Psychology Science, Volume 46, 2004 (1), p. 65-98

Typifying developmental trajectories – a decision making perspective

ALEXANDER VON EYE[1], EUN YOUNG MUN[2], ALKA INDURKHYA[3]

Abstract

Developmental trajectories are defined as curves of repeated observations. Individuals may differ in the starting point, the degree of acceleration or deceleration, the timing of acceleration or deceleration, overall shape, elevation, and scatter of curves. This article discusses methods for typifying developmental trajectories. Two groups of methods are considered. The first group involves assigning individuals to a priori existing trajectories and counting the number of individuals that reflect natural groupings of trajectories based on categorical classifications, using Configural Frequency Analysis (CFA). The second method involves employing methods of cluster analysis. When selecting a method of cluster analysis, the following ten cluster characteristics need to be considered: (1) disjoint vs. overlapping clusters; (2) hierarchical vs. non-hierarchical clustering; (3) agglomerative vs. divisive clustering; (4) exhaustive vs. selective classification; (5) stochastic vs. deterministic clustering; (6) clustering based on correlation vs. distance measure; (7) convex vs. non convex clusters; (8) clustering based on symmetric vs. asymmetric measure; (9) monothetic vs. polythetic classification, and (10) manifest versus latent variable clustering. A review of clustering methods is presented using examples to demonstrate the pros and cons of each method. Discriminant analysis and logistic regression are discussed as methods for subsequent analysis of groupings. Examples are presented using artificial data and empirical data on the development of cigarette smoking in male adolescents.

Key words: cluster analysis, decision making, configured frequency analysis

[1] Alexander von Eye, Michigan State University, Department of Psychology, S-119 Snyder Hall, East Lansing, MI 48824-1117; E-mail: voneye@msu.edu

[2] Eun Young Mun, The University of Alabama at Birmingham, Department of Psychology & The Center for the Advancement of Youth Health, 912 18 Street South, Birmingham, AL 35294-1200

[3] Alka Indurkhya, Harvard School of Public Health, Department of Society, Human Development, and Health, SPH3 624A, Boston, MA 02115

The authors are indebted to Michael Windle for making data available for this article and for commenting on an earlier draft. Eun Young Mun's work was supported in part by NIAAA grant # R 37-AA07861.

Typifying Developmental Trajectories - A Decision Making Perspective

Longitudinal developmental research faces the problem that the number of possible developmental trajectories can be very large. Consider a time series of seven occasions. At each of these occasions, a variable is observed that can assume five scores. The number of possible trajectories on this single variable is $5^7 = 78{,}125$. Suppose that researchers observe three variables. This modest multivariate design increases the number of possible trajectories to $5^{7^3} = 4.768371582031e{+}14$. Or, suppose there are six heterogeneous groups based on quadratic trend-trajectories of one variable over seven measurements. If these trend parameters are dichotomized, a total of $2^{36} = 262144$ possible combinations exist. These numbers of possible trajectories are large enough to make one wonder what can be done to depict the structure of development in a simple yet valid way. Two options will be considered in this article. The first involves a priori specifying typical or important trajectories. This specification may be based on prior knowledge or theory. The second option involves employing grouping procedures. We consider methods of cluster analysis.

1. The Plethora of Developmental Trajectories: A Person-Oriented Perspective

To begin the discussion of depicting the structure of large amounts of information, consider the following example. The variable Physical Aggression Against Peers (PAAP) of a sample of N = 106 students was observed at three occasions in 1983, 1985, and 1987 (Finkelstein, von Eye, & Preece, 1994). The observed raw score profiles appear in Figure 1.

Figure 1 shows that each of the 56 students with complete data displays his or her own, unique trajectory of physical aggression over time. No two trajectories are the same. For the data in Figure 1, we can calculate the descriptors given in Table 1. From the means, we conclude that, physical aggression appears to recede over time.

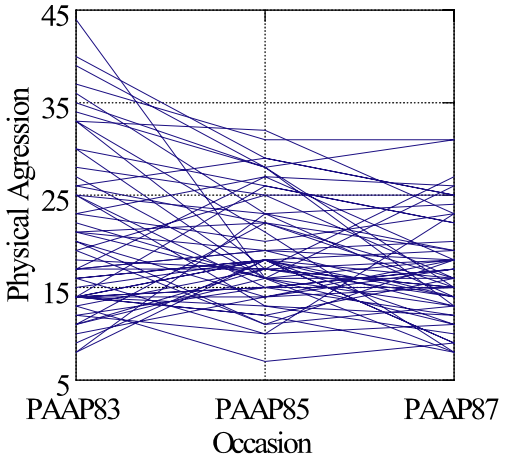

Figure 1:
Longitudinal profila of aggression

Table 1:
Descriptive statistics for data on development of Aggression

	PAAP83	PAAP85	PAAP87
N of cases	106	77	70
Minimum	8.000	7.000	8.000
Maximum	44.000	57.000	31.000
Mean	21.283	19.221	17.029
Standard Dev	8.415	7.212	5.685

The approach pursued when calculating and interpreting sample statistics is known as the *nomothetic approach*. It is the goal of the nomothetic approach to create statements that apply to the entire population. Accordingly, significance testing attempts to generalize sample results to this population. When parameters are significant, researchers conclude that effects exist in this population.

The comparison of the means and standard deviations in Table 1 with the parallel coordinate plot in Figure 1, however, suggests that the data carry far more information than can be meaningfully depicted by descriptors for the entire sample. Lack of strong and homogeneous trends often results in parameters that fail to reach significance. Based on non-significant parameters, researchers tend to conclude either that the effects do not exist or that there was not enough power, given the sample size, the effect size, and the reliability of the measures.

The *Person-Oriented Approach* (Bergman & Magnusson, 1997; Cairns, Bergman, & Kagan, 1998; Magnusson, 1998; von Eye & Bergman, 2003) offers a third interpretation for the lack of significant results at the aggregate level. This approach questions the assumption that the sample was drawn from just one population (for statistical examples see von Eye & deShon, 1998). The Person-Oriented Approach proposes that "the relevant aspect is the *profile* of scores" (Bergman & Magnusson, 1997, p. 293). In addition, each individual is viewed as a whole entity that functions via the interactions of the elements involved. As a result, individuals can be different from each other, and treating them as members of the same parent population without focusing on their differences may lead to a loss of important information.

If researchers consider the concept of interpretable inter-individual differences of profiles, they face the question whether it is possible to identify groups of individuals that can be described using measures of central tendency without loss of important information. In this article, we consider two such methods. The first involves asking how many individuals belong to a priori specified groups. The second involves determining such groups using methods of cluster analysis (cf. Gutiérrez-Peña's distinction between supervised and unsupervised classification in this issue).

2. Specifying A Priori Groups

In this section, we describe two approaches of specifying a priori groups. The first approach uses a priori existing knowledge. The second approach involves cross-tabulating categorical variables.

2.1 Using Existing Knowledge for Group Specification

In many instances, researchers possess a priori knowledge that allows them to a priori expect certain groupings. For instance, in traffic psychology, some researchers entertain the concept of the individual that is accident-prone; in the cartoons, there is the *Born Loser*; in law psychology, there is the concept of the *typical victim*; in developmental psychology, researchers describe the prototypical developmental trajectories of *retarded children*; or in nosology, there is the concept of the *alcoholic*. In these and other instances, it is often possible to provide precise numerical descriptions of hypothesized trajectories.

Consider a learning process that can be described by a first phase of rapid learning progress. This phase is followed by a slowing down, and the process approaches a ceiling, an asymptote. This type of process can be described by an equation of the type $p_n = 1 - (1 - p_1)(1 - \theta)^{n-1}$, where p_n is the probability of a correct response in trial n, p_1 is some prior response probability, n is the learning trial ($n = 1, ...$), and θ is the acceleration parameter. The learning curve rises faster for larger scores of θ (see Hilgard & Bower, 1975). Figure 2 displays two sample learning curves (black symbols). Both curves have parameter $p_1 = 0.2$. The top curve (solid black asterisks) has $\theta = 0.15$, the bottom curve (solid black diamonds) has parameter $\theta = 0.05$.

In a classification process, researchers may ask how many individuals display a learning curve of the types given in Figure 2. Specifically, one may ask how many individuals display a curve with parameter $\theta = \{0.05, 0.1, 0.15 ...\}$. As a matter of course, the probability that an individual will display one of these curves exactly is very slim. Therefore, researchers may wish to define a band around each curve within which an individual may lie. This band can be considered parallel to a confidence interval. Alternatively, if the parameters of a trajectory were estimated using standard statistical methods, confidence intervals can be determined and used as bands. A band of width 2ε yields to an equation of the type $p_n = 1 - (1 - p_1)(1 - \theta)^{n-1} \pm \varepsilon$. Figure 2 exemplifies the idea of trajectories with band widths (gray lines and symbols). For each of the trajectories, a band was drawn with $\varepsilon = \pm 0.1$.

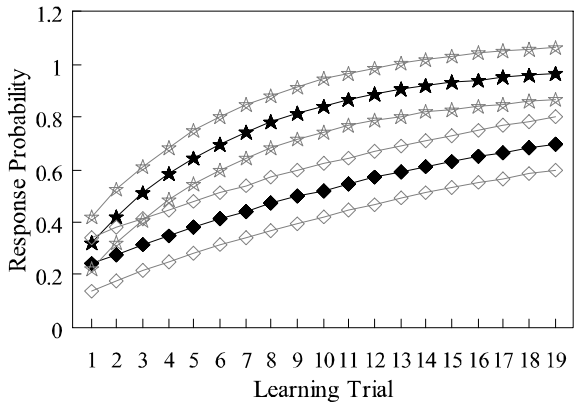

Figure 2:
Learning Curve

When assigning individuals to the existing two trajectories, the following rule can be used for constant p_i:

$$\textit{if} \quad \begin{cases} \theta = 0.15 \;,\; |\delta| \leq 0.1 \textit{ for } 1 < n \leq N & \textit{then } T = 1 \\ \theta = 0.05 \;,\; |\delta| \leq 0.1 \textit{ for } 1 < n \leq N & \textit{then } T = 2 \\ \textit{else} & T = 3 \;, \end{cases} \tag{1}$$

where N is the total number of trials, $\delta = p_{i,n} - p_n$, that is the difference of individual i's response probability at trial n from the model probability, p_n, and T indexes the trajectories, with T = 1 if $\theta = 0.15$; T = 2 if $\theta = 0.05$; and T = 3 else. Alternative assignment rules are conceivable (for Bayesian classification rules, see Gutiérrez-Peña, this issue). These include, for instance, rules where only a percentage of differences δ must be within the band given by ε, and rules where individuals can switch between types of trajectories.

