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

9 788194 008026

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

ICAR

Quantitative Methods for Social Sciences

Vinayak Nikam • Abimanyu Jhajhria • Suresh Pal

Vinayak Nikam ● Abimanyu Jhajhria ● Suresh Pal

Quantitative Methods for SOCIAL SCIENCES

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

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

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 2

CLUSTER ANALYSIS

Arpan Bhowmik, Sukanta Dash, Seema Jaggi and Sujit Sarkar

INTRODUCTION

Statistical science plays a major role in any scientific investigation. Use of appropriate statistical techniques for analyzing the data is very crucial to obtain a meaningful interpretation of the investigation. Throughout any scientific inquiry which is an iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables which leads to the study of multivariate analysis.

Cluster analysis is an important statistical tool with respect to multivariate exploratory data analysis. It involves intricate techniques, methods and algorithms which can be applied in various fields, including economics and other social research. The aim of cluster analysis is to identify groups of similar objects (e.g. countries, enterprises, households) according to selected variables (e.g. unemployment rate of men and women in different countries, deprivation indicators of households, etc.). Cluster analysis is typically used in the exploratory phase of research when the researcher does not have any pre-conceived hypotheses or prior knowledge regarding the similarity of the objects. It is commonly not only the statistical method used, but rather is done in the early stages of a project to help guide the rest of the analysis (Timm, 2002, Hair *et al.*, 2006).

Cluster analysis differs from other methods of classification such as Discriminant analysis where classification pertains to known number of groups and the operational objective is to assign new observations to one of these groups. Whereas cluster analysis is a more primitive tool as in that no assumptions are made about the number of groups or the group structure and the grouping is done based on similarities or distances (dissimilarities).

Cluster analysis is also an important tool for investigation in data mining. For example, in marketing research consumers can be grouped on the basis of their preferences. In short it is possible to find application of cluster analysis in any field of research.

CLUSTERING METHODS

The commonly used methods of clustering are divided into two categories (Johnson and Wichern, 2006).

(i) Hierarchical and

(ii) Non-Hierarchical.

Hierarchical Cluster Analysis

Hierarchical clustering techniques proceed by either a series of mergers or a series of successive divisions. Agglomerative hierarchical method starts with the individual objects, thus there are as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually as the similarity decreases, all sub groups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two sub groups such that the objects in one sub group are far from the objects in the others sub groups. These sub groups are then further divided into dissimilar sub groups. The process continues until there are as many sub groups as objects i.e., until each object form a group. The results of both agglomerative and divisive method may be displayed in the form of two dimensional diagram known as Dendrogram. It can be seen that the Dendrogram illustrate the mergers or divisions that have been made at successive levels.

Linkage methods are suitable for clustering items, as well as variables. This is not true for all hierarchical agglomerative procedures. The following linkage are now discussed:

(i) single linkage (minimum distance or nearest neighbour)

(ii) complete linkage (maximum distance or farthest neighbour)

(iii) average linkage (average distances)

Other methods of hierarchical clustering techniques like Ward's method and Centroid method are also available in literature.

Steps of agglomeration in Hierarchical cluster analysis

The following are the steps in the agglomerative hierarchical clustering algorithm for groups of N objects (items or variables).

i. Start with N clusters, each containing a single entity and an N×N symmetric matrix of distance (or similarities) D = {dik}.

ii. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between most similar clusters U and V be duv.

iii. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters.

iv. Repeat steps (ii) and (iii) a total of N-1 times (all objects will be in a single cluster after the algorithm terminates). Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

The basic ideas behind the cluster analysis are now shown by presenting the algorithm components of linkage methods.

Non Hierarchical Clustering Method

Non Hierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure. Because a matrix of distance does not have to be determined and the basic data do not have to be stored during the computer run. Non hierarchical methods can be applied to much larger data sets than can hierarchical techniques. Non hierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points which will form nuclei of the cluster.

