International Journal of Social Research Methodology

ISSN: 1364-5579 (Print) 1464-5300 (Online) Journal homepage: https://www.tandfonline.com/loi/tsrm20

Clustering in the field of social sciences: that is your choice

Jaime R.S. Fonseca

To cite this article: Jaime R.S. Fonseca (2013) Clustering in the field of social sciences: that is your choice, International Journal of Social Research Methodology, 16:5, 403-428, DOI: 10.1080/13645579.2012.716973

To link to this article: https://doi.org/10.1080/13645579.2012.716973

International Journal of Social Research Methodology, 2013
Vol. 16, No. 5, 403–428, http://dx.doi.org/10.1080/13645579.2012.716973

Routledge
Taylor & Francis Group

Clustering in the field of social sciences: that is your choice

Jaime R.S. Fonseca*

Higher Institute of Social and Political Sciences, CAPP-Centre for Public Administration and Policies, Technical University of Lisbon, Lisbon, Portugal

(*Received 2 March 2012; final version received 25 July 2012*)

Clustering seeks to identify a finite set of clusters to describe data. *Cluster analysis* is partitioning similar objects into meaningful classes, when both the number of classes and their composition are to be determined. Nowadays, we often see illustrations concerning the use of latent class models (LCM) in the field of cluster analysis. They provide a useful probabilistic/statistical method for grouping observations into clusters. In this approach to clustering, each different cluster in the population is assumed to be described by a different probability distribution, which may belong to the same family but differ in the values they take for the parameters of the distribution. The goal of this research is cluster analysis and LCM comparison, and methodologically we considered three datasets: one with solely continuous variables, one with only binary variables and one with mixed variables. In all situations, LCM performed reasonably well; in contrast, cluster analysis achieved both the best (90.7%, only continuous variables) and the worst performance (40%, mixed variables).

Keywords: clustering techniques; hierarchical cluster methods; latent class models; model selection; good number of clusters

1. Introduction

Clustering seeks to identify a finite set of clusters to describe data and cluster analysis is a collection of statistical methods that groups similar objects into homogeneous groups (or clusters). Cluster analytic methods have the ability to rearrange the data, so researchers are more aware of pattern recognition and discovery.

Cluster analysis is the partitioning of similar objects into meaningful classes, when both the number of classes and their composition are to be determined, (Kaufman & Rouusseeuw, 1990). The starting point for cluster analysis is an $n \times n$ similarity matrix whose cells contain indices that for all n objects show the similarity of each and every object with each and every other object on a number of observed variables; the purpose of cluster analysis is to find mutually exclusive groups (clusters, types) of objects in such a way that the objects belonging to the same cluster are as similar as possible and objects belonging to different clusters are as different from each other as possible, (Hagenaars & Halman, 1989). The number of clusters to be retained has traditionally been the Achilles' heel of cluster analysis; one of the more promising ways to identify the appropriate number of clusters to keep is based on a replication and cross-validation procedure (Mandara, 2003).

*Email: jaimefonseca@iscsp.utl.pt

One of the primary assumptions underlying this standard method for calculating distance is that the variables used to classify individuals into groups are continuous in nature, (Anderberg, 1973). Unfortunately, in practice, most of the time, data-sets are characterized by mixed data, which means that they describe individuals by means of both categorical and continuous variables, mostly categorical.

The goal of this article is to introduce discussion on the use of cluster analysis or latent class models (LCM) in social research by using illustrative examples. Thus, we aim to compare hierarchical cluster analysis (HCA) and LCM in terms of simplicity, survey size and accuracy.

2. Usefulness of cluster and latent class analysis in social sciences

In an effort to validate cluster analysis, Clements (1954) applied this technique to the same coefficients which (Kroeber, 1939) had previously clustered by inspection, and his results generally agreed with those obtained by Kroeber; this was viewed as a certain validation of cluster analysis. If clustering procedures parallel procedures utilized in broader society, then an analysis of clustering techniques can yield insights into the manner in which social groups are formed; instead of merely using cluster analysis as a method of studying society, the method itself can be studied on the assumption that it is a microcosm or model of the clustering processes utilized in society at large (Bailey, 1983).

Within the social sciences, cluster analysis has appeared frequently in sociology but not in political science or economics, (Ahlquist & Breunig, 2009). Typological theorizing, for example, has a distinguished tradition in social sciences (Elman, 2005; George & Nennett, 2005). A typology is a hierarchical system of categories used to organize objects according to their similarities and dissimilarities. Therefore, typologies can be either theoretical or numerical, (McQuitty, 1987) and numerical typologies are the predominate approach in the social sciences. Sucoff and Upchurch (1998) analyzed data from a special release of the Panel Study of Income Dynamics that appends census tract information to the individual records of 940 metropolitan black women; using cluster analysis, they created neighbourhood types that reflect the racial and economic composition of neighbourhoods where metropolitan blacks live. Aldenderfer and Blashfield (1984) created neighbourhood typologies using average-linkage cluster analysis of the racial composition and socioeconomic status variables, the major dimensions of neighbourhood stratification; moreover, they selected the average-linkage clustering procedure over other clustering methods because it produced the most internally homogeneous clusters (e.g. all neighbourhoods in one cluster were poor and black, whereas neighbourhoods in another cluster were relatively affluent and white).

Following Vanneman (1977), the family of techniques known as HCA is especially appropriate to the analysis of stratification systems; the attractive feature of hierarchical schemes is that, with these methods, each cluster is itself a combination of smaller clusters; that is, a cluster need not be seen as perfectly homogeneous but can be broken down into its constituent groups. Indeed, typologies play an important role in sociological theory and research; besides Max Weber, many other social scientists from the past and present have created typologies which have had great influence on the directions sociological theory and research have taken, (Hagenaars & Halman, 1989). It seems clear that cluster analysis and sociology are relatively compatible. Not only can cluster analysis be used quantitatively to group

sociological data into relatively homogeneous groups, but the study of clustering algorithms can provide insights into the manner in which members of society are agglomerated or divided into relatively homogeneous social groups (Bailey, 1983). When applied, this means that while we may wish to make a basic distinction between subordinate and dominant classes, we can still recognize internal differentiation within those classes. This is the common method of conceptualizing class structure, (Marx, 1951, p. 62), for example, argues that the French bourgeoisie was divided into the Legitimists (the large landed proprietors) and the Orleanists (the capitalists), with the capitalists themselves divided into the large industrialists and the finance aristocracy; however, in its opposition to the proletariat, the bourgeoisie was itself a single class structure.