Equation 1 can be interpreted as an assignment rule. This rule has the following characteristics:

1. It is an a priori rule, that is, this rule was not specified based on information provided by data but rather on the theoretical formulation of learning curves that differ only in the acceleration parameter, θ.
2. It is a deterministic, non-statistical rule. Thus, individuals are assigned to Trajectory 1 if their learning curves fall within the band for the upper curve. Individuals are assigned to Trajectory 2 if their learning curves fall within the band for the lower curve. If neither curve applies, individuals are assigned to *Trajectory* 3, that is, a trajectory with unspecified characteristics.
3. The rule allows the bands to overlap. Figure 2 shows that Trajectories 1 and 2 can not be perfectly distinguished at Trials 1 through 3. Beginning with Trial 4, the bands are completely separated.
4. Although Figure 2 suggests that, after Trial 3, the two learning curves are perfectly separated, researchers may need information from all Trials, even including Trial 1, because an individual can deviate from one or all a priori specified learning curves at any trial, and the time point of deviation may be of educational or diagnostic importance.

The importance of a priori specification of trajectories as the ones lined out in this section lies first in its connection to substantive theory. The classification system is not determined by data characteristics. In contrast, the classification of individuals will reflect data characteristics. When specifying trajectories or, in more general terms, groupings, a priori, one can ask what percentage of a population can be described by these groupings that were derived from theory and, thus, conform with this theory. In the above example, every individual that is assigned to Trajectories 1 and 2 conforms with the theory that allows derivation of these two trajectories. Individuals assigned to Trajectory 3 may not necessarily contradict the theory but they do not display any of the predicted pathways either.

2.2 Cross-classifying Categorical Variables Yields Natural Groupings

Application of the sample methods introduced in the last section presupposes that theories or prior knowledge exist that allow one to specify trajectories that can be taken by respondents. Sample applications of such knowledge include the description of behavior as following a pattern of a cyclical psychosis, the diagnosis of a learning progress as indicative of retardation, or the description of the developing gambling habits of an individual as a growing addiction.

There is a number of circumstances, however, that prevent researchers from specifying trajectories and their parameters a priori. These circumstances are typically characterized as exploratory research, that is, research where theories still need to be built and prior knowledge is scarce. It is important to emphasize that, although exploratory, this research is not completely theory-free or without prior knowledge. The area of research will always be specified, and the variables under study will always be assumed to be relevant to the phenomenon of interest. If variable selection is completely at random, one may wonder whether a research activity still qualifies as scientific.

In this section, we illustrate how natural groupings can result when categorical, nominal-level or ordinal variables are observed repeatedly. Consider a study that investigates psychiatric diagnoses in a sample of schizophrenic inpatients. The study is carried out to find out whether diagnoses change over time. Suppose the diagnoses are (1) cured; (2) paranoia; and (3) schizophrenia. Suppose also that there are two observation points (the classification strategy will not change if there are more observation points). Table 2 displays the cross-classification of the two diagnoses. Please notice that Category 1, that is, the diagnosis *cured*, can appear only at the second observation. Therefore, the cross-classification has 2 x 3 cells rather than 3 x 3.

Table 2:
Cross-Classification of Schizophrenia Diagnoses at two Occasions
(entries in cells are cell indices)

Time 2 Time 1	Schizophrenic	Paranoid	Cured
Schizophrenic	11	12	13
Paranoid	21	22	23

The entries in Table 2 denote the cell indices. They are not frequencies. For example, 11 denotes the cell that contains those patients who were diagnosed as schizophrenic at both occasions. Cell 22 contains the cases that were paranoid at both occasions. These two cells (shaded) contain cases with stable diagnoses, also called the *persisters* or *stayers*. All other cells contain patients who changed their symptoms or symptom severity. These are the *changers* or *movers* (Clogg, Eliason, & Grego, 1990; von Eye & Schuster, 2002). Many of the moves are evaluated positively. For instance, Cells 13 and 23 contain cases that are con-

sidered cured at the second observation. In contrast, if paranoia is a precursor of schizophrenia, Cell 21 contains those cases that worsened by displaying more and more severe symptoms of schizophrenia.

As can be seen from the example in Table 2, cross-classifying categorical variables creates groupings in a natural way. When repeatedly observed variables are crossed, each of the groups is characterized by a specific pattern of constancy or change. In the example in Table 1, the crossed variables are scaled at the nominal level. When ordinal or higher-level variables are crossed, questions can be asked that are fueled by the specific scale characteristics, and thus, go beyond the questions that can be asked for nominal level variables. For instance, when a symptom is observed not as present versus absent but in regard to severity, cross-classifications of repeated observations allow one to ask questions concerning the development of symptoms to the worse or to the better, whether severity is related to change, or whether treatment success is related to severity (for examples and methods of analysis see Agresti, 1996; Clogg, Eliason, & Grego, 1990; von Eye, 2002; von Eye & Spiel, 1996; von Eye & Schuster, 2002).

The methods for creating groups discussed in Sections 2.1 and 2.2 share in common that they are not based on information provided by the data at hand. Rather, these methods use information from theory or prior results (Section 2.1) or the information provided by the categorical level nature of the variables (see Section 2.2). The information in the data is then projected onto the a priori specified structure. The structure itself exists before data are collected, even without the data. In the next section, we discuss methods for forming groups that use chiefly the information provided by the data. Specifically, we discuss methods of cluster analysis and decisions that must be made when selecting a clustering method.

3. Decisions in the Selection of Clustering Methods

Methods of clustering information are popular for a number of reasons. First, these methods allow researchers to create *structure out of chaos*. In many instances, as was illustrated in Figure 1, it is not deemed sufficient to estimate measures of central tendency or other descriptive measures when it is clear that these measures describe only a small portion of some population, and the rest shows large discrepancies from the parameters. If groups exist, methods of cluster analysis may help detect them. Second, most methods of cluster analysis are non-statistical in the sense that they do not require researchers to make assumptions concerning the parameters of underlying distributions. In addition, significance tests are rarely performed, for lack of null hypotheses that could be tested. Thus, methods of cluster analysis are largely assumption-free and ubiquitously applicable. Third, methods of cluster analysis practically always yield solutions (for examples, see von Eye & Bergman, 2003). Exploratory factor analysis and principal component analysis, methods of multidimensional scaling and most descriptive methods of statistics share this characteristic. As a result, researchers will always be able to come up with a grouping solution. Only in degenerate cases where, for instance, all distances between neighbored cases are the same, there will be no *reasonable* solution, that is, no solution that is easily interpreted.

In particular this third characteristic of cluster analytic methods has met with criticism, for it harbors an element of arbitrariness. If one cannot fail, some say, one wonders whether the analysis is worthy of the label of *scientific*. The present section pursues two goals. The

first is to show that this element of arbitrariness can be minimized when methods of cluster analysis are not blindly applied but a series of decisions has been made and justified. These decisions are critical because they determine the characteristics of the clustering solution. Second, this series of decisions is introduced and illustrated. Section 3.1 presents a definition of the clustering problem. Beginning in Section 3.3.1, we discuss the decisions.

3.1 The Clustering Problem

In this section, we give a general definition of the clustering or classification problem (Blashfield & Aldenderfer, 1988; Bock, 1974, p. 22; Hartigan, 1975). Consider N objects (i.e., cases), O_1, ..., O_N. For these objects, a data matrix, x_{ki}, a similarity matrix[4], d_{jk}, or a relation, $-$, is determined. x_{ki}, d_{jk}, or $-$ depict the similarity structure of the set of objects, $S = \{O_1,...O_N\}$. Searched for is a *classification*, $A = (A_1, A_2,...)$ of S. The subsets of A, that is, A_1, A_2, ... are classes (also called groups or clusters) that (1) reflect the similarity structure of the objects as well as possible, and (2) allow for data reduction. These two requirements are often fulfilled if the objects within a group, A_i, are (1) maximally similar to each other and (2) different classes are easily distinguished from each other, that is, the cross-class dissimilarity is maximized. The first characteristic is called *homogeneity*; the second characteristic is called *separation* of groups (classes, clusters). Mathematically more precise formulations are possible. They typically take characteristics of data and the structure of the desired classification into account. For the sake of simplicity, we only present this definition. The following sections introduce the decisions that need to be made when selecting a method of clustering.

3.2 Decisions About Clustering

Just as when making decisions concerning the most appropriate statistical method for analysis of a particular data set, decisions must be made when selecting a clustering method. These decisions concern the characteristics of the method employed and, as a consequence, the characteristics of the resulting solution. When making these decisions, characteristics of variables and data, and theoretical considerations concerning the desired classification are taken into account.

Before discussing the ten decisions, it is important to emphasize that there is no such thing as *the correct, unique clustering solution*. Rather, employing a clustering method implies that a mathematical structure is superimposed on a data set. The result of a cluster analysis will reflect characteristics of both the data and the mathematical structure. Cluster analysis is almost always able to depict some but not all of the data characteristics. Therefore, a clustering solution can be correct in the sense that it reflects certain characteristics. However, it will always be *incomplete* because there are other characteristics that it cannot reflect.

---

[4] In later sections of this article we will also discuss distance matrices. For the present purposes we assume that the term *similarity matrix* also subsumes distance matrices and other matrices that can be created to describe the relationships between objects. This applies accordingly to the term *similarity structure*, below.

This perspective of the correctness of a clustering solution has two very important implications. The first implication is that researchers may wish to apply more than one clustering method to their data. These methods may provide different solutions. Each of these solutions may be plausible. And, most importantly, each of these possible solutions will be correct in the above sense. All this is said assuming, of course, that no computational errors are committed. The second implication is that many of the method-comparative investigations are problematic. This applies in particular to investigations that involve so-called *plasmodes*, that is, real or artificial data sets of known characteristics. These investigations proceed under the assumption that the plasmodes display a data characteristic that can be reflected by each of the employed methods. This will rarely be the case. Therefore, it does not come as a surprise that the conclusions from the method-comparative investigations failed to be clear-cut. The investigations are statements about *method* characteristics as much as about *data* characteristics. Results can vary with either.

3.3 Ten Decisions

This section discusses ten decisions that must be weighed when selecting a clustering method. One important decision is notably absent from this list, that is, the decision concerning the selection of a suitable software package. Data analysts have their specific preferences, and software packages differ in the kind and number of decisions that they enable researchers to make. Most general purpose statistical software packages (e.g., SPSS, SYSTAT, SAS, S plus) contain a selection of methods. In addition, there exist specialized classification programs such as Wishart's CLUSTAN (Wishart, 1987) and Bergman and ElKhouri's SLEIPNER (1998). For the remainder of this article, we assume that data analysts have access to some appropriate software package, and that this package allows them to perform the desired analyses. The order of the following ten decisions has no effect on the selection process.

3.3.1 Decision 1: Disjoint versus Overlapping Clusters

Disjoint classifications have the following characteristic: Each object O_j belongs to only one subset A_i d α. In contrast, overlapping classifications allow each object O_j to belong to more than one subset. Consider the following example. Two random variables are created for a sample of N = 200 with parameters as given in Table 3.

