Faculteit der Economische Wetenschappen en Econometrie

SERIE RESEARCH MEMORANDA

Classification Techniques in Quantitative Comparative Research: A Meta-Comparison

Peter Nijkamp
Piet Rietveld
Laura Spierdijk

Research Memorandum 1999-56

November 1999

vrije Universiteit amsterdam

Classification Techniques in Quantitative Comparative Research: A Meta-Comparison

Peter Nijkamp Piet Rietveld Laura Spierdijk

MASTER-POINT
Department of Spatial Economics
Faculty of Economics and Econometrics
Free University
De Boelelaan 1105
1081 HV
Amsterdam
The Netherlands

Abstract

This paper emphasizes the importance of quantitative comparative research in the social sciences. For that purpose a great variety of modem classification methods is available. The paper aims to give a selective overview of major classes of these methods and highlights the advantages and limitations of these methods.

Pn895ls

1 Introduction

In our information age there is often an abundance of fragmented data and an overwhelming presence of segmented methodological approaches. There is also an increasing demand for more integrated scientific insights and perspectives. And clearly, there is a need for more research synthesis in the empirical sciences. Empirical research is often based on controlled experimentation, as is witnessed in the research methodology in the natural sciences. In the social sciences however, it appears to be very difficult to apply this research methodology, as both contextual (environmental) conditions and behavioural factors are subject to change. As a consequence, we have witnessed in recent years the emergence of comparative case study research with a view to the identification of common knowledge patterns from distinct classes of information [1], [11]. There is an increasing recognition of the added value of empirical research synthesis in the social sciences.

An important new research methodology in social science research is meta-analysis. Although meta-analysis has originally a limited character, viz. a statistical analysis of previously undertaken quantitative case studies, it is increasingly recognized that meta-analysis is essentially a methodological approach focussed on quantitative, statistical research synthesis. Comparative research on previous research findings is essentially a particular type of meta-analysis.

Comparative research aims to bring to light both common and contrasting elements characterizing a set of phenomena under investigation. Such a research activity serves to gain additional knowledge from syntheses of research findings from previous studies and may lead to generalizable or transferable results. In this sense, comparative research has also a great ability for conditional predictions. Such predictions may be continuous in nature, but -given the uncertainty in many measurement procedures- they often relate to interval data. In many situations the attributes of a phenomenon are represented in distinct classes, so that then comparative research boils essentially down to classification analysis. This paper describes some of the classification techniques that may be useful in a comparative research context.

We will focus here on quantitative, numerically-determined research findings. The starting point of our analysis is a data set, consisting of *n objects*, each object being described by several *variables* or *attributes*. Ideally, the attributes of a phenomenon should be measured on a ratio scale, but also a lower level of measurement can be considered. Attributes having a finite number of possible qualitative values may be transformed into a numerically identifiable form using the method of *categorizing.* For example, if we consider a collection of cars and we wish to look at the distinguishing attribute 'colour', then possible nominal characteristics of the colour could be ‘red’, ‘green’ and ‘blue’. These qualitative values can be transformed into a pseudo-quantitative (nominal) scale by setting e.g. 1 = ‘red’, 2 = ‘green’ and 3 = ‘blue’. However, some classification techniques may yield specific results depending on the way of categorization; that is, in principle different categorizations may lead to different classifications. We will return to this subject when discussing each classification technique considered in the present survey.

In social science research, we often make a distinction between *dependent* and *independent* variables (attributes) so as to allow for explanatory analysis. If such a distinction is possible, there is usually often only one dependent variable. However, we do not exclude the presence of multiple dependent variables in our analysis.

Classification is a prominent research activity in many sciences. And hence, it is no surprise that there is a wide variety of classification methods. The classification techniques that are considered in this paper are the following:

- rough set theory;
- fuzzy set theory;
- cluster analysis;
- discriminant analysis;
- logit and probit models;
- neural networks;
- two-way tables;
- Poisson-regression model.

Clearly, this collection of classification techniques does not pretend to be exhaustive. There exist, of course, other classification techniques which might also be useful in comparative research in the social sciences.

The aim of this paper is to offer an overview of the various classification methods with a view to identifying their weak and strong points in comparative research. This paper is organized as follows. Section 2 discusses rough set theory. In Section 3 we consider fuzzy set theory. The focus of Section 4 is on cluster analysis, while in Section 5 discriminant analysis is discussed. Logit and probit models are the main theme of Section 6, while Section 7 is devoted to neural networks. Section 8 discusses two-way tables, and next Section 9 considers the Poisson-regression model. In Section 10 the methods are compared and this comparison is summarized in a survey table. Section 11 discusses - by way of meta-synthesis - how to find the appropriate classification technique in a specific situation. Finally, in Section 12 a summary and some methodological conclusions are offered for further research on comparative analysis.

2 Rough Set Theory

The first classification technique considered here is fairly recent. Rough set theory (henceforth RST) has its origins in artificial intelligence, and has proven to be useful in several sciences such as decision analysis, economics, econometrics and statitistics. RST is essentially a non-parametric classification technique, having the property that it is *not* based on stochastic data. This section will only provide a short summary of the most important features of RST. For a more detailed description, we refer to e.g. [1], [12], [13] and [14]. The number of publications on RST is rapidly increasing.