K Means Clustering

The K means clustering is a popular non hierarchical clustering technique. For a specified number of clusters K the basic algorithm proceeds in the following steps (Afifi, Clark and Marg, 2004).

i. Divide the data into K initial cluster. The number of these clusters may be specified by the user or may be selected by the program according to an arbitrary procedure.

ii. Calculate the means or centroid of the K clusters.

iii. For a given case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.

iv. Repeat step (iii) for each case.

v. Repeat steps (ii), (iii), and (iv) until no cases are reassigned.

Dendrogram

Dendrogram, also called hierarchical tree diagram or plot, shows the relative size of the proximity coefficients at which cases are combined. The bigger the distance coefficient or the smaller the similarity coefficient, the more clustering involved combining unlike entities, which may be undesirable. Trees are usually depicted horizontally, not vertically, with each row representing a case on the Y axis, while the X axis is a rescaled version of the proximity coefficients. Cases with low distance/ high similarity are close together. Cases showing low distance are close, with a line linking them a short distance from the left of the Dendrogram, indicating that they are agglomerated into a c a low distance coefficient, indicating alikeness. When, the linking line is to the the Dendrogram the linkage occurs at a high distance coefficient, indicating the clusters are agglomerated even though much less alike. If a similarity measure rather than a distance measure, the rescaling of the X axis still produces a diagram linkages involving high alikeness to the left and low alikeness to the right.

Distance measures

Given two objects X and Y in a 'p' dimensional space, a dissimilarity measure sa the following conditions:

1. $d(X,Y) \geq 0$ for all objects X and Y
2. $d(X,Y) = 0$ $X = Y$
3. $d(X,Y) = d(Y,X)$

Condition (3) implies that the measure is symmetric so that the dissimilarity mea that compares X and Y is same as the comparison for object Y verses X. Condition requires the measures to be zero, when ever object X equals to object Y. The obj are identical if $d(X, Y) = 0$. Finally, Condition (1) implies that the measure is n negative.

Some dissimilarity measures are as follows.

Euclidian distance

This is probably the most commonly chosen type of distance. It is simply the geome distance in the multidimensional space. It is computed as,

$$d(X,Y)= \{\sum_{i=1}^{p}(X_i - Y_i)^2\}^{1/2} \quad \text{or}$$

In matrix form

$$d(X,Y)= d(X,Y)= \sqrt{(X-Y)'(X-Y)}$$

Where $X = (X1, X2, \ldots, Xp)$

$Y = (Y1, Y2, \ldots, Yp)$

The statistical distance between the same two observations is of the form

$$d(X,Y) = \sqrt{(X-Y)'A(X-Y)},$$

where $A = S\text{-}1$ and S contains the sample variances and covariances.

Euclidian and square Euclidian distances are usually computed from raw data and n from standardized data.

Square euclidean distance

Square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as:

$$d^2(X,Y) = \{\sum_{i=1}^{p} |X_i - Y_i|^m\}^{\frac{1}{m}}$$

or in matrix form

$$d^2(X,Y) = (X - Y)' (X - Y)$$

Minkowski metric

When there is no idea about prior knowledge of the distance group then one goes for Minkowski metric. This can be computed as given below:

$$d(X,Y) = \{\sum_{i=1}^{p} |X_i - Y_i|^m\}^{\frac{1}{m}}$$

For m = 1, d(X,Y) measures the city block distance between two points in p dimensions.

For m = 2, d(X,Y) becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

City-block (Manhattan) distance

This distance is simply the average difference across dimensions. In most cases, this distance measure yields result similar to the simple Euclidean distance. This can be computed as :

$$d(X,Y) = \sum_{i=1}^{p} |X_i - Y_i|$$

Chebychev distance

This distance measure may be appropriate in case when we want to define the objects as different if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$d(X,Y) = \text{maximum } |X_i - Y_i|$$

Two additional popular measures of distance or dissimilarity are given by the Canberra metric and the Czekanowski coefficient. Both of these measures are defined for non negative variables only. We have

Canberra metric: $d(X, Y) = \sum_{i=1}^{p} \frac{|X_i - Y_i|}{(X_i + Y_i)}$

Czekanowski coefficient $= 1 - \frac{2\sum_{i=1}^{p} \min(X_i, Y_i)}{\sum_{i=1}^{p} (X_i - Y_i)}$

ILLUSTRATION

Given below is food nutrient data on calories, protein, fat, calcium and iron. The objective of the study is to identify suitable clusters of food nutrient data based on the five variables (Chatfield and Collins, 1990).