Researchers may ask whether this sample is *homogeneous* or, whether there exist well separated subgroups. Figure 3 displays the scatterplot of the two variables, VAR(1) and VAR(2).

The circles in Figure 3 indicate the 95% confidence ellipses for two imaginary subgroups, that is, 95% of the cases are expected to lie within the circles. The circles overlap. We thus conclude that there may exist a small group of approximately six objects that are not easily assigned to either group. How can one deal with situations of this kind?

There are several solutions, two of which will be discussed here. A first solution implies creating non-overlapping clusters. In the present example, this can be achieved by, for instance, reducing the diameters of the circles such that they will no longer overlap. The number of objects located outside the circles will increase as the diameters shrink. Thus, the

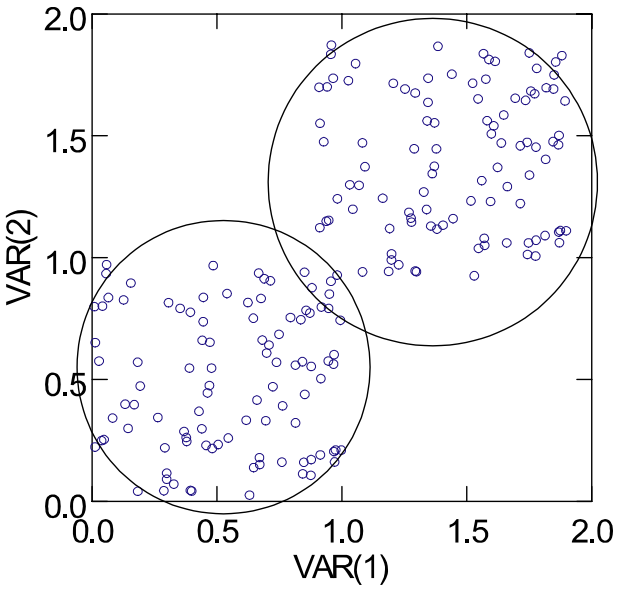

Figure 3:
Cluster structure for two random variates

Table 3: Parameters of two random variates

	VAR(1)	VAR(2)
N of cases	200	200
Minimum	0.011	0.023
Maximum	1.900	1.869
Mean	1.007	0.948
Standard Dev	0.544	0.534

circles will no longer represent the 95% confidence ellipses. However, there will be no overlap. The cases outside can then be assigned to a separate group, be considered members of no cluster, be assigned to one of the two big clusters at random, or be allowed to form separate small groupings that are also disjunct.

The second solution implies considering whether overlapping cluster solutions may be appropriate. There may be cases for which it is reasonable to assume that they belong to more than one group. For instance, an athlete's body build may allow her to perform both as a high jumper and a volley ball player, or a student may be both an average performer and a slow developer.

Thus, researchers face the situation in which they must make a decision as to whether clusters may overlap or not. *Separation* of clusters is a characteristic that describes the degree to which clusters are disjunct. However, the concept of separation is applicable only if

classifications are allowed to create groupings that overlap. If groupings are not naturally disjunct - as is the case when categorical variables are analyzed (see Section 2.2) -, non-overlapping clusters can carry some artificial flavor. In continuous variables contexts, overlapping groupings are very likely to happen. Therefore, application of methods that always create non-overlapping clusters requires justification. Researchers may need to make their decisions as to what type of method to apply based on substantive and theoretical considerations.

When reading applications of clustering methods and when inspecting general purpose statistical software packages, one realizes quickly that researchers prefer methods that yield non-overlapping clusters. One reason for this preference may be that it seems desirable to create groupings where objects can belong to only one cluster. Overlapping groupings are perceived by many as only half-complete solutions. In addition, when in a second analytic step other variables are used to assess the external validity of a grouping, objects with dual citizenship may pose problems for instance when cluster membership is considered a factor in a MANOVA.

Most general purpose statistical software packages do not include modules that allow one to create overlapping clusters. Two programs that can create such cluster solutions include ADCLUS by Arabie and Caroll (1980) and Pyramid by Aude, Diaz Lazcoz, Codani, and Risler (1999; cf. Everitt, Landau, & Leese, 2001).

3.3.2 Decision 2: Hierarchical versus Non-hierarchical Clusters

Classifications can be either hierarchical or non-hierarchical. Hierarchical clustering procedures create series of cluster solutions using a process that either groups objects together into larger and larger groups. These procedures are called *agglomerative*. The main characteristic of these procedures is that smaller numbers of clusters result from merging clusters without loss of cluster members. When divisive procedures are employed, the clustering process is hierarchical if greater numbers of clusters result from splitting clusters. In other words, a clustering method is hierarchical if all objects that belong to the same cluster when the number of clusters is greater are also members of the same cluster when the number of clusters is smaller. One of the most famous applications of hierarchical cluster methods enabled biologists to reproduce the descent of species originally developed by von Linné (1751, 2003). The results of hierarchical clustering processes are typically displayed in the form of dendrograms (see Figures 4 and 5, below).

Clustering methods are non-hierarchical when such a series is not created. Alternatives to hierarchical clustering procedures include methods that identify space density maxima and group objects in the same cluster if their respective distances to the same density center (center of gravity) is smallest (examples of such methods include the well known k-means method and methods discussed by von Eye & Wirsing, 1978, 1980). Non-hierarchical methods create only one solution. The number of clusters in such a solution is either predetermined as in k-means, or results from the number of space density maxima identified according to some a priori specified criteria. In contrast, hierarchical methods create N - 1 solutions. Therefore, a decision as to when to stop merging (in an agglomeration process) or dividing (in a divisive process) is a key part of hierarchical cluster analysis (see Section 3.3.3 for more information).

When deciding whether to employ hierarchical or non-hierarchical methods, one can use two criteria. The first criterion is whether the hierarchy or the series of clusters can be mean-

ingfully interpreted. If this is the case, there is no substitute for hierarchical methods. When, in contrast, there is no way nor intention to interpret the series of cluster solutions, information about the agglomeration or division process can help identify a good solution. It is well known, however, that many measures of the damage that is done when fusing clusters are not always conclusive. When researchers already have hypotheses or assumptions about the number of possible clusters, non-hierarchical solutions may be more parsimonious. Only one or a small number of solutions needs to be calculated, and interpretation is typically straightforward. The second criterion is thus the knowledge that researchers have prior to performing a cluster analysis.

3.3.3 Decision 3: Agglomerative versus Divisive Clustering

In the previous sections, we discussed the distinction between disjunct and overlapping clusters, that is, a distinction related to the product of classification, and the distinction between hierarchical and non-hierarchical solutions, that is, a distinction related to the process of creating clusters. In this section, we discuss a distinction related to the process of clustering in hierarchical clustering. Specifically, we discuss the distinction between *agglomerative* and *divisive* clustering. The former starts the clustering under the assumption that each object, O_j, forms a separate cluster, that is, $S = \{O_1,...,O_N\} = \mathbf{A} = (A1, A2,...)$. When grouping objects with the goal to create a more parsimonious solution, agglomerative methods join those objects that are the most similar to each other first. This is repeated until either some criterion of optimality, for instance, some R^2 threshold, has been reached or until the last two groupings are joined.

Divisive clustering proceeds in the opposite direction. It considers all objects members of one cluster. When creating groups, divisive clustering splits the existing cluster(s) such that the benefit, measured, for instance, in units of decrease in information or increase in R^2, is greatest. This step is repeated until either some criterion of optimality, for instance, some R^2 threshold, has been reached or until each object forms its own cluster. In other words, the situation that $S = \{O_1,...,O_N\} = \mathbf{A} = (A1, A2,...)$ which is the starting point for the agglomerative procedure is the end point of analysis for the divisive procedure.

To illustrate the presentation of the processes of agglomeration and division in the form of *dendrograms* (tree structures), consider the following example. A sample of N = 6 objects is classified using an agglomerative clustering method. This process is depicted in Figure 4.

To illustrate the agglomerative procedure, we read Figure 4 from top to bottom. At the top there are six vertical lines. Each line represents one object. In a first step the two left most objects are joined to form one cluster. Next, the two right most objects are also joined to form a separate cluster. In the third step, Object 3 joins the cluster of the first two. In Step 4, Object 4 joins the first three. At each of these steps a new cluster is created. In the last step, the cluster of Objects 5 and 6 is fused with the cluster of the first 4 objects. Of the N - 1 solutions thus created, researchers interpret the one that best meets the optimality criteria and criteria derived from theory (and plausibility).

To illustrate the divisive procedure, Figure 4 can be used again. This time, however, we read it from bottom to top. The procedure begins by considering all objects members of one cluster. In the following steps this big cluster is subdivided until each object resides in a separate cluster. As for the agglomerative procedure, optimality criteria and criteria of theory and plausibility are used to select from the N - 1 solutions thus created.

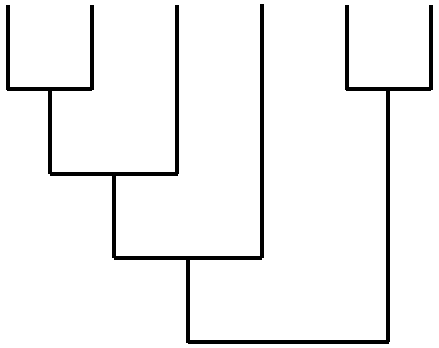

Figure 4:
Dendrogramm

To give an example with a larger data set, consider the data from Figure 3 again. The dendrogram for these data appears in Figure 5. As was suggested in the interpretation of Figure 3, this data set contains two groups of objects. Figure 3 also suggests that these two groups overlap in the neighborhood of the centroid of the entire sample. The clustering method used to create Figure 5 creates disjunct clusters. The dendrogram displayed in Figure 5 gives no hint at the overlap.

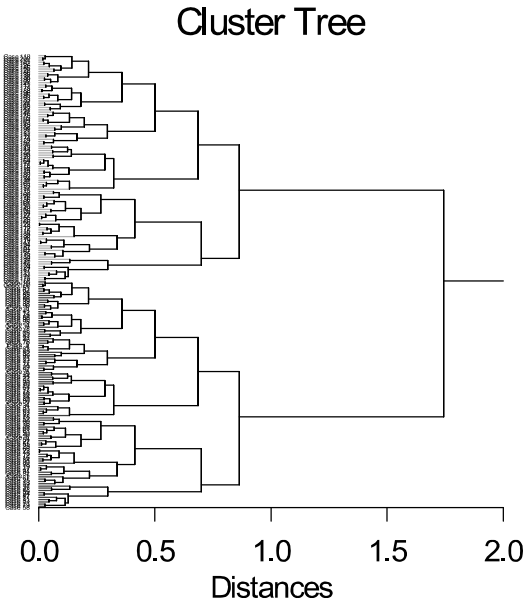

Figure 5:
Dendrogram for Data in Figure 3

When weighing which of the two procedures to select, the agglomerative or the divisive one (if any; see Section 3.3.2 for alternatives), it can be shown that results from both methods are largely the same. Therefore, selection of one of the alternatives may be a matter of preferences and program availability. Most software packages focus on agglomerative methods and divisive methods are rarely included. Single linkage (or nearest neighbor), complete linkage (or furthest neighbor), average linkage, centroid clustering, median clustering, weighted-average linkage, beta-flexible, density linkage, and Ward's method are examples of agglomerative clustering methods available in many software packages.