It seems fruitful to investigate not only the degree to which cluster analysis may be efficacious in identifying naturally occurring social groups, but also the degree to which the study of clustering procedures themselves can help us understand how groups are formed in larger society, (Bailey, 1983). Although favoured by some nineteenth century sociologists, such as Spencer (1864), the explicit study of societal evolution was out of favour in sociology for quite some time before making a mild resurgence in the 1960s (Bellah, 1964; Parsons, 1966). In particular, the question of embourgeoisement and proletarianization can benefit from the recent advances in cluster analysis techniques (Anderberg, 1973; Bailey, 1975; Everitt, 1974; Sneath & Sokal, 1973) have now made these advances more available to sociologists. The ability to identify a natural social group through clustering procedures depends, at least partly, on the degree to which the clustering algorithm used replicates the classification procedures used originally to construct the homogeneous natural social group, (Bailey, 1983). Friedkin (1978) used HCA, plus department literature about the research pursuit of faculty members, to define clusters of faculties with relatively homogeneous research interests. A few analysts have generated groupings more or less inductively, although the most common first-stage analytic strategy by far is cluster analysis. Perhaps more to the point, given the reservations many sociologists have about cluster analysis, a number of methods of cluster validation have been attempted (Abbot & Tsay, 2000).

With regard to gambling, for example, the first approach is to view gambling motivation from a sociological perspective, (Fisher, 1993); because sociology at base pursues research at ‘group’ level, some sociologists have sought to understand why people gamble by focusing on why social groups choose different forms of gambling, (Lee, Lee, Bernhard, & Yoon, 2006). The self-perception scores of rejected children were submitted to a HCA, using the average linkage between groups method based on the squared Euclidian distances (Boivin & Bégin, 1989). Comparing sexual attitudes and behaviours across cultures is a concern of anthropological and sociological research, (Widmer, Treas, & Newcomb, 1998); attitudes toward premarital sex, teenage sex, extramarital sex and homosexual sex in 24 countries were compared, and cluster analysis reveals that there are six groupings of nations with similar moral standards. Arts and Gelissen (2002) summarize several uses of cluster analysis in comparative social policy, on the theme of welfare state regimes. Bambra (2007) used cluster analysis to build upon previous research and resurrect the concept of defamilization, and in contrast to existing work in this area, the analysis produced a five-fold typology of welfare state regimes.

While sociologists’ ability to study processes and causal models has improved dramatically in the past two decades, the mathematical analysis of structures has

only begun to develop, and cluster analysis, particularly HCM, holds great promise, (Vanneman, 1977). However, important practical questions that arise in cluster analysis, such as how many clusters there are and which clustering method should be used, remain unanswered, (Vanneman, 1977); moreover, the statistical properties of these methods are generally unknown, precluding the possibility of formal inference. Despite the frequent use of cluster analysis, particularly for the marketing researcher, little is known about the characteristics of available clustering methods or how clustering methods should be employed, (Punj & Stewart, 1983). Worse still, because there are several clustering methods and proximity measures, for each combination (method, proximity measure) cluster analysis output (dendogram) is quite different, which offers a number of different solutions. Written at differing levels of detail, several sources on HCA are available, such as (Aldenderfer & Blashfield, 1984; Anderberg, 1973; Everitt, 1974; Kaufman & Rouusseeuw, 1990; Nunnally & Bernstein, 1994), for a more mathematically rigorous presentation in theoretical terms.

Latent cluster analysis (LCA) was introduced in sociology in 1950 by Lazarsfeld, who used the technique as a tool for building typologies (or clustering), based on dichotomous observed variables, and which is conceptually similar to cluster analysis; it identifies latent classes based on observed response patterns (Clogg, 1995; Lazarsfeld & Henry, 1968; McCutcheon, 1987). The basic ideas of LCA correspond well with the way social scientists use and define typologies, (Hagenaars & Halman, 1989). LCA is used by Taylor (1983) to examine Philip Converse's propositions about political opinion and non-opinion in the American public. Results support Converse's black-and-white model of attitude stability, which posits the existence of some stable opinion holders and a group of 'non-opinion' respondents, whose positions vary unpredictably over time.

Intergenerational support takes many forms among contemporary American families, including the giving and receiving of money and material resources, care, household assistance, companionship and advice, (Bellah, 1964; Hogan, Eggebeen, & Clogg, 1993) as demonstrated here, these data can be much more simply analyzed by first determining the systematic latent structure of intergenerational exchange that characterizes American families. Silverstein and Bengtson (1997) investigate the structure of intergenerational cohesion by examining social-psychological, structural, and transactional aspects of adult child–parent relations, applying to develop a typology based on three underlying dimensions of intergenerational solidarity: affinity, opportunity structure and function. For Evans and Mills (1998), the LCA presented produces a well-fitting model which identifies a set of (latent) social classes; that is, the data is consistent with a representation of the class structure in which the 97 empirically realized response patterns derivable from the combinations of responses to the job characteristic items, are reduced to just four discrete social classes; this typology of four classes, in turn, largely matches the sorts of distinctions embodied in the Goldthorpe schema.

3. Methods of comparison

3.1. Hierarchical cluster analysis

HCA is an exploratory tool designed to reveal natural groupings (or clusters) within a data-set that would otherwise not be apparent, and it is most useful when the aim

is to cluster a small number (less than a few 100) of objects; otherwise, dendogram can be quite confusing. Very often, the results of a particular cluster technique are rather vague, in the sense that it is not at all clear how many clusters (types) ought to be chosen. Moreover, there are many different cluster techniques (and proximity measures) and, often, different methods yield different outcomes without there being sound reasons for choosing one particular solution (one method) over another. Also the size of the similarity matrix may pose a problem, even for modern computers and programs, (Hagenaars & Halman, 1989). The traditional technique currently used is HCA, which develops a measure of similarity (dissimilarity) between pairs of actors, based on the network structure given. From such a measure, starting with an empty network of N actors and no ties, the process starts with the pair with strongest similarity, and follows by adding ties between pairs of vertices, in order of decreasing similarity. Agglomerative hierarchical clustering models form an initial partition of N clusters (each object is a cluster) and, in stages, proceed to reduce the number of clusters, one at a time, until all N objects are in one cluster. In the first stage, $N-1$ clusters are formed by enumerating the possible fusions of N fusions two at a time and selecting the one which optimizes the chosen criterion; in the second stage, $N-2$ clusters are formed in a similar manner and so on. All hierarchical models can be characterized by the set of partitions (P_0, P_1,…, P_{N-1}) and their corresponding criterion values α_1, α_2,…, α_{N-1}, where the stages 0, 1, …, $N-1$ correspondingly represent N, $N-1$, …, 1 clusters. For partition P_j the associated configuration of clusters can be represented by C_1, C_2, …, C_{N-j}. In this context, a stopping rule which selects the *best* number of clusters based on the distribution of a clustering criterion associated with each hierarchical level is desired. Rules are operational rather than representative of some hypothesized or derived density function, (Mojena, 1977).