2.1 A description

RST is essentially based on a deterministic interval classification. The starting point for RST is a finite set *U* of objects and a finite set A of attributes. Qualitative or nominal attributes should be transformed into quantitative ones as described in the previous section. Such a transformation does *not* influence the classification results. Moreover, it is also required that the values of continuous-valued attributes are transformed into a finite number of distinct classes. For example, consider the attribute 'yearly income' of a person. In principle, this attribute can take any non-negative value. The values of this attribute can be categorized e.g. in the following way:

- 1 = $0 – $ 10.000;
- ***2*** = $10.000 – $ 100.000;
- ***3*** = $ 1.000.000 or more.

At the outset, it should be emphasized that any transformation *does* influence the classification. A sensitivity analysis may be helpful to investigate the robustness of the classification. However, for the time being this problem will be left aside for the sake of simplicity. For more information on this topic we refer to [14].

In the RST terminology, attributes are usually split up into so-called *condition* and *decision attributes.* Condition attributes fulfil the same role as *independent* variables and decision variables have the role of *dependent* variables. We will now concisely describe the essence of RST.

The main principle of RST is the *indiscernibility equivalence relation.* Let ***P*** be a subset of the set of attributes A. Then we call two objects P-indiscernible, if and only they have the same values for all attributes in ***P.*** This equivalence relation generates a partition of the set of objects *U* into equivalence classes of P-indiscernible objects, to which we refer as *P*-elementary sets.

With respect to each subset $X \subset U$ we define the P-lower approximation of *X*, denoted by $\underline{P}X$, as the collection of P-elementary sets which are contained in *X*, and the P-upper approximation, denoted by $\overline{P}X$ as the collection of P-elementary sets which have at least one element in common with *X*. We note that $\underline{P}X$ c $X \subset \overline{P}X$, and hence, objects in $\underline{P}X$ belong with certainty to *X*, while objects in $\overline{P}X$ might belong to *X*.

Now the idea of RST is to approximate each subset X of U by means of the pair $(\underline{P}X$, $\overline{P}X$), with respect to some subset P of A. In case $\underline{P}X = \overline{P}X$, the set X is the union of P-elementary sets. In such a case, one can state with certainty whether an object belongs to Xor not, by considering only the set of attributes P. Moreover, with respect to any subset P of A, we can determine the *accuracy of the approximation* of X as the share of elements of $\underline{P}X$ in $\overline{P}X$. If U is partioned into $U = \{U_1, . . . , U_n\}$ the quality of this classification can be given as the sum of these shares of all sets of this partition.

Taking for granted now the classification of U partitioning the set U into $U = \{ U_1, ... U_n\}$ and the quality of this classification with respect to P, the goal of RST is to find a minimal set of attributes $R \subset P$ that gives the same quality of classification as P. This concept is known in RST *as attribute reduction.* Attribute reduction results in so-called *reducts* of P. The intersection of all reducts of P is known as the core of P. Finally, the total set of the objects may be classified without losing any information, by using only the attributes of R. On the other hand, leaving out one of the core attributes will clearly affect the quality of the classification. Based on these reducts of P, RST derives ultimately decision rules which are a statement of the form '*if* the attributes of the reduct have these values *then* the decision attributes have a given value'. Decision rules can in principle also be generalized to new objects and are thus suitable for transferability and prediction.

2.2 Conclusion

In this brief survey, we have discussed RST which is essentially a non-parametric technique that uses only the deterministic interval information incorporated in the data itself. No a priori assumptions are made about the underlying distribution of the data. The key feature of RST is the indiscernibility relation. RST has various applications to classification problems. By computing reducts of the condition attributes, the relevance of each condition attribute can be evaluated. Attributes not belonging to any reduct may be considered as irrelevant and can thus be left out. Clearly, the reliability of this operation is critically dependent on the representativeness of the findings in the various case studies considered. Leaving out all irrelevant variables leads to a minimal set of attributes with the same quality of description as the original set of attributes. The intersection of all reducts yields a core of attributes from which no variable can be eliminated without deteriorating the quality of the classification. The core variables can be seen as the most important variables. Finally, relationships between objects can be described by means of rules having the form of 'if.. .then' statements. Thus, RST has quite some features that make it interesting as a tool for classification analysis, and hence also for comparative research on case studies with interval results.

3 Fuzzy Set Theory

Another important and popular classification method is based on fuzzy sets. In this section fuzzy set theory (abbreviated as FST) will be concisely discussed. Similar to RST, FST has its roots in artificial intelligence, but it has also applications in other sciences such as psychology, decision analysis, economics, econometrics and statistics. For a detailed overview of fuzzy set theory and its applications, we refer to [10], [15]. Of course, there are numerous publications on FST.

3.1 A description

In ordinary set theory an element x may or may not belong to a set A; that is, there are exactly two mutually excluding possibilities

$$x \in A \quad \text{or} \quad x \notin A$$

In fuzzy set theory, the relation ‘belonging to a set’ is extended towards linguistic information of an imprecise nature. This extension is based on real-life situations, where often it is unclear whether an object belongs to a set or not. For example, if one considers an aged person, one may wonder whether he or she belongs to the class of people being older than 80. If one had to give a number $0 \leq a \leq 1$, indicating the *possibility* that the person belongs to that class, one might perhaps give a = 0.6. This is exactly the aim of FST.