3.3.4 Decision 4: Exhaustive versus Non-exhaustive Clustering

Classifications are *exhaustive* if $\boldsymbol{\alpha}$ contains all objects, O_j. Non-exhaustive methods generate classifications that involve only some of the N objects. From a classification perspective, the objects assigned to clusters are often considered the *most important* objects. The objects that are not assigned to a cluster are often either ignored or assigned to a *poubelle*, that is, a garbage can.

The idea of non-exhaustive classification can be illustrated in two ways. First, consider Figure 3 again. This figure displays the scatterplot of the two variables, VAR(1) and VAR(2), and suggests that there may be two groups of objects. However, there may be an overlap between the two groups. Shrinking the circles that circumscribe the two groups so that there is no overlap will have the side effect that a number of objects will not belong to any of the two groups. Thus, after the shrinking, the clustering is non-exhaustive.

The second way to describe non-exhaustive classifications involves Configural Frequency Analysis (CFA; Indurkhya & von Eye, 1999; Lienert & Krauth, 1975; von Eye, 2002). CFA inspects cross-classifications as described in Section 2.2 and asks whether the number of cases in a cell is greater than or less than expected from some chance model. If the null hypothesis of no discrepancies can be rejected, CFA states that objects belong to a type if more cases than expected were found. If fewer cases were found, objects belong to an antitype. It is a routine result of CFA that only a few types and antitypes emerge, and that the majority of cells does not deviate significantly from expectancy.

Most methods of classification can be used to create exhaustive as well as non-exhaustive solutions. In hierarchical agglomerative dendrograms, a non-exhaustive solution could indicate one or more single-case clusters, each of which is dumped into the poubelle. This can be done accordingly with clusters that contain only two or three objects. In non-hierarchical solutions, one can specify a relatively large number of clusters some of which may then contain only very small numbers of cases. It should be noted that it is not equally likely for each classification method to yield small clusters. Clustering methods that create clusters by letting objects gravitate toward some centroid have a tendency to allow larger clusters to swallow small clusters. Examples of such methods include the centroid method, complete linkage, average linkage, and the popular Ward method (see Section 3.3.3 for more detail).

The Pros and Cons for exhaustive and non-exhaustive classification are obvious. On the pro-side for exhaustive classification is that each object belongs to a group. Thus, the sample subjected to clustering is not reduced in size. However, for some objects, e.g., outliers, it may be artificial to be assigned to a group. On the problematic side for exhaustive clustering is the result that there may be single-case clusters. Often, researchers experience problems with single-case clusters because these are hard to statistically compare with other clusters. For

instance, MANOVA cannot be used for subsequent comparisons when one cell contains only one case.

3.3.5 Decision 5: Stochastic versus Deterministic Clustering

Stochastic models view the data points x_{ki} as realizations of random variables. This seems reasonable if (1) one expects some natural variation within each group A_i; (2) the variables used for creating the classification are measured with error; and (3) the objects, O_1, ..., O_N are a random sample. If all these conditions are met, *stochastic models* allow one to test whether the object set *S* has a group structure. In contrast, *deterministic models* view *S* as a fixed set of objects. The objects' characteristics may vary. However, in deterministic models, these variations are not considered random. As a result, probability statements do not make sense.

The Pros and Cons of these two models are obvious. In favor of stochastic models is that statistical decisions concerning the existence of substructures become possible. A problem with stochastic models is that the null hypothesis used when testing for a particular grouping structure is not always easily specified. In favor of deterministic models is that parametric assumptions are unnecessary. A problem with deterministic models is that the typically existing error structure is ignored.

When asking which of the two approaches, stochastic and deterministic clustering is preferred by users, there is a clear vote for deterministic models. The reasons for this preference are confounded. One reason is that deterministic models may seem easier to interpret. Another reason is that only a few of the general purpose statistical software packages include programs for stochastic clustering (e.g., S-plus).

3.3.6 Decision 6: The Selection of Base Measures

Many clustering methods start from calculating an *N* x *N* matrix that contains information about the relationships between all pairs of objects. Often, these matrices are called *similarity matrices*. Examples of such matrices include correlation matrices, distance matrices, and matrices that contain coefficients that count [in percent] the number of incidences in which two objects' characteristics match. The coefficients used in similarity matrices are called *base measures*. Most popular are measures of correlation, typically Pearson's r, and distance, typically the Euclidean distance. It is well known that *" ... resulting clusters depend more on the underlying similarity criterion than on the physical process of cluster formation"* (Wishart, 1970, p.1). The following paragraphs illustrate this result using the correlation measure, r, and the Euclidean distance as examples. We give two examples. The first example illustrates one shortcoming of Pearson's r, that is, it cannot be calculated when there is no standard deviation. The second example illustrates that clustering based on correlations and clustering based on distances can create classifications that are unrelated to each other such that information about one classification will not carry information about the other.

Correlation measures such as r identify objects as identical if their profiles are parallel. The measure r is not sensitive to differences in standard deviation and mean. Pearson's r is the most widely used measure of correlation. However, it has been criticized because r can be interpreted as the angular distance between the vectors of the two objects under comparison only if the vectors have the same norm (Majone & Sanday, 1971). Distance measures focus on spatial distance, regardless of whether profiles are parallel or not. Correlation and distance scores coincide only if the distance, d = 0. More specifically, we obtain regardless of the size of d

$-1 \leq r \leq +1$ if $d \neq 0$

and

$r = +1$ if $d = 0$.

Consider the data example in Figure 6. The figure displays the four profiles, A, B, C, and D. Figure 6 shows that Profiles A and B are relatively far apart from each other yet parallel. Profile C is parallel to A and B and has a smaller standard deviation. Profile D has a standard deviation of zero, is relatively close to Profile A, and relatively far from Profiles B and C.

Table 4 displays the Pearson r correlations among the four profiles. The measures suggest that the Profiles A, B, and C are parallel. The correlation with Profile D cannot be calculated because $sd_D = 0$. This is indicated by the periods in the last row of Table 4. Table 5 displays the normalized Euclidean Distances among the four profiles. The measures suggest that $d_{AD} < d_{BC} < d_{AB} < d_{AC}$. The distance (dissimilarity) is greatest between A and C and between C and D.

Clustering the data in Figure 6 yields the expected results. Using Ward's method and Euclidean distances for a base measure, Objects A and D are grouped together first, followed by Objects B and C. Using the same method with Pearson's correlation as the base measure, the program is unable to complete the classification because as soon as it tries to evaluate the correlation between Object D and the other objects, it encounters a *missing data point* (see last row in Table 4). Thus, there is no clustering solution. In general, when there is a profile in longitudinal research that suggests no change, it is impossible to include this profile in a cluster solution that is based on correlations.

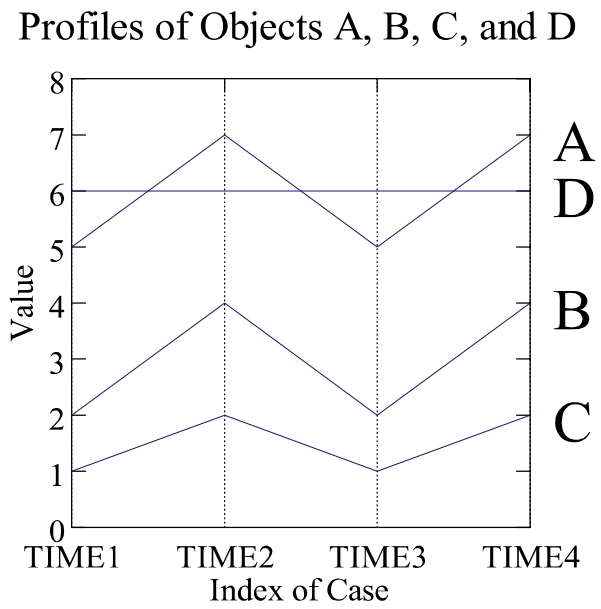

Figure 6:
Profiles of Four Objects

Table 4: Pearson correlations between four Profiles

	A	B	C	D
A	1.000			
B	1.000	1.000		
C	1.000	1.000	1.000	
D	.	.	.	.

Table 5: Normalized Euclidean Distances[5] between four Profiles

	A	B	C	D
A	0.000			
B	3.000	0.000		
C	4.528	1.581	0.000	
D	1.000	3.162	4.528	0.000

In the second example, we show that the base measures of correlations and distance can yield classifications that are independent of each other. A data set with N = 12 cases and t = 5 observation points was created. We call this data set *Chamonix*. The parallel plots of these twelve cases appear in Figure 7.

The figure suggests that, as was the case in Figure 6, the twelve cases differ in both correlation and distance. We now apply the same clustering procedures to these data as to the data in Figure 6. Specifically, we create a hierarchical solution using Ward's method using (1) Euclidean distances and (2) Pearson's correlation r as base measures. Figure 8 displays the dendrogram for the first solution. Figure 9 displays the dendrogram for the second solution.

Figure 8 suggests that, using Euclidean distances, a 2-cluster solution may be most appropriate. Each of the two clusters contains six objects. Fusing the two clusters leads to a major increase of the within-cluster object distances. This is indicated by the distance scale at the bottom of the scale.

Figure 9 suggests that using Pearson's r, a three-cluster solution may be most appropriate. Each of the three clusters contains four cases. Fusing two of the three cluster causes the within-cluster distance, measured in units of correlations, to increase dramatically.

We thus conclude that both using distances and correlations one arrives at relatively clear-cut classifications. However, these classifications are independent of each other. Table 6 displays the cross-classification of the two solutions.

---

[5] The (standard) Euclidean distance between the two objects *f* and *u* is defined as

$$d = \sqrt{\sum (f_i - u_i)^2},$$

where the sum goes over all *i*, that is, all dimensions. The normalized Euclidean distance is defined as

$$dn = \sqrt{\sum \left(\frac{f_i - u_i}{sd_i} \right)^2},$$

that is, by dividing the difference between *f* and *u* in dimension *i* by the standard deviation of this dimension.