Table 1: Food nutrient data on calories, protein, fat, calcium and iron

Food items	Calories	Protein	Fat	Calcium	Iron
1	340	20	28	9	2.6
2	245	21	17	9	2.7
3	420	15	39	7	2
4	375	19	32	9	2.6
5	180	22	10	17	3.7
6	115	20	3	8	1.4
7	170	25	7	12	1.5
8	160	26	5	14	5.9
9	265	20	20	9	2.6
10	300	18	25	9	2.3
11	340	20	28	9	2.5
12	340	19	29	9	2.5
13	355	19	30	9	2.4
14	205	18	14	7	2.5
15	185	23	9	9	2.7
16	135	22	4	25	0.6
17	70	11	1	82	6
18	45	7	1	74	5.4
19	90	14	2	38	0.8
20	135	16	5	15	0.5
21	200	19	13	5	1
22	155	16	9	157	1.8
23	195	16	11	14	1.3
24	120	17	5	159	0.7
25	180	22	9	367	2.5
26	170	25	7	7	1.2
27	170	23	1	98	2.6

Analysis using SPSS

Start by entering the datasheet into SPSS using the steps below.

Step: Go to file→ open→ browse the datasheet→ click open or

Enter all the data in the data editor as shown in Figure1.

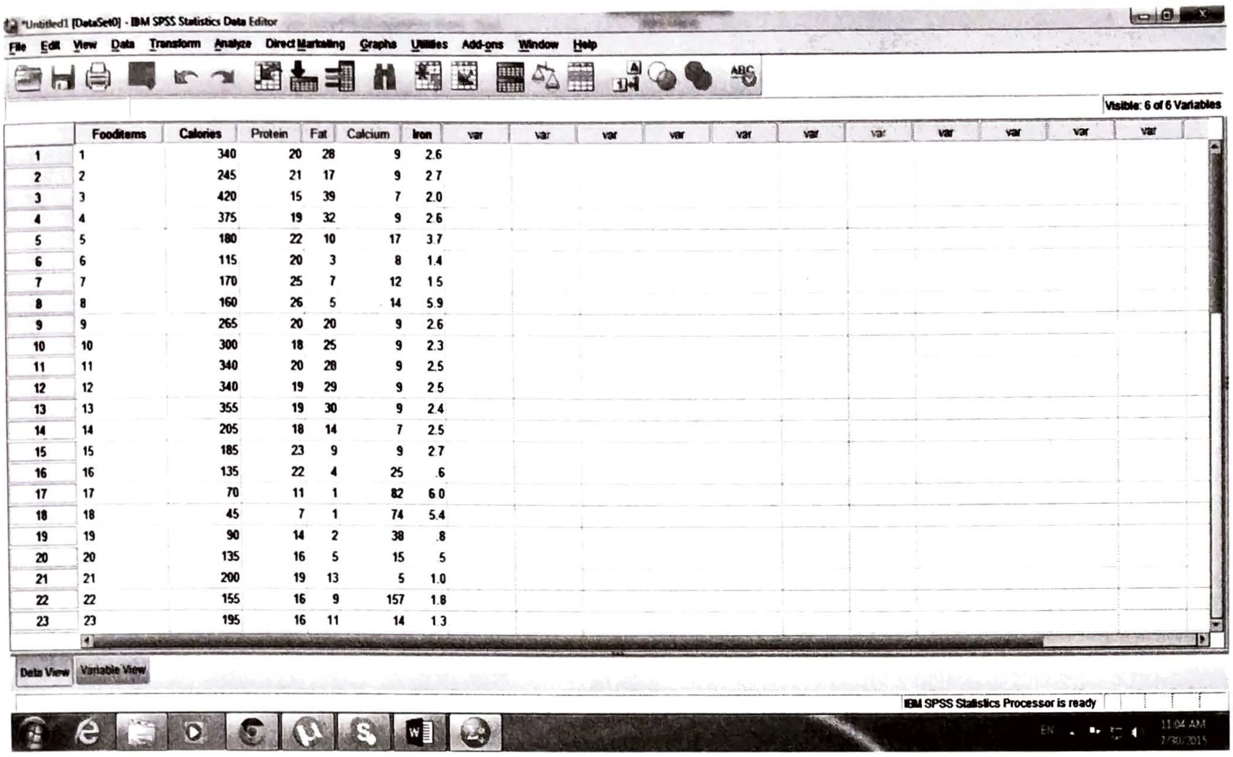


Fig 1: Screen shot after entering the data in data editor

Now click Analyze→ Classify→ Hierarchical Cluster as shown in Figure2.

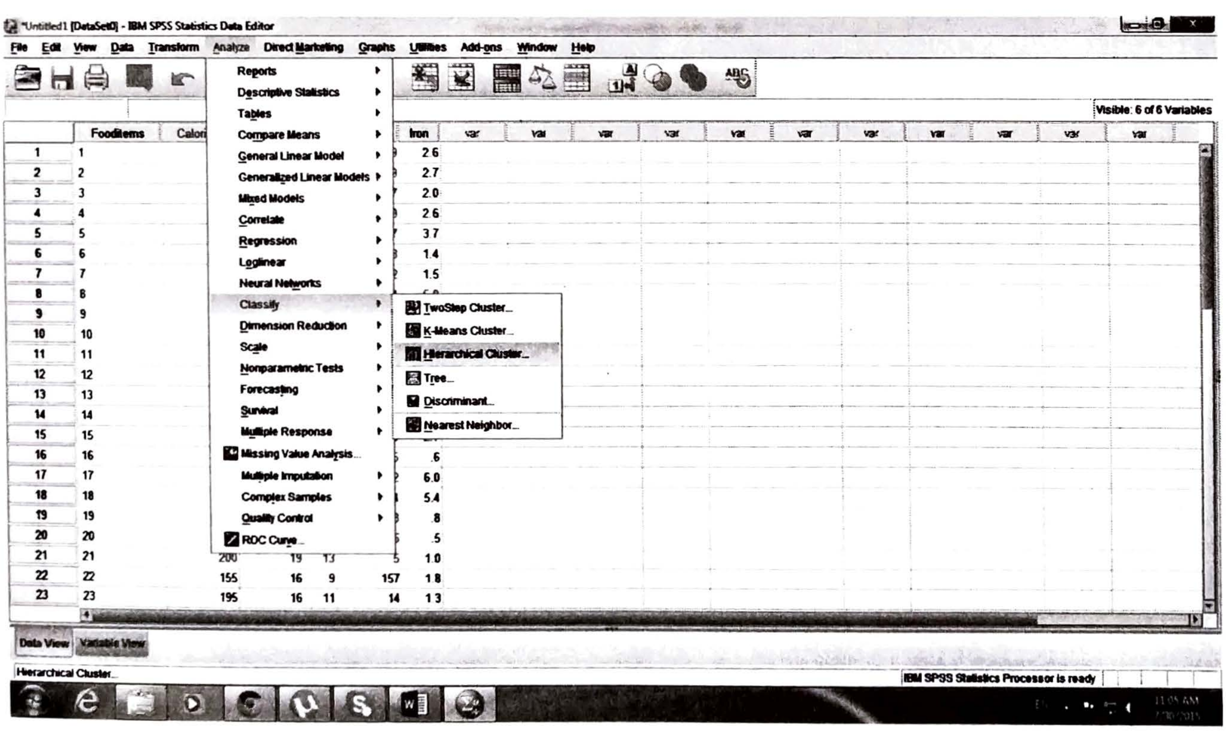


Fig 2: Screen shot of selecting the analysis procedure

Then Identify Name as the variable by which to label cases and Calories, Protein, Fat, Calcium, and Iron as the variables. Indicate that you want to cluster cases rather than variables and want to display both statistics and plots as shown in Fig 3.