HCA begins by separating each object into a cluster by itself. At each stage of the analysis, the criterion by which objects are separated is relaxed, in order to link the two most similar clusters until all of the objects are joined in a complete classification tree (dendogram), which is a graphical summary of the cluster solution. Cases are listed along the left vertical axis, for example, and the horizontal axis shows the distance between clusters when they are joined. One begins by looking for ‘gaps’ between junctions along the horizontal axis. A significant change from one stage to the next implies a partition which should not be undertaken where the classification tree to determine the number of clusters is a subjective process. A good cluster solution sees a sudden jump (gap) in the distance coefficient; the solution before the gap indicates the good solution.

In order to show this rule and how the choice of different clusters would affect an example of a cluster analysis procedure in social science, we will use the example used by Widmer et al. (1998), when comparing sexual attitudes and behaviours across cultures, which is a concern of anthropological and sociological research. From 1–24, the data-set includes the following countries: Australia, Austria, Bulgaria, Canada, the Czech Republic, Germany (East), Germany (West), Great Britain, Hungary, Ireland, Israel, Italy, Japan, the Netherlands, New Zealand, Northern Ireland, Norway, Philippines, Poland, Russia, Slovenia, Spain, Sweden and the USA. They apply cluster analysis and they found six clusters (Table 1). We will present the results of dendograms (Figure 1) for the following combinations method/measure: (1) average linkage (within groups)/Minkowski measure, (2) complete linkage/Chebychev and (3) Ward/squared Euclidean distance and clusters in Table 1.

Table 1. Clusters for the three combinations.

Method/ measure	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Average linkage (within groups)/ Minkowski measure	Czech Republic; Germany (West); the Netherlands; Norway and Slovenia	Ireland; New Zealand and the USA	Austria; Australia; Bulgaria; Canada; Great Britain; Hungary; Israel; Italy; New Zealand; Poland; Russia and Spain	Germany (East) and Sweden	Philippines	Japan
Complete linkage/ Chebychev	Germany (East); Germany (West); the Netherlands; Norway; Slovenia and Sweden;	Ireland; Northern Ireland and the USA	Australia; Bulgaria; Hungary; Italy; New Zealand; Poland and Russia	Austria; Canada; Czech Republic; Great Britain; Israel and Spain	Philippines	Japan
Ward/squared Euclidean distance	Germany (East); Germany (West); the Netherlands; Norway; Slovenia and Sweden	Ireland; Northern Ireland and the USA	Australia; Bulgaria; New Zealand; Hungary; Israel; Italy; Poland; Russia and Spain	Austria; Canada; Czech Republic and Great Britain	Philippines	Japan
Widmer et al. (1998)	Germany (East), Germany (West), Austria, Sweden and Slovenia	The USA, Ireland, Northern Ireland and Poland	The Netherlands, Norway, Czech Republic, Canada and Spain	Australia, Great Britain, Hungary, Italy, Bulgaria, Russia, New Zealand and Israel	Philippines	Japan

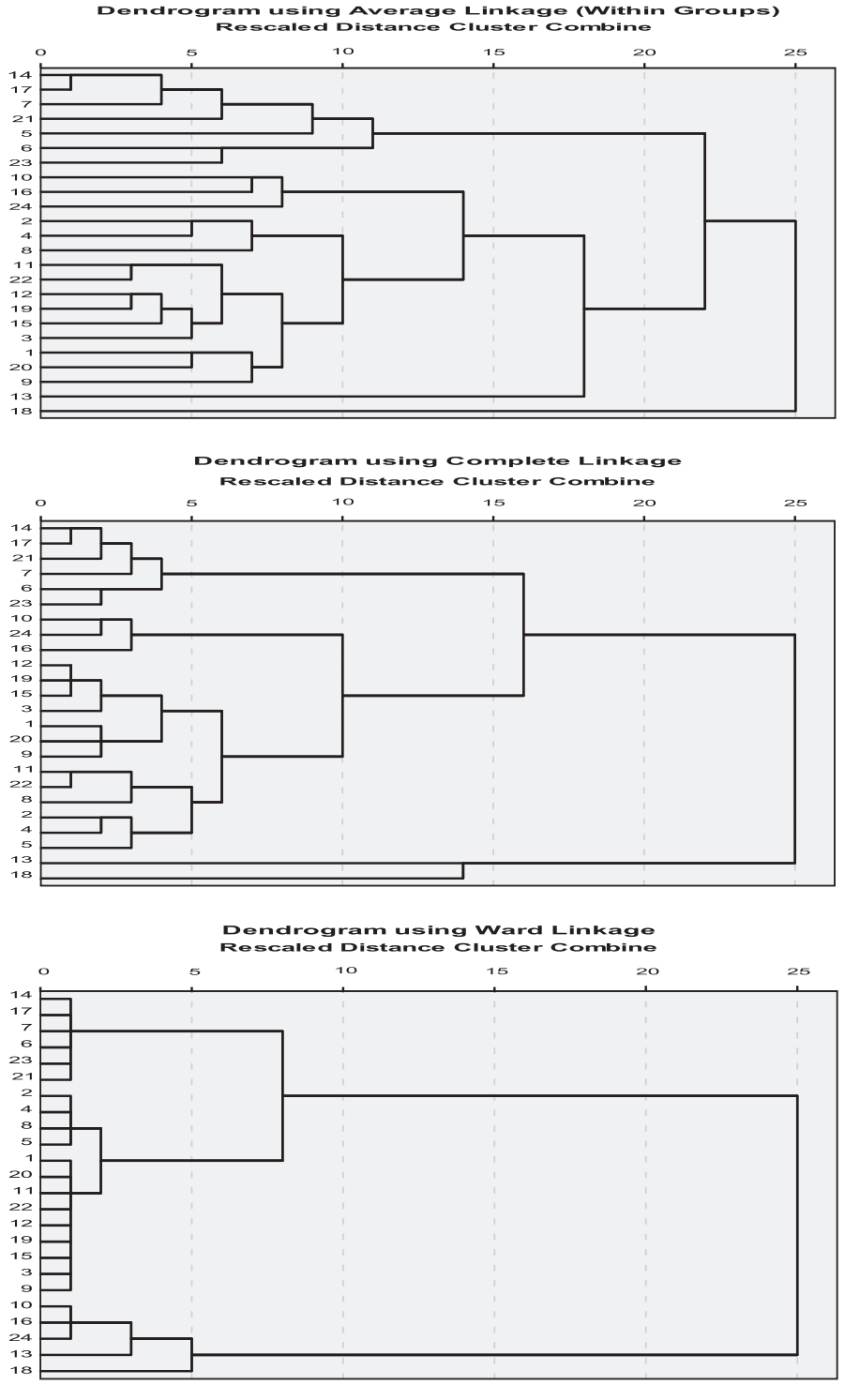

Figure 1. Dendograms (1) average linkage (within groups)/Minkowski measure, (2) complete linkage/Chebychev and (3) Ward/squared Euclidean distance.