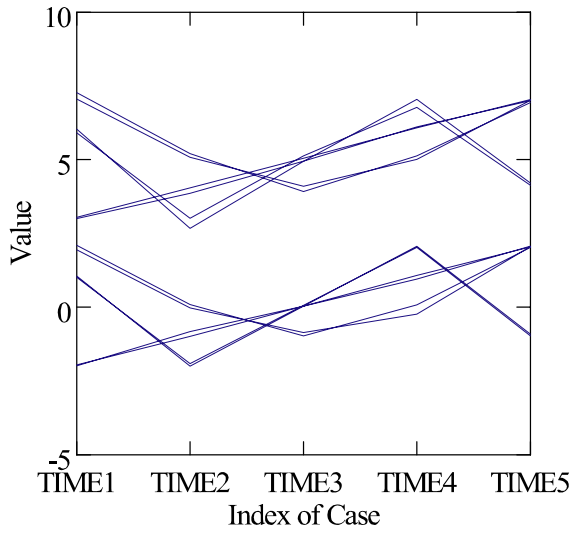

Figure 7:
Parallel plot of the 12 cases in the Chamonix data set

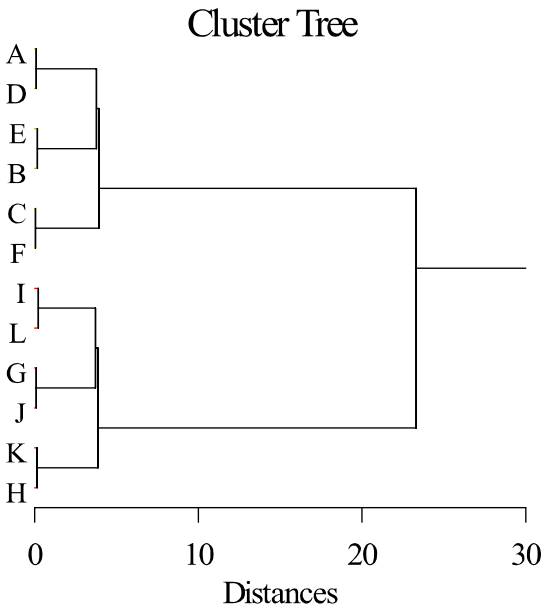

Figure 8:
Classification of the Chamonix data; Ward's method; Euclidean distance

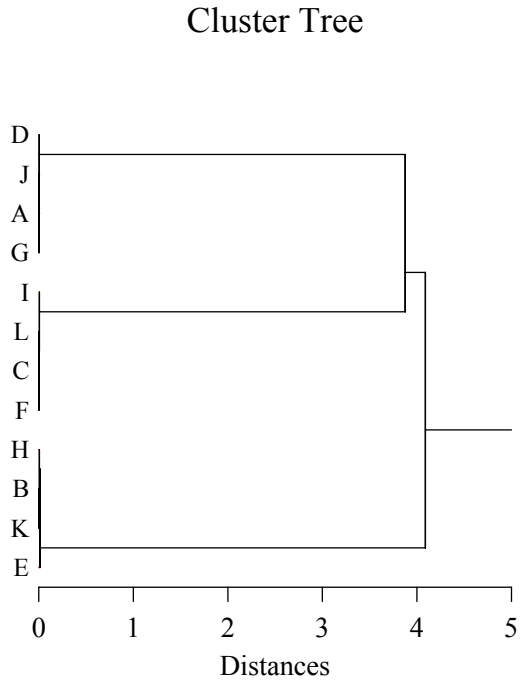

Figure 9:
Classification of the Chamonix data; Ward's method; Pearson's r

Table 6:
Cross-classification of the two classifications of the Chamonix data set

	1	2	3	Total
1	2	2	2	6
2	2	2	2	6
Total	4	4	4	12

The rows of Table 6 display the distribution of cases in the two-cluster solution that is based on the Euclidean distance. The columns display the distribution of cases in the three-cluster solution that is based on Pearson's r. The association between these two solutions can be assessed with the $X^2 = 0.0$ (df = 2; p = 1.0). We thus conclude that the two solutions are independent of each other.

Without additional criteria or knowledge, both solutions are reasonable and defensible. Therefore, when weighing which base measure to use, researchers chiefly rely on substantive information. The Euclidean distance is in many instances a natural choice because it indicates how far objects are from each other in the space of the variables under study. (It should be emphasized that variables must be commensurable before Euclidean distances can be

applied. If variables are not commensurable, the variables with the larger scores and variances dominate the cluster solution.) If, in the context of developmental research, the focus is on the level or magnitude of behavior over time, or, if there exists a natural distinction in level that separates the top from the bottom clusters of developmental trajectories, it seems defensible to use Euclidean or other distance measures as base measures. However, if the focus is on trend characteristics rather than levels, on shape, fluctuation, or curvature similarity, solutions based on correlation coefficients may make more sense.

In some instances, researchers may find it hard to justify a selection of a base measure. In these instances, two strategies can be considered. First, one can create classifications using each of the base measures that are considered. Closer inspection of these classifications may reveal that some are more meaningful than others. An alternative to this approach involves using base measures that combine the characteristics of distance and correlational base measures. An example of such a coefficient is Cattell's (1949) coefficient, r_P, that is sensitive to profile shape, mean differences, and differences in standard deviations. Unfortunately, this coefficient is not included in the best known general purpose statistical software packages.

3.3.7 Decision 7: Convex versus Non-convex Clusters

The shape of clusters is of importance in more than one respect, for two reasons. First, if clusters take a shape that can easily be depicted, e.g., an ellipsoid or a rectangle, verbal description of clusters will also be easy. Second, when new objects are related to an existing cluster structure, one can ask whether these objects are located within the hull that represents a cluster. Therefore, many classification methods create clusters that are convex in shape. Subsets A_i are *convex* if any two objects, O_i and O_j, for $i \neq j$, can be connected by a straight line that is entirely located within A_i. Examples of convex cluster hulls include ellipsoids, squares, circles, rectangles and other types of quadrics (von Eye, 1977; von Eye & Wirsing, 1978, 1980).

There are classification methods, however, that create *non-convex* clusters. For instance, clusters can take the shape of bananas where the connecting straight line between data points can be located outside the cluster hull. Examples of such points include the end points of the banana-shaped structure. The best known classification method that yields non-convex clusters is the single linkage method. Single linkage fuses clusters based on the shortest distance of an element with some other element. Thus, it tends to create chains rather than convex agglomerates. Hartigan (1975) states that "single linkage clusters are famously strung out in long sausage shapes, in which objects far apart are linked together by a chain of close objects" (p. 200) (for graphical representations see Hartigan, 1975, pp. 201 and 202).

In most instances, researchers select convex clusters for they can be described by measures of central tendency or by equation parameters that represent their hull. However, chains can be of interest also, for instance when the chain structure can be interpreted. This is the case, for example, when events such as the "next best step" are grouped, or sequences of decisions where each follows from the preceding one without much reference to the ones before.

3.3.8 Decision 8: Symmetric versus Asymmetric Base Measure

An issue rarely discussed in the context of cluster analysis concerns the symmetry characteristics of base measures. Most typically, researchers employ base measures that are symmetric. Consider, for instance, Pearson's correlation coefficient, r. It is well known that

$r_{AB} = r_{BA}$. In a similar fashion, it holds for the Euclidean distance, d, that $d_{AB} = d_{BA}$, where A and B are two objects. In scaling or in factor analysis, one also assumes that relationships between two objects are symmetric. However, this may not always be the case. Consider the traveling salesperson's situation. She has to determine what the most parsimonious, that is, shortest distance is that allows her to visit all customers. If the distance between Customer A and Customer B is the same in either direction, she faces the standard minimization problem. If, however, the salesperson operates within a one-way street system, the distance from A to B can be much larger than the distance from B to A. This applies in an analogous fashion to measures of similarity. In other words, asymmetric relationships imply that it is possible that $r_{AB} \neq r_{BA}$ or $d_{AB} \neq d_{BA}$.

Typically, researchers assume that similarity relationships are symmetric. For example, the similarity between a BMW and a Porsche is the same, regardless of whether one compares the BMW with the Porsche or the Porsche with the BMW. Consider, however, the similarity between a mother and her daughter or the similarity between an original and the copy. In these cases, the statement that the daughter is similar to her mother (or the copy is similar to the original) naturally makes sense. The inverse statement, that is, that the mother is similar to her daughter (the original is similar to the copy) changes the semantics, if it makes sense at all.

The number of asymmetric measures of distance or similarity is limited. Examples include some PRE measures, for instance, Goodman and Kruskal's (1954) λ, and information theory measures of constraint (Newman & Gerstman, 1952). Matrices that depict symmetric relationships are axial symmetric. There exist log-linear methods for testing axial symmetry in frequency tables (see von Eye & Spiel, 1996).

Thus, clustering distances or similarity measures must take into account the symmetry characteristics of measures. The selection of symmetric over asymmetric measures must be justified from substantive considerations. Unfortunately, to the best of our knowledge, there is not one software package that allows researchers to select asymmetric base measures. The issue has been discussed occasionally in the context of scaling (Kruskal & Wish, 1978) and in the context of definitions of similarity, where symmetry is one of the axioms of similarity (which can be traced back to Frechet, 1906).

3.3.9 Decision 9: Monothetic versus Polythetic Classifications

For the distinction between monothetic and polythetic classifications, we assume that the N objects, O_1, ..., O_N are observed in the p variables, M_1, ..., M_p. The observed measure is x_{ki}, for k = 1, ..., N and i = 1, ..., p. *Monothetic* methods create partitions by only taking one variable into account at a time. To illustrate, consider the two groups, B_1^i and B_2^i, and the binary variable M_i that is used to create the partition. Now suppose also that M_i is binary. Then the splitting into B_1^i and B_2^i can be performed by the following operation

$$B_1^i := \{k \mid k \in A \text{ and } x_{ki} = 1\}$$
$$B_2^i := \{k \mid k \in A \text{ and } x_{ki} = 0\} = A - B_1^i$$

(see Bock, 1974). As a result of this operation the group B_1^i contains all objects O_k 0 α that display a score of $x_{ki} = 1$ in the characteristic under study. Accordingly, one specifies when a variable is quantitative

$$B_1^i := \{k \mid k \in A \text{ and } x_{ki} < c_i\}$$
$$B_2^i := \{k \mid k \in A \text{ and } x_{ki} \geq c_i\} = A - B_1^i,$$

where c_i is a threshold that is either determined before data analysis or determined based on the data.

When a sample is observed in several variables, the partitioning proceeds by taking one variable into account at a time. If variables are categorical, the result can be the same as the result of a cross-classification (see Section 2.2, above). If variables are continuous, the c_i are chosen so that a priori specified optimality criteria are fulfilled. Most typically, the c_i are chosen such that members of a group are as homogeneous as possible (e.g., high correlations or low distances). Classifications are monothetic if "the possession of a unique set of features is both sufficient and necessary for membership in the group thus defined" (Sokal & Sneath, 1963, p. 13).