To clarify FST more precisely, let X be a set of objects and let $A \subset X$. The *membership function* $f_A : X \rightarrow [0,1]$ associates with each $x \in X$ a number in the interval $[0,1]$. This number indicates the *grade (or degree) of membership* of x in A. The set A, with a corresponding membership function f_A , is called a *fuzzy set.* A fuzzy set is empty if and only if its membership function is identically zero on X. Two fuzzy sets A and B are equal, written as A = B, if and only if $f_A(x) = f_B(x)$ for all $x \in X$. The complement of a fuzzy set A is denoted by A’ and its membership function is defined as

$$f_{A'}(x) = 1 - f_A(x).$$

A is a subset of B (A c B) if and only if $f_A(x) \leq f_B(x)$ for all $x \in X$. The union of two fuzzy sets A and B is again a fuzzy set C = A u B with a corresponding membership function:

$$f_C(x) = \max\{f_A(x), f_B(x)\}.$$

The intersection of two fuzzy sets A and B is a fuzzy set C = A n B with a membership function:

$f_C(x) = \min\{f_A(x), f_B(x)\}$.

Other operations on fuzzy sets can be defined in a similar way, but we will confine ourselves to the above operations. FST is often used in decision analysis with imprecise information. Its major advantage is its ability to incorporate linguistic statements which do not have an unambiguous numerical meaning.

3.2 Conclusion

FST provides an extension of the set-theoretical relationship of an object ‘belonging to a set’. In FST it is possible that an object belongs to different classes with various degrees of memberships. This feature makes ST appropriate for classification problems with *vague classes.* Given a collection of *fuzzy* classes and corresponding membership functions, the degree of membership of each membership of each object can be obtained, yielding a *fuzzy* classification. The problem of fuzzy classification then reduces to the problem of finding the fuzzy classes and the corresponding membership functions; for further information we refer to [15]. Especially in exploratory classification analysis FST may play an important role.

4 Cluster Analysis

A standard tool in classification analysis is clustering. The main feature of cluster analysis is that a certain type of structure is imposed on the data. This can be useful, if there is not sufficient information about the underlying patterns in the available data. In such cases cluster analysis can serve as a first exploration of the structure in the data. This section will discuss some important standard clustering methods. However, it should be noted that the class of clustering methods is quite broad. For a good overview of existing clustering algorithms, see e.g. [6] and [8]. Contributions to cluster analysis can be found in almost all disciplines.

4.1 A description

The starting point in cluster analysis is a data set consisting of n objects or points, described by several quantitative variables. There are no problems in transforming qualitative variables into quantitative ones. There are -broadly speaking- three types of clustering methods:

- **Partitioning methods.** In these methods, the aim is to partition the set of n objects into a specified number of disjoint groups, say m, so that each object belongs to one and only one group. For each value of m, one seeks a partition which is optimal in terms of the a priori stated clustering criterion.
- **Hierarchical methods.** One is often interested in investigating the structure in the data at different levels; in particular, one may be interested in how the groups in a partition are

related to each other. Hierarchical cluster analysis aims to address this question. Hierarchical methods are a special case of partitioning methods.

- **Clumping methods.** The groups or classes produced by partitioning methods are normally supposed to have no members in common with each other. Sometimes, this condition is unnecessarily restrictive. In *clumping methods* the groups are allowed to overlap. Such overlapping groups will be called *clumps,* while a division of the set of n objects into clumps such that each object belongs to at least one clump, is called a *covering* of the data set.

In order to illustrate the above distinction, an example of both a hierarchical clustering method and a partitioning method is given below.

- **Single linkage method (hierarchical method).** Suppose we have a data set consisting of n objects, where object i is described by k numerical variables. Let d_{ij} denote the *dissimilarity* between two objects i and j[1] :

$$d_{ij} = \sum_{l=1}^{k} (x_{il} - x_{jl})^2 .$$

Two objects i andj are defined to belong to the same single link cluster at level h if there exists a chain of m • 1 intermediate objects, $i_1, i_2, \ldots, i_m$, linking them such that

$$d_{i_k, i_{k+1}} \le h, \; k = 0, 1, \ldots, m - 1 \qquad (1 \le m \le n-1),$$

where $i_0 = i$ and i_m = j. The value of h controls the scale of the investigation The single link clusters have the important property of being invariant under any monotone transformation of the dissimilarities.

- **Sum of squares method (partitioning method).** This method can be used for the classification of objects which can be presented as points in Euclidean space. Let x_{ik} $(i = 1,\ldots, n;\ k = 1,\ldots, p)$ denote the k^{th} coordinate of the i^{th} point, P_i. The aim is to partition the set of n points into g groups so as to minimize the total within-group sum of squares about the g centroids. In other words, if the centroid of the m^{th} group, which contains the n_m points $\{Pm_i\}_{i=1}^{n_m}$ has coordinates

$$z_{mk} = \frac{1}{n_m} \sum_{i=1}^{n_m} x_{m_i k} \qquad (k = 1,\ldots, p)$$

and if the within group sum of squares of the m^{th} group is

$$S_m = \sum_{i=1}^{n_m} \sum_{k=1}^{p} (x_{m_i k} - z_{mk})^2,$$

then the aim is to find a partition which minimizes

$$S(g) = \sum_{m=1}^{g} S_m.$$

To find this partition, two algorithms can be applied: the agglomerative or the iterative algorithm. For more details we refer to [6].
These are just two illustrative application possibilities. The literature on clustering algorithms is rich and it is virtually impossible to give here a representative survey.

4.2 Conclusion

Many classification methods use clustering techniques. It has been shown in this section that cluster analysis can explore the nature of the data by identifying a certain structure from the data.. Using one of the various clustering methods, a data set can be subdivided into specific classes consisting of elements that 'resemble' each other according to some common, given criterion. Hence, cluster analysis might be useful to classify relatively 'raw' data in an exploratory comparative analysis.

