cepe

Centre for Energy Policy and Economics
Swiss Federal Institutes of Technology

Finding Groups in Large Data Sets

Adrian Müller

CEPE Working Paper Nr. 18
October 2002

CEPE
ETH Zentrum, WEC
CH-8092 Zürich
www.cepe.ethz.ch

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

PAUL SCHERRER INSTITUT PSI

Finding Groups in Large Data Sets

Adrian Müller
Centre for Energy Policy and Economics CEPE
ETH Zürich

October 2002

Abstract

This paper aims to give an overview of methods to find groups in large data sets, such as household expenditure survey data. These methods are grouped in three: cluster analysis, dimension reduction and basic explorative methods. The emphasis is put on a critical analysis and potential drawbacks, especially of inputs that have to be provided by the researcher. These may impose some structure not present in the data, thus defeating the purpose of revealing intrinsic patterns.
In general, the more elaborate methods, such as cluster analysis, are delicate to apply, especially in the context of social sciences. Often, it may be best to limit oneself to more transparent approaches such as comparisons of basic statistics.

1 Introduction

The goal of this paper is to describe some applications to find groups in large data sets, such as household expenditure survey data. Based on the respective increase or decrease in the size of the identified groups of households and on changes in their consumption patterns over the course of time, these groups can be used to describe some aspects of the development of a whole country. These are the reasons to identify such groups in the course of our research that deals with household expenditure data from India.

Finding groups in large data sets clearly is an issue in the investigation of other data as well, and thus this paper may be of interest to researchers dealing with data from a broad range of other sources.

Up to now, methods to find groups in data or, more generally, pattern recognition have not yet widely been used in economic science. With increasing ease in computing, the application of these methods to large data sets has become more feasible. These methods can, used properly, help a great deal in revealing some structure. Thus, it may be useful to have some idea of the potentials and limits of these methods and to learn from other disciplines where they have already been applied a long time.

We want to emphasize that all these methods are full of difficulties. The mere presence of these tools in most statistical packages and the seeming simplicity of their application must not tempt one to use them in a thoughtless manner. It is often more useful and transparent to only employ relatively simple and straightforward approaches, such as comparisons of basic statistics like means and variances.

To give an idea of the methods to find groups in data and to advise caution while using them is a second, more general goal of this paper. It addresses practitioners used to apply statistical tools and dealing with large data sets but not being especially trained in mathematical statistics. Thus, a main focus of the following text is the presentation and discussion of some existing methods so as to do justice to the didactical aims mentioned above.

The general idea is to find groups by making use of some potentially present intrinsic structure of the data and to avoid imposing some more arbitrary structure generated by the researcher, such as income decile groups. To check for the presence of intrinsic structure and to reveal it if it is present, it is best to employ cluster analysis or dimension reduction techniques. These will be described in section 2.3 and 2.4. Since these techniques are not without problems and do not necessarily lead to useful results, we suggest as a third approach to group the data by only relying on basic descriptive statistics. This latter approach may be used if the cluster analysis and dimension reduction techniques do not lead to sensible results. Besides its simplicity, it has the advantage that it can be formulated in a very flexible way and does not require the researcher to arbitrarily choose any parameters thus imposing some unjustified restrictions.

It may be honest to state that we take quite a skeptical view with respect to applications of cluster analysis. This may be reflected in the way we question some applications in the examples we have chosen to back up our discussion with. Generally, we think that these methods have great potential if applied cautiously in situations where all the conditions to successfully apply them are met. However, we consider this to be the case less frequently than hoped for or expected. Applied in cases where they are not adequate, we think that these methods bear the danger to lead to unsound results that might look quite promising, however, and thus tend to hinder or avoid a critical assessment. This is especially the case when the methods, included in most statistical software packages and seemingly easy to apply, are used in a "black-box" manner (as an example, see Huttin (2000), where we suspect this to possibly be the case since not any comment on the methods used is included. In consequence, it is virtually impossible for the reader to assess the results.).

The statistical analysis is done using SAS (SAS-Institute, 2001). We have chosen this program since it has already been in use at our institute, and is able to handle very large datasets quite easily. It is a really powerful tool as soon as one has gotten used to it and found the thorny path through the jungle of its documentation.