In contrast, *polythetic* classifications are created by simultaneously taking into account all variables, M_1, ..., M_p. If all variables are taken into account, "there is no single character that is both sufficient and necessary to every member of the group, yet the group possesses a certain unity" (Sneath, 1965, p. 83). Virtually all classification methods available in the current general purpose statistical software packages create polythetic classifications.

When deciding whether to employ monothetic or polythetic classification methods, researchers use information from substantive theory. Monothetic classifications require objects to possess all of a specified set of characteristics to belong to a class (see Section 2.2, above). In contrast, polythetic classifications only require objects to show similarity (lack of distance) to qualify for group membership. On the Pro side for monothetic methods is that they are exhaustive and easy to perform. They typically involve divisive procedures and use binary variables. Results are typically easy to interpret and are often considered useful for diagnostic purposes. On the Con side is that monothetic classifications are considered "unnatural" by many, and they are impractical for many types of variables, for instance, when cut offs are hard to justify. On the Pro side for polythetic methods is that they yield less artificial classifications. However, they are numerically more complex. In fact, when sample sizes are very large, it is still impractical to apply programs for polythetic procedures when they require that the entire matrix of base measures be in the computer memory at the same time. Monothetic and polythetic clustering methods are often used or described in the context of divisive hierarchical methods.

3.3.10 Decision 10: Manifest Variable versus Latent Variable Grouping

The methods of creating groups and clusters discussed thus far in this article operate at the level of *manifest*, that is, observed variables. Cases are grouped based on their relative distances or similarities. The resulting groupings and clusters are described also at the level of manifest variables. In contrast to these methods, there exist methods that use the relations among manifest variables to create *latent* variables, that is, unobserved variables. The latent variables are used to explain the covariation among the observed variables. In this section, we briefly review two variants of latent variable grouping, *latent class analysis* (LCA; Lazarsfeld, 1950; Rost & Langeheine, 1997), and latent class mixture models (LCGM; Muthén, 2001; Nagin, 1999).

LCGM is a method of latent growth curve analysis that models heterogeneous latent classes of trajectories. Two methods have been proposed. Muthén (2001) uses baseline measures to cluster the data into classes. Then, the latent variable structure is used to model class trajectories. The method is Bayesian as the class trajectories are conditional on class membership. For diagnostics, this method uses posterior predictors of class membership.

The second method, the semiparametric, group-based approach introduced by Nagin (1999) can model data based on the zero-inflated Poisson, censored normal, and binary logit distributions. Unobserved latent classes are set to explain individual differences along the aggregated single developmental trajectory in a population. Fixed-effect growth is hypothesized within each class so that all individuals of one latent class are hypothesized to have the same within-class developmental trajectory over time. The trajectories are class-specific, that is, each class has a trajectory that differs from the trajectories of each other class. When data are normally distributed, LCGM tends to extract more classes than for non-normal data (Bauer & Curran, 2003).

LCA of categorical variables uses the concept of *local independence*. The joint probability of the observed frequencies is expressed as the product of the marginal probabilities, given a latent class. In different words, once the latent class is known, the observed variables are independent. Note, however, that local independence is not a common characteristic of all latent class models. Mixed Markov models do not pose the restriction of local independence when they describe response patterns. Mixed Markov Models are of particular importance for the analysis of longitudinal data (Langeheine & van de Pol, 1993).

Looking at the individuals in the analysis, LCA models do share in common that all individuals within a latent class have the same response probabilities for the categories included in the analysis. That is, individuals within the same latent class are treated identical and thus different than individuals in other latent classes. This concept is obviously stricter than the concepts used in cluster analysis in which individuals within a cluster are, on average, more similar to each other than to individuals in other clusters.

From a technical perspective, methods of LCA can be viewed as finite mixture models in which the component distributions are multivariate Bernoulli. Parameter estimation is typically done using maximum likelihood estimation by way of the EM algorithm. Many researchers consider LCA the categorical variable analogue to factor analysis (cf. Molenaar & von Eye, 1994). In the present context of typifying, the analogy to cluster analysis is more interesting: LCA is also considered a method of cluster analysis for categorical variables.

4. A Data Example

In this section, we present a data example that reflects the ten decisions using data from the Lives Across Time (LAT) study. The LAT is an ongoing prospective longitudinal study of adolescent and adult development currently in its 16th year and sixth wave of data collection (for more detail see Windle & Windle, 2001). Participants were recruited from suburban public high school districts in Western New York when they were in 10th and 11th grade (N = 1,219). The sample used in the data example consists of a smaller subset of males who answered all five assessments of cigarette smoking (N = 250). Cigarette smoking was asked four times with a six-month interval (Waves 1 - 4) and followed up once again when they were young adults (Wave 5). Daily cigarette smoking during the last 30 days prior to assess-

ment was reported on a seven-point scale that reflects quantity of smoking: 0 = *None*, 1 = *less than one*, 2 = *1 to five cigarettes*, 3 = *half pack*, 4 = *one pack*, 5 = *one and a half packs*, and 6 = *2 packs or more*. At an aggregate level, cigarette smoking seems to increase over time, and the average level of their smoking is below one cigarette a day (see Table 7). However, aggregation does not make sense since it is generally accepted that there are different groups of people based on their smoking pattern. For example, there are words to describe people of distinctive smoking behavior (e.g., smokers, nonsmokers, experimenters, chippers, chain-smokers) that not only reflect the categorical distinction but also signify different levels of involvement or dependency. It is very plausible that these groups of smokers and nonsmokers have distinctively different developmental trajectories during adolescence and young adulthood. We illustrate the process of finding groups following the ten decisions. SAS 9 (SAS Institute Inc., 2002) was used for the subsequent analyses.

Table 7:
Descriptive statistics for data on daily cigarette smoking among males

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
N of cases	250	250	250	250	250
Minimum	None (0)	None (0)	None (0)	None (0)	None (0)
Maximum	Two packs or more (6)[a]	Two packs or more (6)	One and a half packs (5)	Two packs or more (6)	Two packs or more (6)
Mean	.428	.492	.524	.624	.972
Standard Dev	1.222	1.166	1.159	1.268	1.556

[a]Numbers in parentheses are the ranks used for analyses (see also Table 9)

4.1 Disjoint versus Overlapping Clusters

The first decision asks whether individuals can be members of two or more clusters at the same time (i.e., overlapping clusters) or whether an individual trajectory of smoking should be placed in one and only one cluster (i.e., disjoint clusters). With the current data of smoking, it is reasonable to expect that clusters are disjoint. For example, chain smokers cannot at the same time be chippers. And nonsmokers cannot simultaneously be smokers. Belonging to one cluster precludes one from belonging to other clusters. Therefore, finding disjoint clusters is appropriate for the data.

4.2 Hierarchical versus Non-hierarchical Clusters

The second decision asks whether developmental trajectories of smoking among males can naturally be grouped hierarchically. The current data example is hierarchical in the sense that, for example, once smokers are grouped together from nonsmokers, they can further be grouped into smaller groups of smokers based on the quantity of smoking over time. Assuming that the smoking pattern is relatively stable over time, it makes sense that more homogeneous developmental trajectories can be identified from a mixture of heterogeneous smoking

trajectories. In addition, there are no clear-cut theoretical justifications for the number and kind of smoking trajectories. Thus, a hierarchical cluster analysis suits well.

4.3 Agglomerative versus Divisive Clustering

In essence, agglomerative and divisive clustering methods should result in same results. The only difference is that divisive clustering demands more processing time and memory. So we choose an agglomerative clustering method. From the various agglomerative methods listed previously, we chose centroid clustering. While Ward's method of minimum sum of squares is a popular choice, it tends to prefer a solution of clusters with equal size. Given the data characteristics where uneven size clusters are expected, we chose centroid clustering since it is known to give the best results when the size of clusters is different (Everitt, 1987, cf. Everitt et al., 2001, p.65). In centroid clustering, the distance is defined as the squared Euclidean distance between mean vectors (i.e., centroids).

4.4 Exhaustive versus Non-exhaustive Clustering

Exhaustive clustering involves all data observations including outliers or a few cases that do not fit well with any clusters. We first examined possible outliers using single linkage clustering since this method often yields singletons (i.e., one case per cluster) that have the biggest minimum distance from its nearest neighbor. Figure 10 shows the dendrogram of single linkage clustering analysis and Table 8 displays the last 10 generations of the cluster history that shows chaining and the presence of singletons. The frequency in Table 8 increases by one or two as the number of clusters decreases, which is indicative of chaining and singletons. Reading from top to bottom of the table, it is clear that a few cases were isolated until the last moments when they were pulled toward the existing clusters one at a time, creating chains. From Table 8, observations such as 169, 145, and 44 may be considered as outliers. At the right side of Figure 10, it is visible that a few cases remained as sin-

Table 8:
The Last 10 Generations of the Single Linkage Cluster History

Number of Clusters	Clusters Joined		Frequency	Minimum Distance between Clusters
10	CL11	CL16	241	0.7146
9	CL10	OB147	242	0.7146
8	CL9	OB60	243	0.7989
7	CL8	OB185	244	0.7989
6	CL7	OB198	245	0.7989
5	OB55	OB56	2	0.8752
4	CL6	OB44	246	0.9453
3	CL4	OB145	247	0.9453
2	CL3	CL5	249	1.0719
1	CL2	OB169	250	1.3838

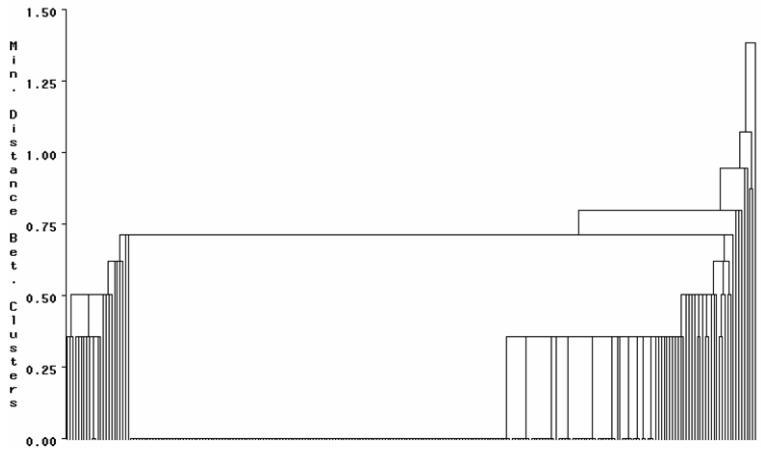

Figure 10:
Single Linkage with Outliers Included ($N = 250$)

4.5 Stochastic versus Deterministic Clustering

gletons until the last moments when they were finally merged one by one with the majority. These cases showed extreme swings in their cigarette smoking pattern over time. For example, observation 169 had a smoking profile that started out as a non-smoker but, within six months, smoked two packs a day only to reduce it to less than one cigarette and then to inch up to 1-5 cigarettes a day in subsequent waves, which reflects changes in and out of smoking status during middle adolescence. Many hierarchical agglomerative methods are sensitive to outliers and therefore we chose to eliminate 1% of observations with the radius of the sphere = 2, assigning the excluded outliers to a *poubelle*.