3.2. Latent class model

In model-based clustering, it is assumed that the objects under study are generated by a mixture of probability distributions, with one component corresponding to each class, (Zhang, 2003). These statistical models allow us to test if a group of unobserved classes (latent) conveniently justifies the association among the observed variables. In this context, a specific solution, constituted by a group of latent classes, is reasonable when it leads to the minimization of the association among observed variables, within each class. This minimization is the result of the basic assumption of independence or conditional independence.

As such, postulating a heterogeneous population, constituted by S groups or homogeneous sub-populations, the LCM is defined by the variable Y with S categories or latent types of students, described through the observed variables, X_1, X_2, …, X_p, with I_1, …, I_p categories, respectively. Let $\lambda_{i1i2, \ldots, iP}$ be the probability for a certain individual to belong to the categories (i_1, i_2, …, i_P), relatively to the conjoint variable (X_1, X_2, ..., X_p), with $i_1 = 1, \ldots, I_1, \ldots, {}_{iP}$. In these conditions, supposing the existence of a latent variable Y, with S categories, explaining the relationships among the observed variables, the probability $\lambda_{i1i2 \ \ldots \ iP}$ can be defined by the model

$$\lambda_{i_1 i_2 \ldots i_P} = \sum_{s=1}^{S} \lambda_Y(s)\ \lambda_{X_1|Y=s}(i_1)\ \lambda_{X_2|Y=s}(i_2)\ \cdots\ \lambda_{X_P|Y=s}(i_P),$$

where $\lambda Y(s)$ represents the probabilities of $Y = s$, probabilities that an individual belongs to the latent class s ($s = 1, \ldots, S$), that is, the probabilities of the latent classes, also designated by relative sizes or mixture proportions, which estimate the likelihood that individuals belong to each one of the classes.

$\lambda_{X_p|Y=s}(i_p)$, $p = 1, \ldots, P$, represents the conditional probability that the variable Xp is in the category ip, knowing that the latent variable Y is on level s. In estimating LCM, the estimates of the probabilities of the latent classes or relative sizes and certain individual's conditional probabilities are of fundamental importance in their structure, to take values in certain categories of the observed variables, given that it is a member of a class of the latent variable.

The proportions of the latent classes describe the distribution of probability of the latent classes or typologies; they become useful in the description of the typologies' prevalent within the population and in the comparison of those prevalent among sub-populations.

For a more complete description of estimation of LCM, see (Fonseca, 2010, 2011; Fonseca & Cardoso, 2007; McLachlan & Peel, 2000).

In relation to methodologies for selecting the appropriate LCM, we propose traditional information criteria: we used the information criterion Akaike's information criterion (AIC_3), which is most suitable for data-sets with only categorical variables and data-sets with mixed variables, and Bayesian information criterion (BIC), the most appropriate for a data-set with solely continuous variables, (Fonseca, 2010; Fonseca & Cardoso, 2007). Therefore, we will select a model that is the best for presenting the minimum value for AIC_3 (BIC) or an elbow, instead.

We retain six clusters in the three situations, in order to be comparable with Widmer, Treas and Newcomb's results, and we can see several differences in cluster composition.

4. Methodology

In terms of HCA and LCM comparison, we will apply these models to uncover the data pattern in an understandable way. In relation to cluster analysis, we intend to use several combinations of clustering methods and (dis)similarities measures; for latent class analysis, we will use information criteria for model selection.

In order to obtain comparable results, we will use real data-sets with known number of clusters. Three data-sets were used in this research: (1) Fisher's data-set with only continuous variables; (2) Store, data-set with only categorical variables and (3) North Central Wisconsin, data-set with continuous and categorical variables (mixed data). These data-sets were selected because of the knowledge we have concerning the true number of clusters, 3, 2, 3, respectively.

There are several quite different methods for extracting communities, and Statistical Package for Social Sciences-18 provides a clustering programme that implements a variety of hierarchical agglomerative procedures: when applying cluster analysis for each data-set, we will use three proximity measures (Squared Euclidean distance, Chebychev distance and Minkowski distance for both continuous and mixed data-sets and Binary squared Euclidean distance, pattern differences and Rogers Tanimoto for categorical data-set, Table 2) with each one of the four clustering methods (average linkage (between groups), average linkage (within groups), complete linkage and Ward method), based on Milligan's (1981) study.

Because several solutions are possible, at least one for each combination method/dissimilarity measure, hierarchical clustering may not be a good choice in order to detect community structure. We suppose that for all combinations (method and measure) cluster analysis identifies the true number of clusters.

5. Data analysis and discussion of results

Broadly speaking, the typical procedure for conducting a cluster analysis method includes the following steps: select a sample of entities; define a set of variables according to which the entities are measured, select a combination (cluster method and proximity measure) to use; calculate measures of proximity between all the entities; group the entities together based on their proximity scores using the selected clustering algorithm and create a graphic depiction of the groupings that emerge, in order to facilitate interpretation of the results. Choosing the variables for the clustering is as fundamental a problem to cluster analysis as the choice of method. We applied the two methods to data-sets with the same set of variables – clustering variables.

Table 2. Proximity measures used.

Measure	Summary description
Squared Euclidean distance	$\sum_{k\neq i,j}(x_{i,k}-y_{j,k})^2$
Chebychev distance	$d_{ij}=\max_k\lvert x_{i,k}-y_{j,k}\rvert$
Minkowski distance	$p\sqrt{\sum_{k\neq i,j}(x_{i,k}-y_{j,k})^2}$
Binary squared Euclidean distance	$\mathrm{BSEUCLID}_{(x,y)}=b+c$
Pattern difference	bc/($n^{**}2$) (from a fourfold table)
Rogers and Tanimoto	$\mathrm{RT}_{(x,y)}=(a+d)/(a+d+2(b+c))$

By reformulating cluster analysis as a problem in estimation for mixed distributions or LCM, no 'similarities' or 'distances' need to be assumed a priori, (Wolf, 1970) the closest analogy to a 'similarity' in mixed distributions is the probability of membership of a point in a cluster; however, this probability is the result of an iterative solution to the likelihood equations rather than an arbitrarily given function.