5 Discriminant Analysis

Discriminant analysis is another technique for comparative research. In this section we will discuss some classification algorithms from discriminant analysis. Loosely speaking, discriminant analysis seeks to find a *discriminant function* that serves as a rule for finding the class an object belongs to. As a condition for applying discriminant analysis the classes have to be known in advance. Discriminant analysis serves as a tool for determining the class an object belongs to, *not* for determining the different classes. In this section we will briefly discuss two discriminant algorithms: the Bayesian minimum error rule and the minimax rule. Other discriminant algorithms can be found in [5], [6] and [8]. Discriminant analysis is certainly a standard technique in comparative social science research.

5.1 A description

Similar to the situation considered in the previous sections, the starting point in this section is a data set consisting of n objects, which are described by a set of attributes. It is assumed that all attributes are real-valued. Qualitative attributes can in principle be transformed into qualitative ones, but this seriously affects the results of the classification. In case of qualitative attributes, another type of discriminant analysis should preferably be used, namely *discrete discriminant analysis; see e.g.* [5].

5.1.1 Bayesian minimum error rule

Suppose we have a data set consisting of n objects. Each object belongs to one and only one of the (disjoint) classes w_k, $k = 1,\ldots,N$. Each object x is characterized by means of a k-dimensional vector of attribute values, x. These attributes assume values in a set $\Omega \subset R^k$. The set Ω can be partitioned into N subsets Ω_k, $k = 1,\ldots,N$. We define a *decision rule* as a statement of the form:

$$x \in \Omega_k \Rightarrow x \in w_k.$$

Suppose we know the probability that an object belongs to w_k. We denote this probability by $P(w_k)$. We call these probabilities the *a priori probabilities,* as they do not depend on x and are known before we make any observation. Having a vector x of information on the attribute values of each object x, we can use the following rule to classify each object:

$$\forall j \neq k \qquad P(w_k|x) > P(w_j|x) \Rightarrow x \in \Omega_k.$$

This rule is known as *Bayes minimum error rule.* The related probabilities $P(w_k | x)$ $(k = 1,\ldots,N)$ are known as the *posteriori probabilities,* as they can only be calculated if the values of x are known. We can compute them by means of the *Bayes theorem:*

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)}.$$

The decision rule then simplifies to

$$p(x/w_k)\,P(w_k) > p(x/w_j)\,P(w_j), \qquad j \neq k \Rightarrow x \in \Omega_k.$$

Unfortunately, the probabilities $p(x | w_i)$ and $P(w_i)$ are not always known. In that case, they need to be estimated from the data.

5.1.2 Minimax rule

Although the Bayes minimum error rule minimizes the overall error, in fact we might be interested in some other relevant assignment criterion. So far we have assumed that misclassifying an w_i-object as an w_k-object for $i \neq$ k is the same for all i and k. Clearly, this is not always the case. This concept has been formalized in terms of a cost function, C_{ij}, which is the cost of misclassifying an object from class w_i as belonging to class w_j. If $x \in w_i$, then the expected cost equals

$$r_i = \sum_{j=1}^{N} C_{ij} \int_{\Omega_j} p(y|w_i)dy.$$

Thus the overall expected cost r equals

$$r = \sum_i r_i P(w_i) = \sum_i \sum_j \int_{\Omega_j} C_{ij} P(w_i)\, p(x|w_i)dx$$

$$= \sum_i \int_{\Omega_j} \{\sum_i C_{ij} P(w_i)\, p(x|w_i)\}dx.$$

This will be minimized if we define Ω_k. such that $x \in \Omega_k$, whenever

$$\forall j \neq k \quad \{\sum_i C_{ik} P(w_i) p(x|w_i\} < \{\sum_i C_{ij} P(w_i) p(x|w_i\}.$$

This is the *Bayes minimum risk rule.* It sometimes happens that we do not know the a priori probabilities. Since each of the three decision rules above uses these probabilities, this may pose a problem, especially when the sample sizes for each class are not proportional to the class probabilities. The minimax rule is designed to minimize the maximum possible risk. The risk r equals

$$r = \sum_i P(w_i) \sum_j \int_{\Omega_j} C_{ij}\, p(x|w_i)dx.$$

For the sake of simplicity, we assume that we only have two classes. Then r is a linear function of $P(w_1)$ and therefore the maximum occurs when $P(w_1)$ = 0 or 1. Thus the maximum is either

$$\sum_j \int_{\Omega_j} C_{2j}\, p(x|w_2)dx, \text{ or } \sum_j \int_{\Omega_j} C_{1j}\, p(x|w_2)dx.$$

If we assume the cost function to be such that $C_{11} = C_{22} = 0$, then the maximum becomes either

$$\int_{\Omega 1} C_{11} p(x|w_2)dx, \text{ or } \int_{\Omega 2} C_{12} p(x|w_2)dx.$$

And thus

$$\max \left(\int_{\Omega 1} C_{11} p(x|w_2)dx, \int_{\Omega 2} C_{12} p(x|w_2)dx \right)$$

takes it minimum value if

$$\int_{\Omega 1} C_{11} p(x|w_2)dx = \int_{\Omega 2} C_{12} p(x|w_2)dx.$$

This formula can easily be generalized to the case of k classes, for $k = 1,\ldots,N$. We can write the above rules more generally as

$$p_k(x) > q_j(x), j \neq k \Rightarrow x \in \Omega_k.$$

We call the function q a *discriminant function,* which explains the name *discriminant analysis.* This approach has gained a high popularity in many fields of social science research.