The fifth decision involves which one of stochastic and deterministic clustering analysis makes more sense. There are several ways to pursue stochastic clustering analysis. Fuzzy clustering produces a solution where some of the cases have probabilities of zero and one that they will belong to a certain cluster and for some other cases the probabilities are in between zero and one so that the case in question can belong to multiple clusters or none. Model-based clustering is also available with MCLUST software (Fraley & Raftery, 2002) written for the S-Plus 6 software package (Insightful Corporation, 2001). Model-based clustering utilizes the expectation-maximization (EM) algorithm for maximum likelihood to determine partitions and uses hierarchical agglomerative clustering solution as initial values. Multivariate normal mixtures are used to describe data with the possible addition of Poisson distribution to model noise or outliers. A third option is to use Nagin's semi-parametric model-based analysis (Nagin, 1999). In the current data example, we do not make any assumptions about the mixing proportions or the underling density distributions (e.g., multivariate normal versus multivariate *t*) as is required in the latter two; we are only interested in finding groups from the observed data without having to make any assumptions. Therefore, we chose deterministic cluster analysis.

4.6 Selection of Base Measure

The base measure can be correlation-based similarity or distance measures such as Euclidean distance for continuous data. Euclidean distance was chosen as a proximity data in the current data example because 1) the exact numerical value on smoking carries more meaning than the relative standing in cigarette smoking and 2) the Euclidean distance measure is typically used as a proximity measure for centroid cluster analysis (Everitt et al., 2001, p. 62).

4.7 Convex versus Nonconvex Clusters

It is difficult to visualize clusters of *p*-multidimensional data. One way to handle is to perform principal component analysis to reduce *p* dimensions into one or two. Principal component analysis allows one to preserve the multivariate structure while condensing *p*-dimensional data into a smaller set of principal components. The current data have five repeated measures of smoking with bi-variate correlations, ranging from 0.516 to 0.865 among them. Principal component analysis resulted in the first two principal components attributable for total 89.7% of the variance of the five repeated measures: 77.9% and 11.8%, respectively for the first and second principal components. A scatter plot between the two principal component scores is shown in Figure 11. From this scatter plot, it appears that 1) there are no apparent banana-like elongated non-convex clusters, 2) there are probably two or more clusters, and 3) a few outliers exist (e.g., the three observations near the bottom of the figure). The outliers are to be eliminated from the subsequent cluster analysis (see also Decision 4.4: Exhaustive versus Non-exhaustive Clustering, and Figure 14).

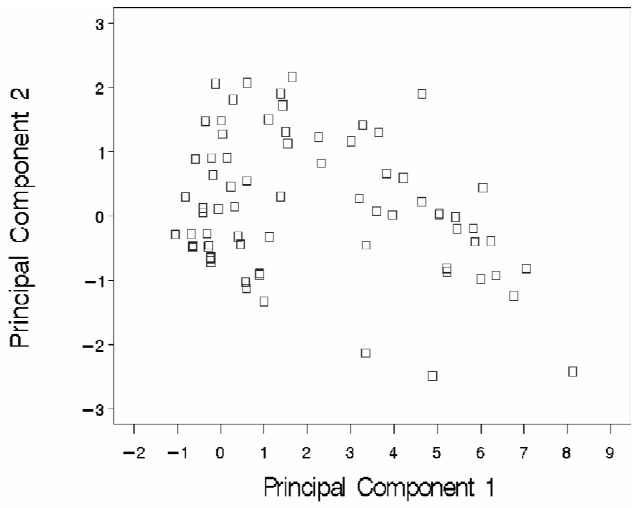

Figure 11:
Scatter plot of principal components ($N = 250$)

4.8 Symmetric versus Asymmetric Base Measure

Typical dissimilarity measures for continuous data are symmetrical base measures, including Euclidean distance or correlation-based measures. From the current data example, Table 9 shows three typical cases. Correlations between cases 1 and 2, between 1 and 3, and between 2 and 3 are .612, .250, and .408, respectively. Squared Euclidean distances between cases 1 and 2, between 1 and 3, and between 2 and 3 are 5, 66, and 49, respectively. Cases 1 and 3 are most dissimilar and cases 1 and 2 are most similar of all pairs of observations. In both of these measures, we assume that the dissimilarity or distance between case 1 and case 2 is the same as the dissimilarity or distance between case 2 and case 1, which makes sense in the current data example. So, symmetrical base measures are suitable for the current data.

Table 9:
Typical Data Points and Symmetrical Measures

Case	W1	W2	W3	W4	W5
1	Never (0)[a]	Never (0)	Never (0)	Never (0)	< 1 (1)
2	Never (0)	Never (0)	Never (0)	1-5 Cig.(2)	1-5 Cig. (2)
3	One pack (4)	Half pack (3)	One pack (4)	One pack (4)	One pack (4)

[a] Numbers in parentheses are ranks used for analyses (see also Table 7)

4.9 Monothetic versus Polythetic Classifications

Monothetic classifications are utilized in divisive hierarchical clustering analysis using binary data while polythetic classifications can be used for both binary and interval-scaled data. S-Plus (Insightful Corporation, 2001) can handle the divisive monothetic and polythetic cluster analysis by the *mona* and *diana* functions, respectively. The current smoking data ranged from zero to six, thus monothetic divisive classification is not suitable. Hierarchical agglomerative cluster method chosen for the data is a polythetic classification method since at each grouping step all variables are considered simultaneously.

4.10 Manifest Variable versus Latent Variable Groupings

Latent variable analyses of classification - Latent class analysis and latent class growth mixture analysis explicitly assume mixtures of densities from which observations spring up. Moreover, the latent variable approach is different from manifest variable analysis in the sense that it factors in measurement errors of the observed variables used in classification analysis and it yields stochastic classification for each individual or object. Thus, the decision of manifest versus latent variable groupings overlaps to a certain extent with the decision of stochastic versus deterministic classifications. In addition to the simplicity of manifest variable cluster analysis (there is no need to estimate mixing proportions and distributional characteristics of the mixture densities as discussed in Decision 4-5), manifest variable analysis makes sense for repeatedly measured data because repeated measures share the scale

characteristics and in part, measurement errors. Moreover, latent variable analysis involves additional theoretical discussion concerning the nature of the latent variables - dimensions or typologies (i.e., categorical versus metrical) since mathematically they are interchangeable at the latent variable level (Molenaar and von Eye, 1994). For the current data example, manifest variable analysis makes sense because we intend to find homogeneous groups in data without making any assumptions on the nature of latent variables, distributions, or measurement errors.

This series of decisions leads us to using centroid cluster analysis with squared Euclidean distance as a proximity measure. Given that all five variables shared the same metric, and variability was approximately consistent across time (see Table 7), raw scores were used (as opposed to weighted or standardized scores), to compute squared Euclidean distance. Eight observations were trimmed based on the decision to eliminate 1% outliers with radius = 2. Figure 13 presents the dendrogram for the results. The history of the last 15 cluster solutions is provided in Table 10. Both the dendrogram and the history of cluster analysis suggest a three-cluster solution. The semipartial R^2 indicates the decrease in the proportion of variance accounted for by joining two clusters, and the squared multiple correlation, R^2 indicates the proportion of variance accounted for by the cluster solution. The three-cluster solution shows that 80.2% of the variance was accounted for. The proportion of variance lost by moving from the four-cluster solution to the three-cluster solution was only 0.027. Values of the cubic clustering criterion and pseudo F show local maxima at three clusters. The pseudo t^2 also shows the optimal level at three clusters.

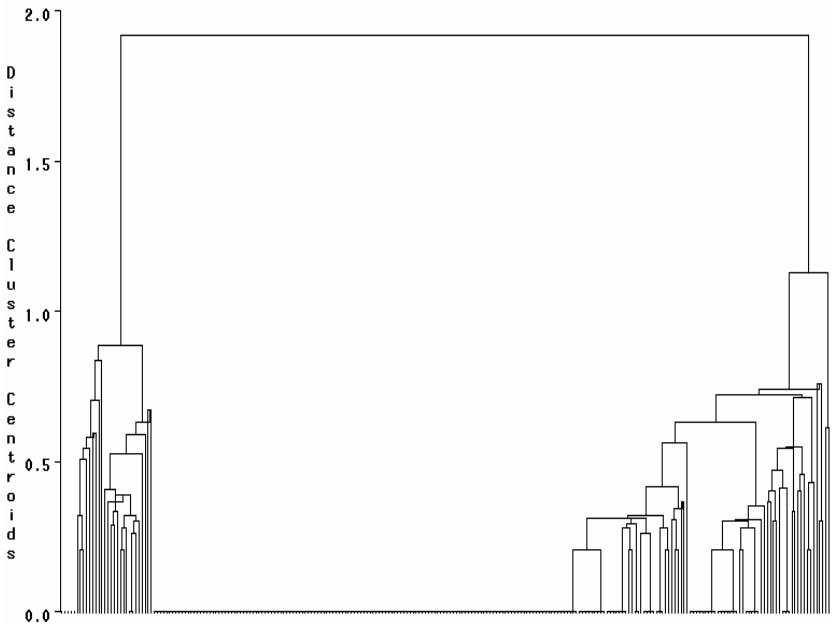

Figure 13:
Dendrogram of Centroid Clustering Analysis

Table 10:
Centroid Cluster Analysis (N = 242)

Clusters Joined		Freq.	Semi partial R^2	R^2	CCC	Pseudo F	Pseudo t^2	Distance
CL23	obs.	14	0.002	.922	7.29	191	2.6	0.54
CL42	CL28	7	0.004	.918	7.22	196	10.2	0.56
CL20	CL14	12	0.008	.910	6.21	193	8.3	0.57
CL24	CL18	181	0.014	.896	4.19	181	52.3	0.58
CL22	obs.	5	0.002	.894	4.87	195	3.1	0.59
CL13	CL16	14	0.005	.889	5.11	206	3.3	0.60
obs.	obs.	2	0.002	.887	6.29	230	.	0.61
CL11	obs.	6	0.003	.885	7.47	257	2.2	0.61
CL15	CL9	16	0.006	.879	6.82	285	5.5	0.62
CL12	CL17	187	0.019	.860	5.94	290	54.1	0.63
CL8	obs.	7	0.003	.857	8.18	355	2.0	0.64
CL10	CL41	32	0.029	.829	8.02	383	30.9	0.66
CL5	**CL7**	**23**	**0.027**	**.802**	**10.4**	**483**	**18.6**	**0.81**
CL6	CL4	219	0.172	.629	2.86	408	265	0.87
CL3	CL2	242	0.629	.000	0.00	.	408	1.91

Note. CCC = Cubic Clustering Criterion

Figure 14 illustrates three clusters and outliers in two dimensional space using principal component scores and Table 11 shows descriptive information. The first cluster, indicated by hollow circles in Figure 14, can be characterized as non-smokers (n = 187). The second cluster, illustrated by solid squares, can be considered as non-smokers during middle adolescence who turned out to be regular smokers in young adulthood (n = 32). The third cluster, solid circles, indicates regular smokers all throughout middle adolescence and young adulthood (n = 23) who smoke somewhere between one half and one pack of cigarettes a day throughout the observation period. Outliers (n = 8), displayed by hollow squares, show an elevated level of smoking as a group but standard deviations across time are also substantial.