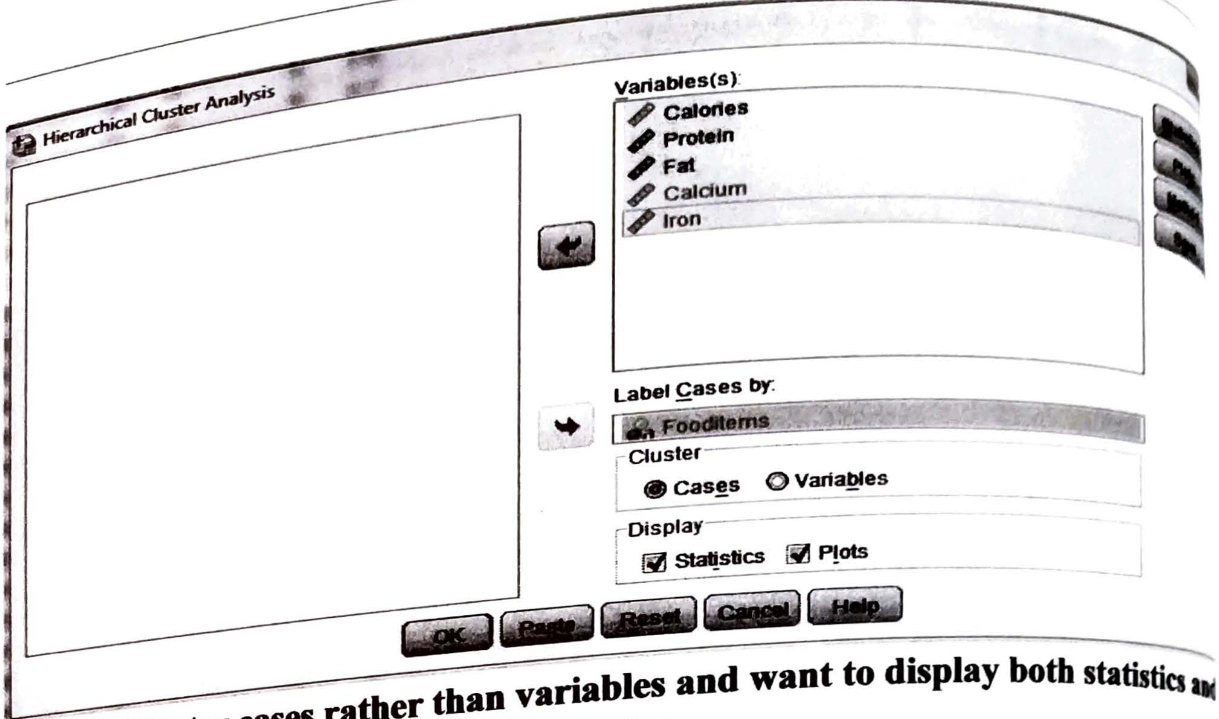


Fig 3: Cluster cases rather than variables and want to display both statistics and plots

Click Statistics and indicate that you want to see an Agglomeration schedule with 2

