0.015	-0.12	0.18	0.06	-0.40	2.52	0.75
Official	0.014	0.053	0.23	0.52	0.05	0.25	0.80	0.07
Driving	0.020	0.102	0.74	8.08	0.43	0.65	7.65	0.34
Finances	0.016	0.030	0.27	0.88	0.16	-0.62	5.56	0.84
Insurance	0.020	0.058	0.65	6.15	0.58	-0.47	4.02	0.31
Repairs	0.024	**0.313**	1.53	**40.73**	0.71	0.86	15.88	0.23
Holidays	0.023	0.196	0.25	1.08	0.03	-1.44	**42.45**	**0.96**
Column								
Wife	0.026	0.301	-0.84	44.46	**0.80**	0.37	10.31	0.15
Alternating	0.011	0.118	-0.06	0.10	0.00	0.29	2.78	0.11
Husband	0.017	**0.381**	1.16	**54.23**	0.77	0.60	17.79	0.21
Jointly	0.022	0.315	0.15	1.20	0.02	-1.03	**69.12**	**0.98**

Correspondence Plot

First two PCs are retained to develop the bi-plot (2d-graph) and three PCs for the tri-plot (3D-plot) as they account the maximum variation among the k components. For example, in case of household's data the bi-plot is explaining 88.60 per cent of variation of data and it can be visualized from the plot that the spread of points in horizontal direction is relatively more than vertical direction, which justifies 48.69 and 39.91 per cent of variation of PC1 and PC2 respectively (Fig 2).

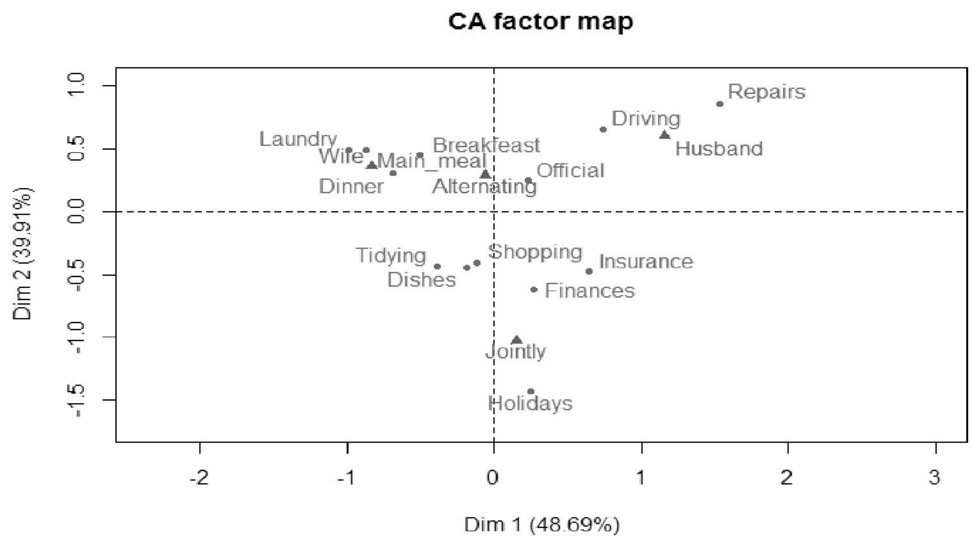

Fig 2: Correspondence plot of correspondence analysis

Inference of Correspondence plot

- Wife and alternating groups together contribute towards laundry, breakfast, main meal and dinner; while Husband and alternating groups together contribute for official, driving and repair tasks. Alternating group is a little biased towards wife group with respect to husband group in doing tasks as it is in the upper left quadrant with wife and highly biased towards wife with respect to jointly group.
- Jointly group is far from wife, alternating and husband group.
- The contribution of alternating is almost negligible compared to wife, husband and jointly group as it is nearest to origin.
- Wife most often do laundry, dinner, breakfast and main meal.
- Husband does official, repairs and driving more often out of which most repairs are done by husband only.
- Jointly they do holidays, shopping, insurance, dishes, tidying and finance tasks more often out of which holidays are most enjoyed jointly.

Finally, it is advised to carefully design and analyze to reach the required objectives to solve any real world problem and the workflow to reach correspondence analysis results is mentioned in the following workflow.

Problem

↓

Identification of objectives

↓

Adequate study design for data analysis

↓

Data

↓

Data collection process

↓

Multivariate analysis (if more than one variable is required to be studied simultaneously)

↓

If data is categorical and categorical variables are interdependent

↓

Correpondence analysis

↓

Results

Fig 3: Workflow for correspondence analysis

REFERENCES

Benzécri, J. P. (1973), *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. Paris, France: Dunod

Other Suggested Readings

Greenacre, Michael (2007), *Correspondence Analysis in Practice, Second Edition*. London: Chapman & Hall/CRC.

Hirschfeld, H.O. (1935), A connection between correlation and contingency. *Proc. Cambridge Philosophical Society*, 31, 520–524.

Härdle W. and Simar L. (2007), *Applied Multivariate Statistical Analysis, Second edition*. Berlin: Springer-Verlag.

Hair, J., W. Black, R. Anderson and B. Babib (2018), *Multivariate Data Analysis, eighth edition*. London: Cengage Learning EMEA

Härdle, Wolfgang Karl & Hlávka, Zdenek (2007), *Multivariate statistics: Exercises and solutions*. New York: Springer-Verlag. doi: 10.1007/978-0-387-73508-5.

http://www.sthda.com/english/rpkgs/factoextra