5.1. Fisher's continuous data-set

We start this study with a continuous data-set, Fisher's data-set. We will analyze data by means of cluster analysis and LCM. We intend to display the matrix confusion or cross-tabs for true and estimated number of clusters (Tables 3–15), in order

Table 3. Class × average linkage (between groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	38	12	50
Total	50	88	12	150

Table 4. Class × average linkage (within groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	23	27	50
Class 3	0	49	1	50
Total	50	72	28	150

Table 5. Class × complete linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	23	27	50
Class 3	0	49	1	50
Total	50	72	28	150

Table 6. Class × ward linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	49	1	50
Class 3	0	15	35	50
Total	50	64	36	150

Table 7. Class × average linkage (between groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	40	10	50
Total	50	90	10	150

Table 8. Class × average linkage (within groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	14	36	50
Total	50	64	36	150

Table 9. Class × complete linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	21	29	50
Class 3	0	44	6	50
Total	50	65	35	150

Table 10. Class × ward linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	15	35	50
Total	50	65	35	150

Table 11. Class × average linkage (between groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	14	36	50
Total	50	64	36	150

Table 12. Class × average linkage (within groups).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	23	27	50
Class 3	0	49	1	50
Total	50	72	28	150

Table 13. Class × complete linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	23	27	50
Class 3	0	49	1	50
Total	50	72	28	150

Table 14. Class × ward linkage.

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	50	0	50
Class 3	0	14	36	50
Total	50	64	36	150

Table 15. Class × LCM (89.3%).

	Cluster			
Count	1	2	3	Total
Class 1	50	0	0	50
Class 2	0	48	1	50
Class 3	0	14	36	50
Total	50	62	37	150

to know the percentage correctly classified by each one of the used combinations (method and measure) and LCM, for comparison.

In this study, we intend to compare the two techniques only in terms of classification, not in terms of model selection. So we apply cluster analysis where the number of clusters is known; in relation to LCM, we use information criteria, BIC or AIC_3 for model selection.

5.1.1. Squared Euclidean distance

Firstly, we used squared Euclidean distance with average linkage method, trying to uncover the known pattern (true number of classes: three). By using the cluster

analysis outcomes, we ran a cross-tab (Table 3) with *class* (true number of clusters) and *clusters* (the number of clusters by cluster analysis). As we can see from this result, cluster analysis correctly classified 74.7% of cases.

Secondly, we used squared Euclidean distance with average linkage method and we can see from this result that cluster analysis correctly classified 49.3% of cases (Table 4).

Thirdly, we used squared Euclidean distance with complete linkage method and we can see from this result that cluster analysis correctly classified 49.3% of cases (Table 5).

Next, we used squared Euclidean distance with the Ward linkage method and we can see from this result that cluster analysis correctly classified 89.3% of cases (Table 6).

5.1.2. Chebychev distance

Firstly, we used Chebychev distance with average linkage method, in order to uncover the known pattern. With the cluster analysis outcomes, we ran a cross-tab (Table 7) with *class* and *clusters*. As we can see from this result, cluster analysis correctly classified 73.3% of cases.

Secondly, we used Chebychev distance with average linkage method and we can see from this result that cluster analysis correctly classified 90.7% of cases (Table 8).

Thirdly, we used Chebychev distance with complete linkage method and we can see from this result that cluster analysis correctly classified 51.3% of cases (Table 9).

Lastly, we used Chebychev distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 90% of cases (Table 10).

5.1.3. Minkowski distance

Firstly, we used Minkowski distance with average linkage method, in order to uncover the known pattern. With the cluster analysis outcomes we ran a cross-tab (Table 11) with *class* and *clusters*. As we can see from this result, cluster analysis correctly classified 90.7% of cases.

Secondly, we used Chebychev distance with average linkage method and we can see from this result that cluster analysis had correctly classified 49.3% of cases (Table 12).

Thirdly, we used Chebychev distance with complete linkage method and we can see from this result that cluster analysis correctly classified 49.3% of cases (Table 13).

Lastly, we used Chebychev distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 90.7% of cases (Table 14).

5.1.4. Latent class model

Here, we estimated LCM. We started the process with the baseline model (homogeneity) and proceeded with the estimation of 2–4 LCM. The BIC information criterion (used in this case because all the indicators are continuous) selected a three LCM. From the confusion matrix (Table 15), we can see that this model correctly classified 89.3% of cases.

First of all, we can highlight the wide range of percentages of recovering the cluster structure of data: from 49.3 to 90.7%. Even for data-sets with only continuous variables, some combinations (method and measure) offer excellent results, others very poor ones. The best results are associated with Ward linkage method.

In relation to LCM outcomes, this is very similar to the best cluster analysis result (89.3%).

5.2. Store – categorical data-set

This is a data-set with binary variables, and there are several measures concerning the use of cluster analysis in these situations. We cannot use measures such as Jaccard, Phi four-point correlation, Sokal and Sneath measures, Yule's measures, Ochiai measure or Anderberg's. As a result, we decided to use binary squared Euclidean distance, pattern difference and Rogers' and Tanimoto measure, as follows.

5.2.1. Binary squared Euclidean distance

Once again, we began with squared Euclidean distance and average linkage method, in order to uncover the known pattern. With the cluster analysis outcomes, we ran a cross-tab (Table 16) with *class* and *clusters*. As we can see from this result, cluster analysis correctly classified only 37.5% of cases.

Second, we used squared Euclidean distance with average linkage method and we can see from this result that cluster analysis correctly classified 42.6% of cases (Table 17).

Third, we used squared Euclidean distance with complete linkage method and we can see from this result that cluster analysis correctly classified 37.5% of cases (Table 18).

Finally, we used squared Euclidean distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 67.3% of cases (Table 19).

Table 16. Store × average linkage (between groups) (37.5%).

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	247	147	394
Total		313	255	568

Table 17. Store × average linkage (within groups).

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	218	176	394
Total		284	284	568

Table 18. Store × complete linkage.

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	247	147	394
Total		313	255	568

Table 19. Store × ward method.

		Cluster		
		Department	Speciality	Total
Store	Department	128	46	174
	Speciality	140	254	394
Total		268	300	568

5.2.2. Pattern difference

Here, we started using the new pattern difference measure and average linkage, which achieved some of the poorest results: only 41.5% of recovering data (Table 20).

Secondly, we used pattern difference measure and average linkage method and we can see from this result that cluster analysis correctly classified only 45% of cases (Table 21).

Thirdly, we used pattern difference measure with complete linkage method and we can see from this result that cluster analysis correctly classified 42.8% of cases (Table 22).

Finally, we used pattern difference measure with Ward linkage method and we can see from this result that cluster analysis correctly classified 37.7% of cases (Table 23).

Table 20. Store × average linkage (between groups).

		Cluster		
		Department	Speciality	Total
Store	Department	124	50	174
	Speciality	282	112	394
Total		522	46	568

Table 21. Store × average linkage (within groups).

		Cluster		
		Department	Speciality	Total
Store	Department	174	0	174
	Speciality	312	82	394
Total		486	82	568

Table 22. Store × complete linkage.