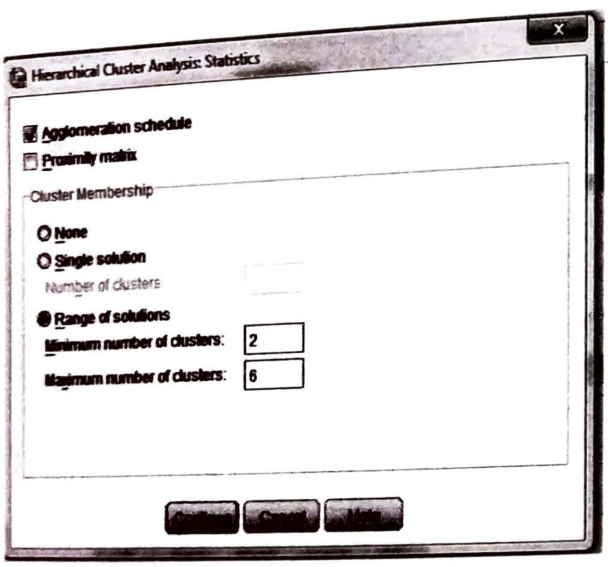


Fig 4: Hierarchical cluster analysis statistics

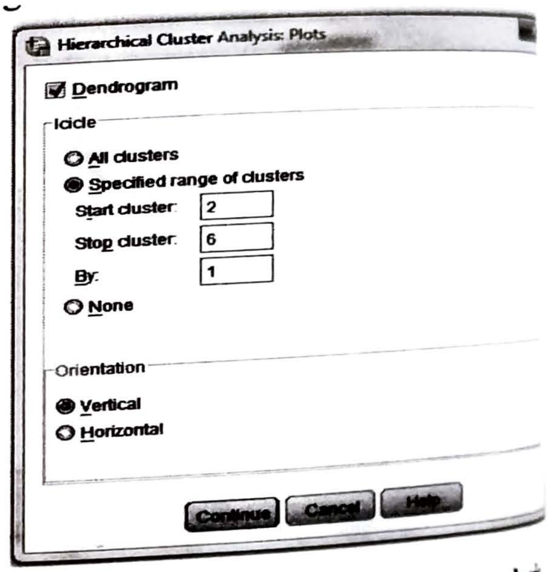


Fig 5: Hierarchical cluster analysis plot

4, and 5 cluster solutions. Click Continue as shown in Fig 4

Click plots and indicate that you want a Dendogram and a verticle Icicle plot with 2 and 4 cluster solutions. Click Continue as shown in Fig 5

Click Method and indicate that you want to use the Between-groups linkage method clustering, squared Euclidian distances, and variables standardized to z scores (so each variable contributes equally). Click Continue as shown in Fig 6.