An alternative choice would be to use the freeware R (R-project, 2001), which we consider equally powerful and for which there exists an excellent on-line documentation.

In the remaining part of this paper, the ideas and methods that can be applied to identify groups in data will be discussed in section 2. Some remarks and conclusions will then be presented in section 3.

2 Ideas and Methods

In this section we will first introduce some general notions. These will be illustrated by examples from the data set we are concerned with in our research, i.e. several annual rounds of the National Sample Survey (NSS) expenditure data on Indian households (NSSO, 1998).

2.1 How to look at the Data

Each round of the NSS data includes values for several hundred variables, v_1 to v_p, giving the entries of a vector $\vec{v}$, ascertained for some ten thousands of households, h_1 to h_n, that can be collected in a vector $\vec{h}$. These data can be organized as a matrix X, the values grouped by variables defining the columns $\vec{v}_j, j = 1...p$, i.e. vectors with n entries, the values grouped by households correspondingly defining the rows $\vec{h}_i, i = 1...n$. The matrix entries $x_{ij} = h_{ij} = v_{ji}$ thus give the value the j-th variable takes for the i-th household.

The simplest way to look at this data is to imagine it as a large cloud of points in a high-dimensional linear space[1] $S \cong \boldsymbol{R}^p$: each variable spans one coordinate axis or dimension, the values $v_{ji} = x_{ij}, j = 1...p$, the variables take for one household $\vec{h}_i$ define its position along these axes and thus its position in the space spanned by all the variables. We give a two-dimensional example: take the values every household reports for the two variables 'monthly per capita expenditure on rice' and 'monthly per capita expenditure on wheat' – this can be depicted as a cloud of points in a plane (see Figure 1).

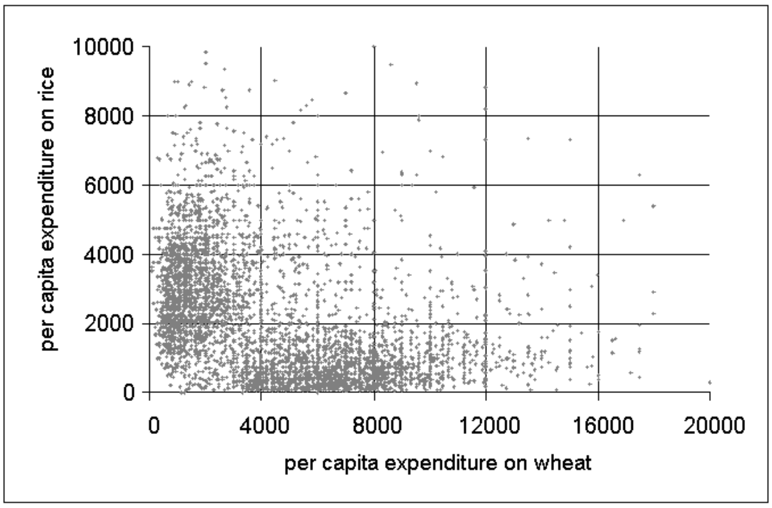

Figure 1:Scatter-plot of monthly per capita expenditure on rice vs. monthly per capita expenditure on wheat (in Rupees/100). Date for the states of Andhra Pradesh and Gujarat, India, NSS Round 50 (1993-1994).

This cloud in a high dimensional space can show a wide range of forms and structures – the idea is to let the data reveal its intrinsic structure to get an unprejudiced view of it. The main hope is to find distinct 'clusters' of households: parts of the cloud showing a high density separated by areas of relatively lower density.

[1] To look at the observations as elements of a linear space imposes already some structure on the data. This is appropriate in most of the cases, but the possibility of a more general approach should be borne in mind. This is the case for binary variables, for example, but these can easily be included in such a formalism.