		Cluster		
		Department	Speciality	Total
Store	Department	123	51	174
	Speciality	274	120	394
Total		522	46	568

Table 23. Store × ward method.

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	246	148	394
Total		312	256	568

5.2.3. Rogers' and Tanimoto measure

First, we used the new measure Rogers' and Tanimoto measure and average linkage and they only achieved 37.5% of recovering data (Table 24).

Second, we used Rogers' and Tanimoto measure with average linkage method and we can see from this result that cluster analysis correctly classified only 72.7% of cases (Table 25).

Third, we used Rogers' and Tanimoto measure with complete linkage method and we can see from this result that cluster analysis correctly classified 37.5% of cases (Table 26).

Finally, we used Rogers' and Tanimoto measure with Ward linkage method and we can see from this result that cluster analysis correctly classified 57.6% of cases (Table 27).

Table 24. Store × average linkage (between groups).

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	247	147	394
Total		313	255	568

Table 25. Store × average linkage (within groups).

		Cluster		
		Department	Speciality	Total
Store	Department	174	0	174
	Speciality	155	239	394
Total		366	202	568

Table 26. Store × complete linkage.

		Cluster		
		Department	Speciality	Total
Store	Department	66	108	174
	Speciality	247	147	394
Total		313	255	568

Table 27. Store × ward method.

		Cluster		
		Department	Speciality	Total
Store	Department	128	46	174
	Speciality	195	199	394
Total		320	248	568

5.2.4. Latent class model

From the application of LCM estimation and AIC_3 information criterion for model selection, results show us that this model correctly classified 72% of cases.

To sum up, in relation to Store data-set (with only binary variables), cluster analysis performance ranges from 28.2 (average linkage) to 72.7% (complete linkage); LCM performed relatively well, with 72% of cases correctly classified.

5.3. North central Wisconsin – mixed data-set

This is the most complex example (the mixed case) because there are no studies that suggest the most appropriate measures for this case, in case of cluster analysis. As such, we decided to use the same measures we used for data-sets with continuous variables.

5.3.1. Squared Euclidean distance

Firstly, we use the squared Euclidean distance and average linkage, which only achieved 32.9% of recovering data (Table 28).

Secondly, we used squared Euclidean distance with average linkage method and we can see from this result that cluster analysis correctly classified 40% of cases (Table 29).

Table 28. Stratum × average linkage (between groups).

	Cluster			
	1	2	3	Total
Stratum 1	50	0	0	50
Stratum 2	56	0	0	56
Stratum 3	46	2	1	49
Total	152	2	1	155

Table 29. Stratum × average linkage (within groups).

	Cluster			
	1	2	3	Total
Stratum 1	48	4	0	50
Stratum 2	52	12	0	56
Stratum 3	35	2	2	49
Total	135	18	2	155

Table 30. Stratum × complete linkage.

	Cluster			
	1	2	3	Total
Stratum 1	47	3	0	50
Stratum 2	49	7	0	56
Stratum 3	26	21	2	49
Total	122	31	2	155

Table 31. Stratum × ward method.

	Cluster			
	1	2	3	Total
Stratum 1	44	6	0	50
Stratum 2	42	14	0	56
Stratum 3	15	32	2	49
Total	101	52	2	155

Thirdly, we used squared Euclidean distance with complete linkage method and we can see from this result that cluster analysis correctly classified 36.1% of cases (Table 30).

Finally, we used squared Euclidean distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 38.7% of cases (Table 31).

Cluster analysis performed very poorly with all clustering methods and squared Euclidean distance, ranging from 32.9% (average linkage) to 40% (Ward).

5.3.2. Chebychev distance

Firstly, we used the Chebychev distance and average linkage and they only achieved 32.9% of recovering data (Table 32).

Secondly, we used Chebychev distance and average linkage method and we can see from this result that cluster analysis correctly classified 34.8% of cases (Table 33).

Thirdly, we used Chebychev distance with complete linkage method and we can see from this result that cluster analysis correctly classified 32.9% of cases (Table 34).

Table 32. Stratum × average linkage (between groups).

	Cluster			
	1	2	3	Total
Stratum 1	50	0	0	50
Stratum 2	56	0	0	56
Stratum 3	46	2	1	49
Total	152	2	1	155

Table 33. Stratum × average linkage (within groups).

	Cluster			
	1	2	3	Total
Stratum 1	48	2	0	50
Stratum 2	52	4	0	56
Stratum 3	35	12	2	49
Total	135	18	2	155

Table 34. Stratum × complete linkage.

	Cluster			
	1	2	3	Total
Stratum 1	50	0	0	50
Stratum 2	56	0	0	56
Stratum 3	46	2	1	49
Total	152	2	1	155

Table 35. Stratum × ward method.

	Cluster			
	1	2	3	Total
Stratum 1	44	6	0	50
Stratum 2	42	14	0	56
Stratum 3	15	32	2	49
Total	101	52	2	155

Finally, we used Chebychev distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 38.7% of cases (Table 35).

Again, the performances are very poor when the four clustering methods with Chebychev distance are used.

5.3.3. Minkowski distance

We began using Minkowski distance with average linkage and they achieve only 29% of recovering data (Table 36).

Table 36. Stratum × average linkage (between groups).

	Cluster			
	1	2	3	Total
Stratum 1	44	0	0	50
Stratum 2	42	0	0	56
Stratum 3	15	2	1	49
Total	152	2	1	155

We then proceeded with Minkowski distance with average linkage method and we can see from this result that cluster analysis correctly classified 34.2% of cases (Table 37).

Thirdly, we used Minkowski distance with complete linkage method and we can see from this result that cluster analysis correctly classified 36.1% of cases (Table 38).

Finally, we used Minkowski distance with Ward linkage method and we can see from this result that cluster analysis correctly classified 38.7% of cases (Table 39).

Table 37. Stratum × average linkage (within groups).

	Cluster			
	1	2	3	Total
Stratum 1	48	2	0	50
Stratum 2	52	4	0	56
Stratum 3	37	11	1	49
Total	137	17	1	155

Table 38. Stratum × complete linkage.

	Cluster			
	1	2	3	Total
Stratum 1	47	3	0	50
Stratum 2	49	7	0	56
Stratum 3	26	21	2	49
Total	122	31	2	155

Table 39. Stratum × ward linkage.

	Cluster			
	1	2	3	Total
Stratum 1	44	6	0	50
Stratum 2	42	14	0	56
Stratum 3	15	32	2	49
Total	101	52	2	155

Table 40. Stratum × LCM.

	Cluster			
	1	2	3	Total
Stratum 1	44	6	0	50
Stratum 2	38	18	0	56
Stratum 3	0	0	49	49
Total	82	24	49	155

Once again, cluster analysis performs very poorly, ranging from 29 (average linkage) to 38.7% (Ward).