5.2 Conclusion

In this section two discriminant methods have been discussed: the Bayesian minimum error rule and the minimax rule. In general, discriminant analysis seeks to find a discriminant function that serves as a rule for deciding on the class an object belongs to. To apply discriminant analysis, the classes should be known in advance. Discriminant analysis then serves as a tool for deciding which class an object belongs to. Seen from this perspective, discriminant analysis is a powerful tool in comparative research.

6 Neural Networks

A more recently developed research methodology is neural networks. Neural network analyses originates also from artificial intelligence. As in the case of RST, they have proven to be successful in other sciences, such as economics, econometrics and statistics. In this section a short description of the main features of neural networks will be given; for further details the reader is is referred to [3], [4], [9]. There is a rising tide of publications on neural networks.

The starting point in this section is again a data set consisting of n objects, described by a set of characterizing attributes. The attributes should be transformed into a quantitative form, while real-valued attributes are allowed in neural networks.

6.1 A description

Generally speaking, a neural network consists of a set of computational units, often called cells, and a set of one-way data connections. At certain times a unit examines its inputs and computes a signed number, called an *activation,* as its output. The new activation is then passed along those connections leading to other units. Each connection has a signed number, called a weight, that determines whether an activation that travels along it influences the receiving cell to produce a similar or a different activation according to the sign (+ or -) of the weight. The size of the weight determines the magnitude of the influence of a sending cell's activation upon the receiving cell; thus a large positive or negative weight gives the sender's activation a more significant effect on the receiving cell than a smaller weight.

Neural networks are built as a parallel of the functioning of the human brains. The cells correspond to our *neurons,* an activation corresponds to *neural firing rates,* connections correspond to *synapses* and connection weights correspond to *synaptic strength.*

There are many types of neural networks, e.g. backpropagation networks, radial bias networks and Hopfield networks. In global terms, they can be subdivided into two classes: networks that need *supervised learning* and networks that need *unsupervised learning.* Supervised learning consists in showing the network both the input and the desired output, whereas unsupervised learning only needs the input. In the sequel, we suppose that we have to do with a supervised learning network.

The main question we want to answer is how to use a neural network for classification purposes. Suppose we have a data set consisting of n objects, characterized by m independent variables (and their values) and k dependent variables (usually k = 1). Firstly, we need to *train* our neural network. That means that we show a specific part of our data set (e.g. 75% of it), randomly chosen to the network (both independent and dependent variables). In this way we can train the network in such a way that it 'predicts' the value of the dependent variables as good as possible. Then we show the test set to the network, but this test set only consists of the values of the independent variables. The network then predicts the values of the dependent variables and it is to the user to compare those values to the real values of the dependent variables. If the network classifies most elements (e.g. 95%) of the test set correctly, then the network will probably work satisfactorily on other test sets as well.

Neural networks have been applied in many choice and management situations, especially in case of large data sets. They have proven to be a powerful tool in exploratory research, and may be helpful for classification analysis.

6.2 Conclusion

In this section neural networks were discussed. Neural networks are built according to the human brains and consist mainly of interconnected cells. Neural networks allow for the prediction of the dependent variables if the values of the independent variables are given. Thanks to their computational power, neural network can effectively deal with large datasets.

7 Logit and Probit Models

The next class of methods dealing with categorical variables is the class of logit and probit models. Logit and probit models belong to the family of *discrete choice models.* In these cases, the independent variables are allowed to be real-valued, while the dependent variables must be binary - or at least categorically - valued. Although the binary restriction on the dependent variables may seem quite restrictive, this is certainly not the case, as also nested approaches are allowed. The collection of binary and categorically valued variables incorporates not only the broad class of all variables reflecting a 'yes or no' answer, but also qualitative multi-state responses (e.g., in data bases for survey questionnaires). A survey of these methods is contained in [2].

This section will first consider the one-dimensional logit and probit model. Then the one-dimensional case will be generalized to the multi-dimensional case. These methods have become very popular in the modern statistical and econometric literature.

7.1 A description

Suppose we have observation $y_1, \ldots, y_n$ of a dependent variable y that can only take values in $\{0,1\}$. Furthermore, suppose this variable y is determined by a response variable y^* in the following way. The response variable y* linearly depends on a vector of independent variables x called regressors, i.e.

$$y^* = \beta' x + u$$

Let now y be determined by x according to the condition:

$$Y = \begin{cases} 1, & \text{if } y^* > 0 \\ 0, & else. \end{cases}$$

Combining (1) and the above expression yields

$$P(y_i = I) = P(u_i > -\beta x_i)$$

$$= 1 - F(-\beta' x_i)$$

where P (.) is a probability distribution function. In the above expression F(·) represents the distribution function belonging to u_i. If we assume the u_i's to be independent identically distributed (henceforth i.i.d.) random variables with the *logistic* distribution function given by

$$F(z) = \frac{\exp(z)}{1 + \exp(z)},$$

then

$$F(-\beta' x_i) = \frac{\exp(-\beta' x_i)}{1 + \exp(-\beta' x_i)} = \frac{1}{1 + \exp(\beta' x_i)}.$$