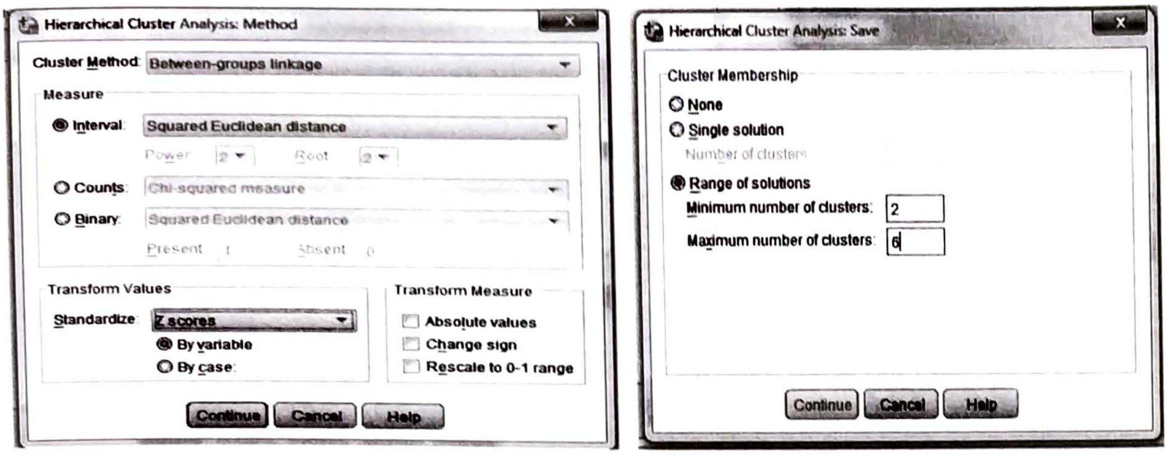
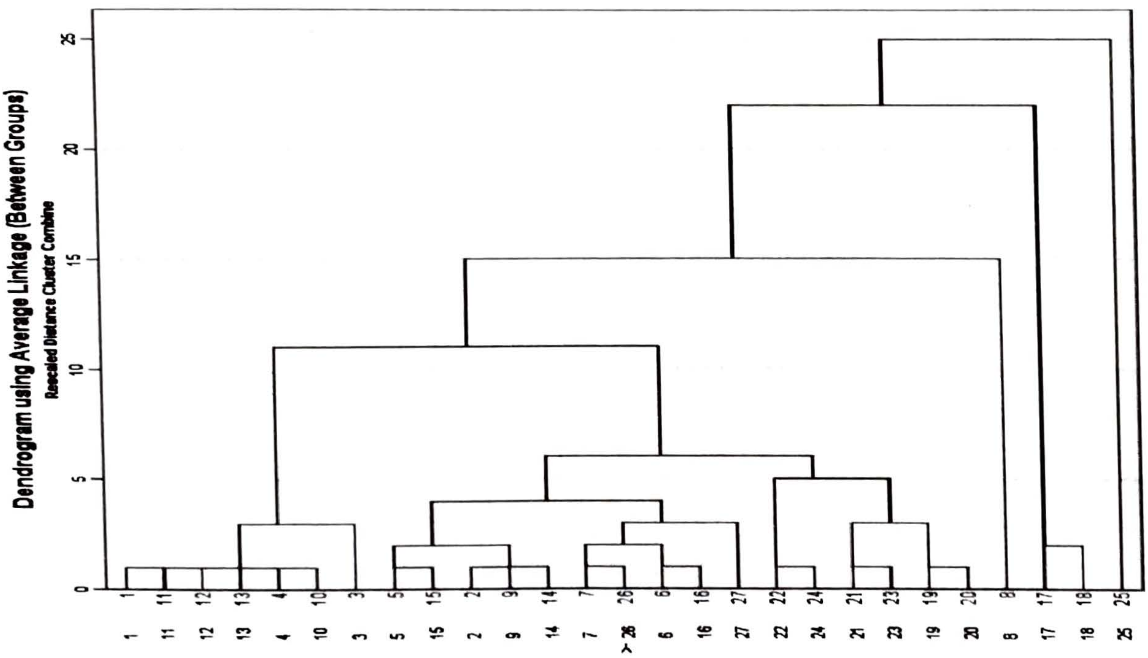


Fig 6: Hierarchical cluster analysis method

Click Save and indicate that you want to save, for each case, the cluster to which the case is assigned for 2, 3, 4, 5 and 6 cluster solutions. Click Continue, OK as shown in Fig 7

SPSS starts by standardizing all of the variables to mean 0, variance 1. This results in all the variables being on the same scale and being equally weighted.

Dendrogram



Interpretation

The main objective of our analysis is to group the food items on the basis of their nutrient content based on the five variables such that food items with in the groups are homogeneous and between the groups are heterogeneous.

Table 2: Interpretation

Number of groups	Food items
Two groups	Group-1 (1,11,12,...,18) Group-2 (25)
Three groups	Group-1 (1,11,...,8) Group-2 (17,18) Group-3 (25)
Four groups	Group-1 (1,11,...,20) Group-2 (8) Group-3 (17,18) Group-4 (25)
Five groups	Group-1 (1,11,...,3) Group-2 (5,15,...,20) Group-3 (8) Group-4 (17,18) Group-5 (25)
Six groups	Group-1 (1,11,...,3) Group-2 (5,15,...,27) Group-3 (22,24,...20) Group-4 (8) Group-5 (17,18) Group-6 (25)

Illustration (Using survey data from social science)

Given below is a part of the data based on a study which was conducted to understand the socio-economic implication of climate and vulnerability of farmers in arid ecosystem of Rajasthan by Sarkar (2014). Two districts Jodhpur and Jaisalmer were selected from arid ecosystem and 100 farmers were selected randomly for the present study. However, for the present chapter, in order to demonstrate the similarity in terms of adaptive behaviour of the farmers, the cluster analysis was performed by considering variables like awareness, attitude towards climate change, egalitarianism, risk perception w.r.t. 20 farmers.