5.3.4. Latent class model

Now we use LCM estimation with AIC_3 information criterion for model selection, and we can see from the results that this model performs better, correctly classifying 71.6% of cases (Table 40)!

We have just seen that cluster analysis performs worse than LCM in relation to mixed data-sets, whatever the combination used (method and measure).

6. Conclusion

In relation to Fisher's data-set, we can summarize the main findings as follows.

Using squared Euclidean distance leads to similar results both with within groups (Table 4) and complete linkage (Table 5). However, they are quite different from the results achieved with between groups (Table 3) and Ward linkage (Table 6). As for Chebychev's distance, it leads to similar results both with the within groups method (Table 8) and Ward linkage (Table 10). However, they are quite different from the results achieved with between groups (Table 7) and complete linkage (Table 9). In relation to Minkowski, it leads to similar results both with within groups (Table 12) and complete linkage (Table 13). However, they are quite different from the results achieved with between groups (Table 11) and Ward linkage (Table 14), which are similar.

In Table 41, we display all percentages of structure recovery for all data-sets and method/measure combinations.

In relation to Fisher's continuous data-set, we can see that Chebychev distance with the within groups method and Minkowski distance with both between groups and Ward methods achieving the best performance (90.7%), followed by squared Euclidean distance and Ward method, and LCM with 89.3%.

As far as categorical data-set Store is concerned, the best performance goes with Rogers' and Tanimoto measure and the within groups method (72.7%), followed by LCM with 72%.

Finally, as for mixed data-set, North Central Wisconsin, the best performance is obtained by LCM with 71.6%, followed by squared Euclidean distance with the within groups method with 40%.

To sum up, we can conclude that HCA is a very sensitive technique to both the choice of clustering method and the (dis)similarity; LCM perform well in all situations, categorical data-set, continuous data-set and mixed data-set, both in selecting

Table 41. Performances' performance.

Data-set	Method	Measure	Percentage of structure recovery
Fisher's data-set	Between groups	Squared Euclidean distance	74.7
	Within groups		49.3
	Complete linkage		49.3
	Ward		89.3
	Between groups	Chebychev distance	73.3
	Within groups		90.7
	Complete linkage		51.3
	Ward		90
	Between groups	Minkowski distance	90.7
	Within groups		49.3
	Complete linkage		49.3
	Ward		90.7
	Latent class Model		89.3
Store	Between groups	Binary squared Euclidean distance	37.5
	Within groups		42.6
	Complete linkage		37.5
	Ward		67.3
	Between groups	Pattern difference	41.5
	Within Groups		45
	Complete linkage		42.8
	Ward		37.7
	Between groups	Rogers and Tanimoto	37.5
	Within groups		72.7
	Complete linkage		37.5
	Ward		57.6
	Latent class Model		72
North Central Wisconsin	Between groups	Squared Euclidean distance	32.9
	Within groups		40
	Complete linkage		36.1
	Ward		38.7
	Between groups	Chebychev distance	32.9
	Within groups		34.8
			32.9

(*Continued*)

Table 41. (*Continued*).

Data-set	Method	Measure	Percentage of structure recovery
	Complete linkage		
	Ward		38.7
	Between groups	Minkowski distance	29
	Within groups		34.2
	Complete linkage		36.1
	Ward		38.7
	Latent class model		71.6

the true pattern (true number of clusters), and in percentage of recovering, 89.3, 72 and 71.6%, respectively. Moreover, the researcher does not need to choose a method or a measure from the wide range methods and measures available, and interpret several dendograms, in order to become aware the data structure.

Rather than relying on predetermined cut-off points, this multivariate approach assumes an underlying categorical latent variable that determines an individual's class membership.

Unlike HCA, the LCM is model based or probabilistic, which implies that the model can be replicated with an independent sample, (Muthén & Muthén, 2000). These probabilistic/statistical models allow us to test if a group of unobserved classes (latent) justifies the association among the observed variables.

On the issue of data analysis, the LCM approach to clustering offers some advantages when compared to other, more traditional techniques:

(1) An important difference between standard cluster analysis techniques and LC clustering is that the latter is a model-based approach. This means that a statistical and probabilistic model is postulated for the population from which the data sample is obtained. An advantage of using a statistical and probabilistic model is that the choice of the cluster criterion is less arbitrary and the approach includes rigorous statistical tests, (Magidson & Vermunt, 2002).
(2) There is no need to standardize variables. Before performing hierarchical clustering, analysts must standardize variables to have equal variance to avoid obtaining clusters that are dominated by variables having the most variation. Such standardization does not completely solve the problems associated with scale differences since the clusters are unknown and so it is not possible to perform a within cluster standardization. In contrast, the LC clustering solution is invariant of linear transformations on the variables, so standardization of variables is not necessary, (Vermunt & Magidson, 2002).
(3) Determination of the number of clusters. Hierarchical clustering methods provide no assistance in determining the number of clusters. In contrast, LC clustering provides various diagnostics such as theoretical information criteria or likelihood ratio test, which can be useful in determining the number of clusters, (Dillon & Kumar, 1994; McLachlan & Peel, 2000).

(4) Inclusion of variables of mixed scale types. Hierarchical clustering methods are limited to interval scale quantitative variables. In contrast, extended LCM can be estimated in situations where the variables are of different scale types. Variables may be continuous, categorical (nominal or ordinal) or counts or any combination of these (Vermunt & Magidson, 2002). If all variables are categorical, one obtains a traditional LC model (Goodman, 1974).

(5) Inclusion of demographics and other exogenous variables. A common practice following a hierarchical clustering is to use discriminant analysis to describe differences between the clusters on one or more exogenous variables. In contrast, the LC cluster model can be easily extended to include exogenous variables (covariates). This allows both classification and cluster description to be performed simultaneously using a single uniform maximum likelihood estimation algorithm (Fonseca & Cardoso, 2007; Vermunt & Magidson, 2002).

Notes on contributor

Jaime R.S. Fonseca, PhD, is an assistant professor. Research interests include: multivariate data analysis, mixed methods in social sciences, mixed methods in health sciences, mixed methods in criminology, latent class models, information criteria, market segmentation and customers satisfaction. Books/articles published: *Estatística Matemática* (2000); *Estatística Matemática* (2001); *Retail clients latent segments, in progress in artificial intelligence* (2005); Supermarket customers segments stability, *Journal of Targeting, Measurement and Analysis* (2007); Mixture-model cluster analysis using information theoretical criteria, *Intelligent Data Analysis* (2007); The application of mixture modelling and information criteria for discovering patterns of coronary heart disease, *Journal of Applied Quantitative Methods* (2008); Customer satisfaction study via a latent segment model, *Journal of Retailing and Consumer Services* (2009); *Análise de Dados Univariados e Multivariados* (2010); *Students' attitudes towards mathematics, in Encyclopedia of the Sciences of Learning (ESL)* (2011); Why does segmentation matter? Identifying market segments through a mixed methodology, *European Retail Research* (2011); How satisfied are citizens with public hospitals' service?, *International Journal of Healthcare and Quality Assurance, Emerald* (2012); *Atitudes e comportamentos do eleitorado de Portugal e suas semelhanças com países latino-americanos in Comunicação Política e Comportamento Eleitoral em América Latina* (2012); *How latent class models matter to social network analysis and mining: Exploring the emergence of community, in the influence of technology on social network analysis and mining* (2012).