This model is called the *logit model.* If we assume the u_i's to be i.i.d. random variables with a $N(0, \sigma^2)$ distribution, we obtain the probit model. If F(·) represents the distribution function of the u_i's, then

$$F(-\beta' x_i) = \Phi\left(\frac{-\beta' x_i}{2\sigma^2}\right),$$

where Φ(·) is the distribution function of the standard normal distribution given by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{x^2}{2\sigma^2}\right) dt.$$

Suppose that we want to estimate the parameter β in either the logit or the probit model by means of ordinary least squares (OLS). Then it appears that we have to deal with *heteroscedasticity.* Since u_i either equals $-\beta'$ x_i or 1 - $\beta' x_i$ with probabilities $F(-\beta' x_i)$ and 1 - $F(-\beta' x_i)$ respectively, the conditional variable of u_i, given x_i, is given by

$$\text{var}\,(u_i \mid x_i) = \beta' x_i (1 - \beta' x_i).$$

The above expression for var (u_i | x_i) indicates the presence of heteroscedasticity. In practice, the usual estimator, the OLS estimator, may be *inconsistent.* Thus we have to rely on another estimator, *e.g.*, the *maximum likelihood estimator* of β.

The multidimensional probit model is obtained by assuming that the u_i's are i.i.d. random variables, having a multivariate normal distribution with mean 0 and positive definite covariance matrix Σ.

Both models have extensively been used in quantitative social science research, mainly as explanatory tools in a behaviourial context. But their categorical input makes them also suitable as classification methods, with even a clear predictive capability.

7.2 Conclusion

In this section we have discussed the logit and probit model as typical examples of discrete regression models. In case of the logit and probit model, the dependent variable should be binary valued, whereas the independent variables are allowed to be real-valued. Both the logit and probit model can be used in binary classification problems. The observations $y_1, \ldots, y_n$ and regressors $x_1, \ldots, x_n$ can be used to estimate the parameters β. Given a new observation x_{n+1}, the value of the corresponding y_{n+1} can be predicted. In this context, these models are also useful for classification purposes.

8 Two-way Tables

There are also various multidimensional analytical tools. In this section we will consider two-way tables. This classification technique differs substantially from the other methods in this paper, since it does not yield a classification. The starting point of the two-way tables is an existing classification. However, the method may provide some information about the underlying statistical pattern of the attributes of phenomena.

8.1 A description

In this subsection we will offer a simplified description of two-way tables. Consider a collection of n objects that can be classified according to two attributes a_1 and a_2. Attributes a_1 and a_2 have k and r values, respectively. The classification can be summarized in a matrix X_n, called a k x r contingency table. In this matrix the entry $X_{n\ ij}$ denotes the number of objects having attribute a_1 with value i and attribute a_2 with value j. The i-th row of the matrix X_n is denoted by $X_{n,i}$ and its j-th column is denoted by $X_{n,j}$. Let α_i denote the probability that the first attribute has a value i and β_j the probability that the second attribute has a valuej. Let P_{ij} denote the probability that attribute a_1 assumes value i and that attribute a_2 assumes valuej. It is now interesting to know whether the two categories are *independent*, i.e. whether

$$p_{ij} = \alpha_i \beta_j .$$

Thus, we want to test the null *hypothesis of independence*

$$H_0: p_{ij} = \alpha_i \beta_j .$$

Since we do not know the probabilities α_i and β_j we can estimate them by $\hat{\alpha}_i = X_{n,i\cdot}/n$ and $\hat{\beta}_j = X_{n\cdot,j} / n$. The null hypothesis is rejected for large values of

$$D_n^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(X_n - n\hat{\alpha}_i\hat{\beta}_j)^2}{n\hat{\alpha}_i\hat{\beta}_j}.$$

From statistics it is known that D_n^2 has a $\chi^2_{(k-1)(r-1)}$ -distribution, i.e. a chi-square distribution with $(k-1)(r-1)$ degrees of freedom. Therefore, the null hypothesis is rejected if

$$D_n^2 > \chi^2_{(k-1)(r-1),0.95} \quad ,$$

where $\chi^2_{(k-1)(r-1),0.95}$ denotes the 95%-quantile of the chi-square distribution with $(k-1)(r-1)$ degrees of freedom. Such a two-way (or, in general, multi-way) table analysis -often coined contingency table analysis- is also often used in the context of log-linear statistical analysis.

8.2 Conclusion

In this section we have discussed some basic principles of two-way tables. Although this method does not yield a classification, it may provide some interesting information about independencies in an existing classification. The method can reject or accept the hypothesis of independence. In this first case, the method makes clear that two attributes are not related to each other, while the acceptance of the null hypothesis indicates a dependence between the attributes. This technique is essentially an exploratory method in the framework of comparative analysis.

9 The Poisson-regression Model

Finally, the Poisson-regression model will be discussed. This model is typically appropriate to model *count data,* i.e., the number of events in a certain period. Suppose the variable Y represents the number of events in a specific period, where the number of events depends on a vector of regressors denoted by x. The relation between Y and x could be modelled by a direct linear or non-linear regression. However, the log-linear model yields much better results in case the dependent variables are discrete in nature and the values of Y tend to be close to zero. For example, the Poisson-regression model may be used to model the number of car accidents at a specific location for each hour; see, for example, [7]. The vector of regressors might then contain e.g. the average velocity of the passing cars and the number of passed cars. This is a rather well established method.