Table 3: Illustration

Farmers' ID	Awareness	Attitude	Egalitarism	Risk perception
1	26	60	37	60
2	18	43	25	58
3	25	67	40	65
4	23	53	34	57
5	20	41	37	41
6	16	37	38	50
7	23	60	38	65

Farmers' ID	Awareness	Attitude	Egalitarism	Risk perception
8	19	41	27	50
9	23	41	26	64
10	26	61	37	60
11	18	48	25	59
12	26	67	40	67
13	23	53	35	57
14	20	41	37	41
15	16	37	38	48
16	25	59	38	66
17	19	40	27	50
18	23	42	27	64
19	26	68	36	61
20	16	42	25	58

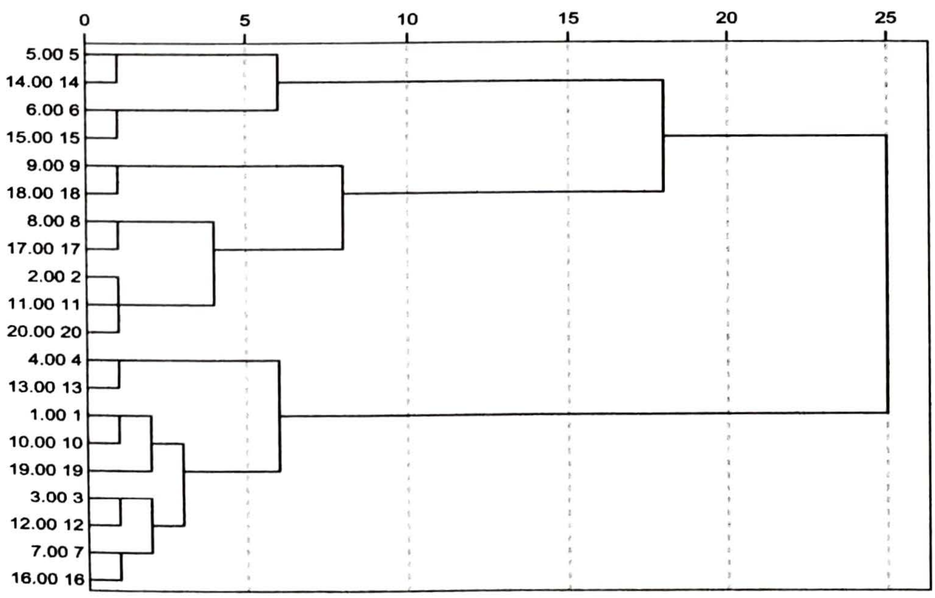
Here, the purpose of the cluster analysis is to group the farmer based on their adaptive behaviour so that appropriate action can be suggested for the farmers who are lagging behind. Two groups were formed viz. adaptors and non-adaptors. The results are summarized as follows:

Table 4: Adopters and non-adopters

Groups	Farmers' ID
Adopters	1,3,4,7,10,12,13,16 and 19
Non-adopters	2,5,6,8,9,11,14,15,17,18,20

Dendogram

Dendrogram using Average Linkage (Between Groups) Rescaled Distance Cluster Combine



REFERENCES

Afifi, A., V. A. Clark and S. Marg (2004), Computer Aided Multivariate Analysis. CRC Press, USA.

Chatfield, C. and A. J. Collins (1990), Introduction to Multivariate Analysis. Chapman and Hall Publications.

Johnson, R. A. and D. W. Wichern (2006), Applied Multivariate Statistical Analysis. 5th Edn., London, Inc. Pearson Prentice Hall.

Sarkar, S. (2014), Assessment of Climate Change Led Vulnerability and Simulating the Adaptive Behaviour of Farmers in the Himalayan and Arid Ecosystems. Ph.D. Thesis, IARI, New Delhi.