References

Abbot, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research, 29*, 3–33.

Ahlquist, J.S., & Breunig, C. (2009). Country clustering in comparative political economy. Max Planck Institute for the Study of Societies, MPIfG discussion paper, No. 09/5, Cologne, Germany. Retrieved from http://hdl.handle.net/10419/36527

Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage.

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York, NY: Academic Press.

Arts, W., & Gelissen, J. (2002). Three worlds of welfare capitalism or more? A state-of-the-art report. *Journal of European Social Policy, 12*(2), 137–158.

Bailey, K.D. (1983). Sociological classification and cluster analysis. *Quality and Quantity, 17*, 251–268.

Bailey, K.D. (1975). Cluster analysis. In D.R. Heise (Ed.), *Sociological methodology 1975* (pp. 59–128). San Francisco, CA: Jossey-Bass.

Bambra, C. (2007). Going beyond the three worlds of welfare capitalism: Regime theory and public health research. *Journal of Epidemiol Community Health, 61*(12), 1098–1102.

Bellah, R.N. (1964). Religious evolution. *American Sociological Review, 29*, 358–374.

Boivin, M., & Bégin, G. (1989). Peer status and self-perception among early elementary school children: The case of the rejected children. *Child Development, 60*, 591–596.
Clements, F.W. (1954). The relationship of thyrotoxicosis and carcinoma of the thyroid to endemic goitre. *The Medical Journal of Australia, 4*, 894–897.
Clogg, C.C. (1995). Latent class models. In G. Arminger, C.C. Clogg, & M.E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York, NY: Plenum.
Dillon, W.R., & Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview, chapter 9. In R.P. Bagozi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge: Blackwell.
Elman, C. (2005). Typologies in qualitative studies of international politics. *International Organization, 59*, 293–326.
Evans, G., & Mills, C. (1998). A latent class analysis of the criterion-related and construct validity of the goldthorpe class schema. *European Sociological Review, 14*, 87–106.
Everitt, B. (1974). *Cluster analysis*. London: Heinemann Educational.
Fisher, S. (1993). Gambling and pathological gambling in adolescents. *Journal of Gambling Studies, 9*, 277–288.
Fonseca, J.R.S. (2010, March 29–31). *On the performance of information criteria in latent segment models estimation with categorical segmentation base variables, Proceedings of ICMSE 2010. International Conference on Mathematical Science and Engineering*, World Academy of Science, Engineering and Technology, WASET, Rio de Janeiro, Brazil.
Fonseca, J.R.S. (2011). In Dirk Morschett, Thomas Foscht, Thomas Rudolph, Peter Schnedlitz, Hanna Schramm-Klei Bernhard Swoboda (Eds.), *Why does segmentation matter? Identifying market segments through a mixed methodology, European retail research* (Vol. 25, pp. 1–26). Gabler Verlag: Springer Fachmedien.
Fonseca, J.R.S., & Cardoso, M.G.M.S. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis, 11*, 155–173.
Friedkin, N.E. (1978). University social structure and social networks among scientists. *AJS, 83*, 1444–1465.
George, A.L., & Nennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.
Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.
Hagenaars, J.A., & Halman, L.C. (1989). Searching for ideal types: The potentialities of latent class analysis. *European Sociological Review, 5*, 81–96.
Hogan, D.P., Eggebeen, D.J. & Clogg, C.C. (1993). The structure of intergenerational exchanges in american families. *American Journal of Sociology, 98*, 1428–1458.
Kaufman, L., & Rouusseeuw, P.J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: Wiley.
Kroeber, A.L. (1939). *Cultural and natural areas of native, North America*. Berkeley, CA: University of California Press.
Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.
Lee, C.-K., Lee, Y.-K., Bernhard, B.J., & Yoon, Y.-S. (2006). Segmenting casino gamblers by motivation: A cluster analysis of Korean gamblers. *Tourism Management, 27*, 856–866.
Magidson, J., & Vermunt, J.K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*, 36–43.
Mandara, J. (2003). The typological approach in child and family psychology: A review of theory, methods, and research. *Clinical Child and Family Psychology Review, 6*, 129–146.
Marx, K. (1951). *The eighteenth Brumaire of Louis Bonaparte*. (D. DeLeon, Trans.). New York, NY: New York Labor News.
McCutcheon, A.L. (1987). *Latent class analysis* Sage University Paper. Newbury Park, CA: Sage.
McLachlan, G.F., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
McQuitty, L.L. (1987). *Pattern-analytic clustering: Theory, method, research and configural findings*. Lanham, MD: University Press of America.

Milligan, G.W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psycometrika, 46*, 187–199.
Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20*, 359–363.
Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24, 882–891.
Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
Parsons, T. (1966). *Societies: Evolutionary and comparative perspectives*. New York, NY: Prentice-Hall.
Punj, G., & Stewart, D.W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, XX*, 134–148.
Silverstein, M., & Bengtson, V.L. (1997). Intergenerational solidarity and the structure of adult child–parent relationships in American families. *American Journal of Sociology, 103*, 429–460.
Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy*. San Francisco, CA: W.H. Freeman.
Spencer, H. (1864). *First principles*. New York, NY: Appleton.
Sucoff, C.A., & Upchurch, D.M. (1998). Neighborhood context and the risk of childbearing among metropolitan-area black adolescents. *American Sociological Review, 63*, 571–585.
Taylor, S.H. (1983). Adjustment to threatening events. *A Theory of Cognitive Adaptation, American Psychologist, 38*(11), 1161–1173.
Vanneman, R. (1977). The occupational composition of American classes: Results from cluster analysis. *American Journal of Sociology, 82*, 783–807.
Vermunt, J.K., & Magidson, J. (2002). Latent class cluster analysis. In J.A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class, analysis* (pp. 89–106). Cambridge: Cambridge University Press.
Widmer, E.D., Treas, J., & Newcomb, R. (1998). Attitudes toward nomarital sex in 24 countries. *Journal of Sex Research, 35*, 349–358.
Wolf, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329–350.
Zhang, N.L. (2003). *Structural EM for hierarchical latent class models* (Technical Report HCUSTCS03-06). Hong Kong: Hong Kong University of Science & Technology.