9.1 A description

Let Y be a stochastic variable representing the number of events in a certain period. Suppose Y is drawn from a Poisson distribution with parameter $\lambda > 0$ related to a set of independent variables x. The primary equation of the model is

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

Moreover, it is assumed that λ is defined by

$$\log \lambda = \beta' x.$$

The above relation for λ is called the log-linear model. It follows that the expected number of events per period is now given by

$$E\,[Y|x] = \mathrm{var}[Y|x] = \lambda = \exp \beta'\lambda.$$

Given a sample $y_1,\ldots,y_n$ and regressors $x_1,\ldots,x_n$, the coefficient β can be estimated by means of the maximum likelihood estimator. Clearly, this method is also suitable for predictive purposes on the basis of comparative data analysis.

9.2 Conclusion

In this section the Poisson-regression model has been discussed. This model is typically appropriate for the modelling of count data, where the number of events is likely to depend on a vector of regressors. In general, this model yields better results than a classical linear or non-linear regression model, especially if the count data take small values and the regressors are discrete in nature. It is a valuable method for quantitative comparative methods.

10 A Comparison of Methods

In the previous sections we have concisely reviewed a wide variety of classification methods. Despite commonalities they also appeared to have specific features. In this section we will now compare -by way of meta-experiment- the classification techniques considered in this paper. For a concise summary, the reader is referred to Table 1.

Table 1. A comparison of different classification techniques

	RST	FST	Cluster analysis	Discriminant analysis	Logit/Probit Model	Neural networks	*k* x *r* tables	Poisson regression
Stochastic data	no	no	no	yes	Yes	no	yes	yes
Relative importance of attributes	yes	no	no	yes	yes	no	-	yes
Incomplete and imprecise data	yes	yes	no	no	no	yes	no	**no**
Appropriate for large data sets	**no**	**no**	yes	yes	yes	yes	yes	yes
Appropriate for large small sets	yes	yes	yes	no	no	yes	no	no
Prediction (p) or classification (c)	c	c	c	p	p	p		p
Classification of unseen objects	yes	yes	no	yes	yes	yes	-	yes

10.1 A comparison of classification techniques

From a first inspection of the various classification methods, it should be observed that discriminant analysis, the logit (probit) model and the Poisson-regression model are based on a stochastic interpretation of the data concerned. In case of discriminant analyis, such an assumption is necessary to compute the probabilities mentioned in Section 5. Often a normal distribution is assumed, since this assumption simplifies computations. In the logit (probit) model, usually logistic (normal) errors are assumed. The other techniques do not make any assumptions about the underlying distribution of the data.

In the discussion of the classification techniques we have seen that it is sometimes necessary to transform qualitative variables into quantitative ones. For cluster analysis and RST, such a transformation does not influence the results of the classification. However, such a transformation does influence the results of discriminant analysis. In this case discrete discriminant analysis yields better results. For RST it is also required that real-valued attributes are transformed into discrete value variables by means of some way of categorizing. This kind of categorization strongly influences the results of RST. Furthermore, logit and probit models can only deal with dependent variables that are binary valued. It has been made clear that this is not a very strong restriction, since the class of binary valued variables incorporates all ‘yes or no’ variables. In case of the Poisson-regression model, the dependent variables are assumed to be discrete valued, whereas the regressors are allowed to be real-valued.

The neural network approach does not aim to identify classes or clusters in the data; it can do something stronger. It can predict the values of the dependent variable. The same holds for the logit and probit model and the Poisson-regression model.

Cluster and discriminant analysis have in common that they split up the data set in different classes. For discriminant analysis the different classes have to be known a priori, whereas cluster analysis aims to identify these classes.

RST, FST and discriminant analysis, neural networks, logit and probit models and the Poisson-regression model can classify a new object on the basis of known information on the attributes. This cannot be done by means of cluster analysis.

RST, FST as well as neural networks have the property that they can deal with incomplete, imprecise and missing data. This is of course an important advantage in social science research.

FST is somewhat different from the other classification techniques, since it involves fuzzy classes and yields a fuzzy classification, that is, a classification with vague or imprecise classes. It is possible that an object belongs to different classes with various degrees of membership.

The two-way tables have the property that they can tell something about a possible dependence between two attributes. However, this technique does not find a classification itself, it merely says something about an existing classification.

RST yields rules that reflect patterns observed in the data. These rules can be either exact or inexact, where inexact rules refer to inconsistencies in the data. Rules can be very useful for relatively small data sets. For large data sets, a large quantity of rules will likely be generated. In that case, the rules will be difficult to interpret and will therefore not contribute to a better understanding of regularities in the data.

Neural networks can deal very well with large data sets and yield results in a relatively short computation time, whereas cluster analysis and discriminant analysis usually take more time. Nevertheless, with modern computers most computations can be done in a very short time span.

Finally, RST, discriminant analysis, the logit (probit) model and the Poisson-regression model can indicate the relative importance of each attribute. RST even yields additional information in the form of reducts and a core. A reduct is a group of variables that can fully explain the dependent variables. The core consists of those variables that as a minimum requirement must be taken into account in explaining completely the dependent variables.

10.2 Conclusion

On the basis of various qualitative judgement criteria, we have compared the classification techniques that were introduced in the previous sections. We have seen that there are various important differences between the techniques employed. This will have serious implications for the choice of classification techniques in an actual research situation, as will be shown in the next section.

11 The Choice of Classification Techniques

After the review of a multiplicity of classification methods, we will in this section discuss the question how to find the most appropriate classification technique in a specific situation. The key factors in this decision process are, of course, the aim of the classification and the nature of the data set. In the sequel this will be dealt with more precisely.