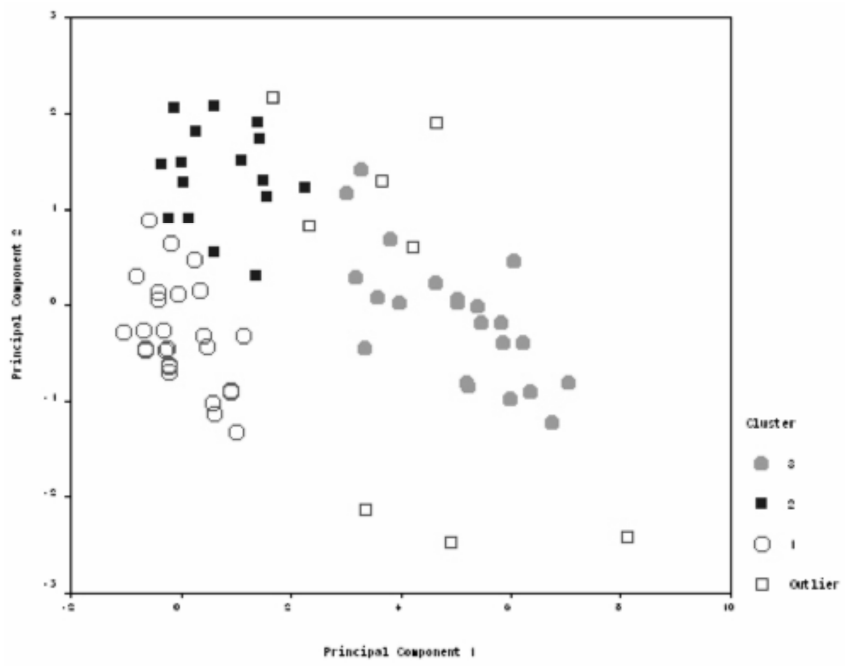

Figure 14:
Scatter Plot of Principal Components by Clusters (N = 250)

Table 11:
Average cigarette smoking among males by clusters (standard errors in parentheses)

Cluster	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Cluster 1 (n = 187)	.06 (.29)	.11 (.36)	.11 (.38)	.12 (.44)	.16 (.46)
Cluster 2 (n = 32)	.13 (.42)	.09 (.30)	.31 (.69)	.97 (1.28)	3.25 (.76)
Cluster 3 (n = 23)	3.22 (1.28)	3.17 (1.03)	3.39 (.72)	3.48 (.73)	3.65 (.71)
Outliers (n = 8)	2.13 (1.96)	3.38 (1.85)	2.75 (1.67)	2.88 (1.81)	3.25 (2.32)

5. Summary and Discussion

This article discusses two methods of classification. The first group of classification is concerned with theory-driven, a priori classifications. Empirical data serve to see whether a priori classifications coincide with natural groupings. The second group of classification methods is concerned with finding groups in data. Ten decisions of classification are proposed and discussed in this article, and empirical data are used to illustrate how the ten decisions help guide selection from the pool of clustering methods. The clusters describe trajectories of smoking in male adolescents. The series of ten decisions takes into account data characteristics and also the mathematical characteristics of clustering methods. Clearly different classification results are possible from the same data. However, none is inherently wrong; each reflects characteristics of the decisions involved in cluster analysis. Discriminant analysis or logistic regression analysis can subsequently be employed to examine group differences.

References

1. Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.
2. American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders (4th ed.). American Psychiatric Association, Washington, DC.
3. Arabie P., & Caroll, J.D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. Psychometrika, 45, 211 - 235.
4. Aude, J.C., Diaz Lazcoz, Y., Codani, J.J., & Risler, J.L. (1999). Applications of the pyramidal clustering method to biological objects. Computers and Chemistry, 23, 303 - 325.
5. Bauer, D.J., & Curran, P.J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. Psychological Methods, 8, 338 - 363.
6. Bergman, L.R., & Magnusson, D. (1997). A person-oriented approach to research on developmental psychopathology. Development and Psychopathology, 9, 291 - 319.
7. Bergman, L.R., & ElKhouri, H.M. (1998). SLEIPNER - A statistical package for pattern-oriented analyses. Vs. 2. Stockholm: Stockholm University, Department of Psychology.
8. Blashfield, R.K., & Aldenderfer, M.S. (1988). The methods and problems of cluster analysis. In J.R. Nesselroade, & R.B. Cattell (eds.), Handbook of multivariate experimental psychology, 2nd ed. (pp. 447 - 473). New York: Plenum.
9. Bock, H.H. (1975). Automatische Klassifikation [Automatic Classification]. Göttingen: Vandenhoeck & Ruprecht.
10. Cairns, R.B., Bergman, L.R., & Kagan, J. (eds.)(1998). Methods and models for studying the individual. Thousand Oaks, CA: Sage.
11. Cattell, R.B. (1949). r_p and other coefficients of pattern similarity. Psychometrika, 14, 279 - 298.
12. Clogg, C.C., Eliason, S.R., & Grego, J.M. (1990). Models for the analysis of change in discrete variables. In A. von Eye (ed.), Statistical methods in longitudinal research, Vol. II (pp. 409 - 441). San Diego, CA: Academic Press.
13. Everitt, B. S. (1987). Introduction to optimization methods and their applications in statistics. Chapman & Hall CRC, London.
14. Everitt, B.S., Landau, S., & Leese, M. (2001). Cluster analysis (4th ed.). London: Arnold.

15. Finkelstein, J. W., von Eye, A., & Preece, M. A. (1994). The relationship between aggressive behavior and puberty in normal adolescents: A longitudinal study. Journal of Adolescent Health, 15, 319 - 326.
16. Fraley, C., & Raftery, A. E. (2002). MCLUST: Software for model-based cluster analysis, discriminant analysis, and density estimation. Unpublished technical report no. 415, Department of Statistics, University of Washington.
17. Frechet, M. (1906). Les dimensions d'un ensemble abstrait. Mathematiques Annales, 68, 145-168.
18. Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross-classifications. Journal of the American Statistical Association, 49, 732 - 764.
19. Gutiérrez-Peña, E. (2003). Bayesian classification methods. Psychology Science; this issue.
20. Hartigan, J.A. (1975). Clustering algorithms. New York: John Wiley.
21. Hilgard, E.R., & Bower, G.H. (1975). Theories of learning. Englewood Cliffs, NJ: Prentice Hall.
22. Indurkhya, A., & von Eye, A. (2000). The power of tests in Configural Frequency Analysis. Psychologische Beiträge, 42, 301 - 308.
23. Kruskal, J.B., & Wish, M. (1978). Multidimensional scaling. Beverly Hills, CA: Sage.
24. Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. Sociological Methods & Research, 18, 416 - 441.
25. Lienert, G.A., & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. Educational and Psychological Measurement, 35, 231 - 238.
26. Magnusson, D. (1998). The logic and implications of a person-oriented approach. In R.B. Cairns, L.R. Bergman, & J. Kagan (eds.)(1998). Methods and models for studying the individual (pp. 33 - 63). Thousand Oaks, CA: Sage.
27. Majone, G., & Sanday, R.P. (1971). On the numerical classification of nominal data. In P. Kay (ed.), Explorations in mathematical anthropology (pp. 226 - 241). Boston.
28. Molenaar, P.C. M., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), Latent variables analysis - Applications for developmental research (pp. 226 - 242). Newbury Park, CA: Sage.
29. Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling. In L.M. Collins, & A.G. Sayer (eds.), New methods for the analysis of change (pp. 291 - 322). Washington, DC: American Psychological Association.
30. Nagin, D.S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. Psychological Methods, 4, 139 - 157.
31. Newman, E.B., & Gerstman, L.S. (1952). A new method for analyzing printed English. Journal of Experimental Psychology, 44, 114 - 125.
32. SAS 9 (SAS Institute Inc., 2002). The SAS System for Windows. SAS Institute Inc., Cary: NC.
33. Sneath, H.A. (1965). A method for curve seeeking from scattered points. Computer Journal, 8, 383 - 391.
34. Sokal, R.R., & Sneath, H.A. (1963). Principles of numerical taxonomy. San Francisco: Freeman.
35. S-Plus (Insightful Corporation, 2001). S-Plus 6 for Windows Guide to Statistics. Insightful Corporation, Seattle: WA.
36. von Eye, A. (1977). Über die Verwendung von Quadriken zur einbeschreibenden Klassifikation. Biometrical Journal, 19, 283-290.

37. von Eye, A. (1990). Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications. Cambridge, UK: Cambridge University Press.
38. von Eye, A. (2002). Configural Frequency Analysis: Methods, models, and applications. Mahwah, NJ: Lawrence Erlbaum.
39. von Eye, A., & Bergman, L.R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. Development and Psychopathology, 15, 553 - 580.
40. von Eye, A., & DeShon, R.P. (1998). The highly gifted: definitions and methodological implications. High Ability Studies, 9, 23 - 41.
41. von Eye, A., & Schuster, C. (2002). Log-linear models for change in categorical variables. Applied Developmental Science, 6, 12 - 23.
42. von Eye, A., & Spiel, C. (1996). Standard and non-standard log-linear symmetry models for measuring change in categorical variables. The American Statistician, 50, 300 - 305.
43. von Eye, A., & Wirsing, M. (1978). An attempt for a mathematical foundation and evaluation of MACS, a method for multidimensional automatical cluster detection. Biometrical Journal, 20, 655 - 666.
44. von Eye, A., & Wirsing, M. (1980). Cluster search by enveloping space density maxima. In M. M. Barritt & D. Wishart (Eds.), Compstat 1980, Proceedings in Computational Statistics (pp. 447 - 453). Wien: Physica.
45. von Linné, C. (2003). Linnäus' Philosophica Botanica. New York: Oxford University Press (Translation by S. Freer).
46. Windle, M., & Windle, R.C. (2001). Depressive symptoms and cigarette smoking among middle adolescents: prospective associations and intrapersonal and interpersonal influences. Journal of Consulting and Clinical Psychology, 69, 215 - 226.
47. Wishart, D. (1970). The treatment of various similarity criteria in relation to Clustan 1A. St Andrews, Scotland.
48. Wishart, J. (1987). Clustan user manual, 4th ed. Edinburgh: University of St. Andrews.