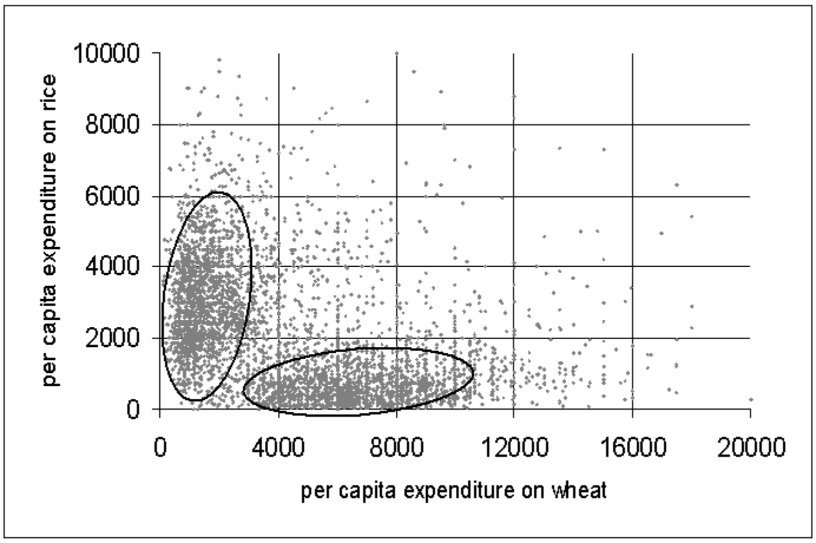

Figure 2: Scatter-plot as in Figure 1, additional ellipses pointing out areas of higher density defining clusters.

Take the example from above: in the NSS data we have found a lot of households showing higher values for expenditure on rice and lower ones on wheat - or vice versa. But there are not that many households with more or less balanced expenditures on these two items. Thus, we can see two distinct clusters of higher density in the plane spanned by these two variables (see Figure 2).
Besides finding these clusters by looking for areas of high density, they could also be identified by looking at the data from different directions, i.e. by investigating projections onto subspaces of lower dimension (see Figure 3).

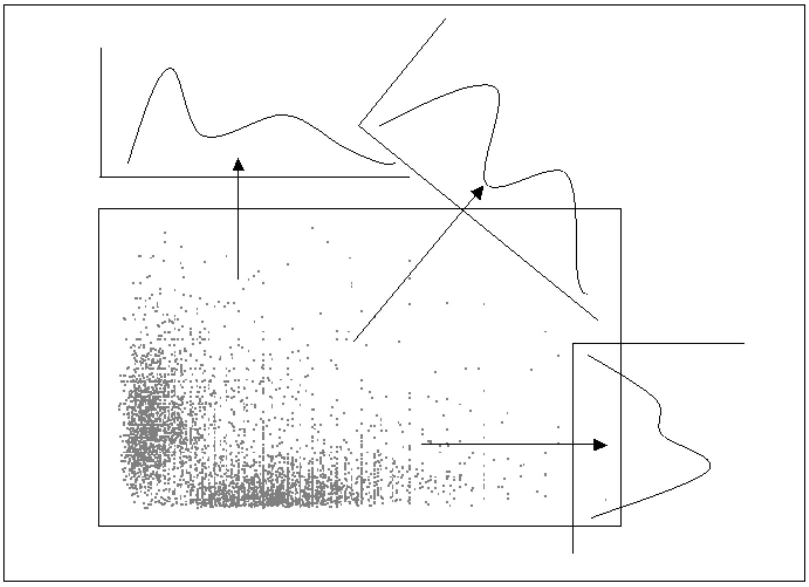

Figure 3: Scatter-plot as in Figure 1, frequency curves visualizing different projections on one-dimensional sub-spaces.

These two examples may serve to explain the basic features of the two groups of statistical methods presented below.

If, however, we cannot see any structure at all (Figure 4), or this structure is so intricate that the statistical methods are not able to reveal it (Figure 5), or the only structure present is due to some discrete variables (Figure 6) and thus trivial, we may be forced to restrict ourselves to more basic methods such as calculating descriptive statistics or comparing means by use of analysis of variance, since applying the other methods would result in groupings that cannot be backed up by any structure intrinsically present in the data and thus would mislead further research.