11.1 Which classification technique to use

As mentioned before, the choice of classification technique depends on the kind of classification that is needed. For example, if one wants to know the relative importance of each attribute, then RST, discriminant analysis, the logit (probit) model or the Poisson-regression model is likely to be a good choice.

The size of the data set should also be taken into account when choosing an appropriate classification technique. Although RST has some very special features, it is not very appropriate for large data sets, since it gives the information about the data set in a rather unstructured form, namely in the form of rules. If a small data set of 10 objects yields 5 rules, these can easily be interpreted. However, a large data set of 1000 objects yielding 400 rules will cause a problem, since it is not clear what to do with 400 rules. Conversely, discriminant analysis only works well for large data sets. This has to do with the stochastic assumptions about the underlying distribution of the data. The same holds for the logit and probit model. In these two models the maximum likelihood estimator of β has to be computed. This estimator has some nice asymptotic properties that are only achieved for large samples.

If the aim of the classification is an exploration of the structure in the data, without any a priori knowledge of this structure, then cluster analysis is a good option. This technique is both appropriate for small and large data sets, although the computation time increases with the number of objects. Since cluster analysis gives an indication about the structures in the data, it can serve as a preparatory step preceding e.g. a discriminant analysis.

Neural networks are especially appropriate for large data sets, with possibly missing or incomplete values. Since neural networks work relatively fast, they can effectively handle large data sets. Experiments also point out that neural networks yield relatively good results for extremely small data sets. Usually, discriminant analysis fails in such situations (see e.g. [8]).

Neural networks and RST are techniques that have their origins in artificial intelligence and are especially appropriate for incomplete and imprecise data, like many other techniques from that part of cognitive science. FST is an appropriate tool in case a fuzzy classification is required. FST, also originating from artificial intelligence, can effectively deal with incomplete and imprecise data. If the dependent variables are binary, the logit (probit) model might be appropriate. Of course, assumptions about the underlying distributions of the data should be tested statistically. If these tests fail, other methods should be tried. Such alternative methods, e.g. RST and neural networks, can also deal with binary classifications. The Poisson-regression model resembles the logit and probit model, but can handle a non-binary dependent variable. However, the Poisson-regression model is typically appropriate for count data, representing the occurrence of certain events in time. If something has to be said about the dependence between different attributes in an existing classification, then two-way tables might be useful.

11.2 Conclusion

Based on the previous considerations, it seems wise to try -whenever possible- different classification techniques and to compare the results, while taking into account the structure of the data and the aim of the classification. A sensitivity analysis might be useful in this context as well.

13 Retrospect

The number of classification methods for comparative research is vast. In this paper we have selectively considered a collection of classification techniques: RST, FST, cluster analysis, discriminant analysis, logit and probit models, neural networks, two-way tables and the Poisson-regression model. These classification techniques have in common that they can be used in quantitative comparative research.

A comparison of these classification techniques shows that there are various fundamental differences between the techniques. Therefore, one has to choose carefully a certain classification technique, taking into account the structure of the data, the size of the data set and the aim of the classification.

In this paper we have discussed the properties of various classification techniques and the differences between those techniques. As mentioned in the introduction, classification techniques may be very useful in comparative research. This holds for the broad field of social science research, but certainly also for those disciplines which have an explicit spatial connotation, such as geography, regional science and transportation science. This paper aims to stimulate further research in the application of classification techniques in comparative research in all these fields.

Literature

[1] Bergh, J.C.J.M. van den, K. Button, P. Nijkamp and G. Pepping, *Meta-Analysis in Environmental Economics,* Kluwer, Dordrecht, 1997

[2] Cramer, J.S., *The Logit Model,* Edward Arnold, London, 1991

[3] Fischer, M., and S. Gopal, Neurocomputing, *Environment & Planning A,* vol. 25, 1993, pp. 757-760

[4] Gallant, *S.I., Neural Network Learning and Expert Systems,* MIT Press, Cambridge, Mass., 1995

[5] Goldstein, M., and W.R. Dillon, *Discrete Discriminant Analysis,* John Wiley, New York, 1978

[6] Gordon, A.D., *Classification,* Chapman-Hall, New York, 198 1

[7] Greene, W.H., *Econometric Analysis,* Prentice-Hall, Englewood Cliffs, 1997

[8] Hand, D.J., *Discrimination and Classification,* John Wiley, New York, 198 1

[9] Himanen, V., P. Nijkamp and A. Reggiani (eds.), *Neural Networks in Transport Applications,* Avebury, Aldershot, UK, 1998

[10] Munda, *G., Fuzzy Set Theory for Environmental Evaluation,* Physika Verlag, Heidelberg, 1995.

[1 1] Nijkamp, P., and Pepping, G., A Meta-Analytic Exploration of the Effectiveness of Pesticide Price Policies in Agriculture, *Journal of Environmental Systems, vol. 26,* no. 1, 1998, pp. l-25

[12] Orlowska, E., *Incomplete Information: Rough Set Analysis,* Physika-Verlag, Heidelberg, 1998

[13] Pawlak, Z., *Rough Sets, Theoretical Aspects of Reasoning about Data,* Kluwer, Dordrecht, 199 1

[14] Slowinski, R., *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory,* Kluwer, Dordrecht, 1991

[15] Zadeh, L.A., *Fuzzy Sets and Applications,* John Wiley, New York, 1987