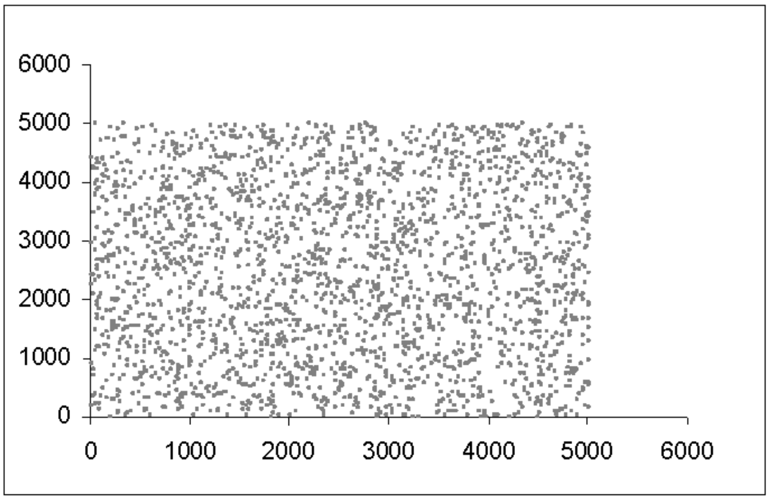

Figure 4: Two-dimensional uniformly distributed random data, generated by the author.

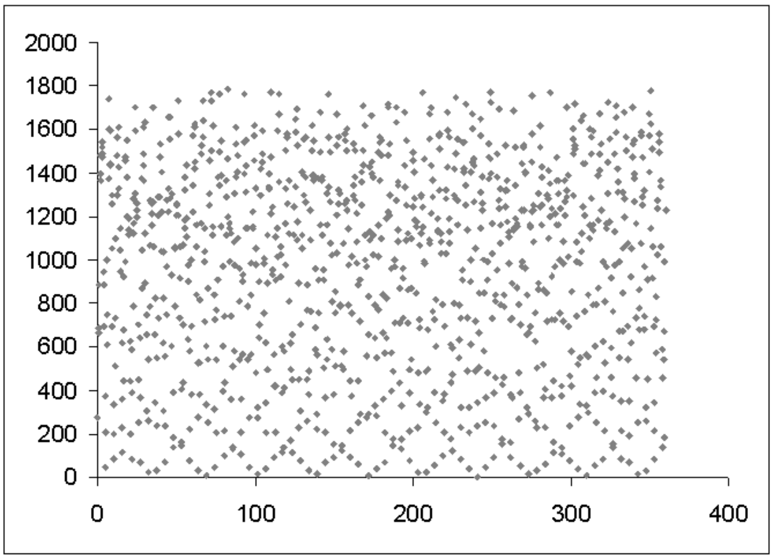

Figure 5: Data with some structure, generated by the author.

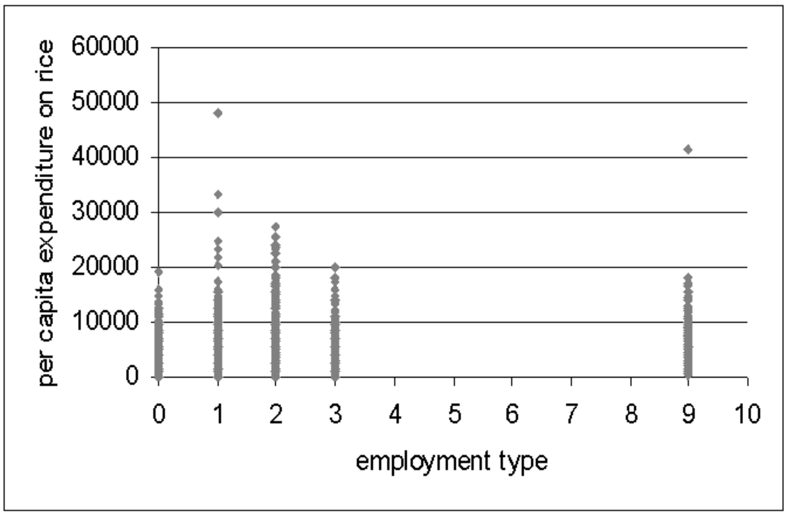

Figure 6: Scatter-plot of employment type of the head of the household vs. monthly per capita expenditure on rice (in Rupees/100). Date for the states of Andhra Pradesh and Gujarat, India, NSS Round 50 (1993-1994).

2.2 Introduction to the Statistical Methods

The best method to reveal structure in the data would still be simply to look at it, since the human brain is better than any computer at detecting patterns. However, this is only possible in two and three dimensions. With more dimensions, it is well worth looking at scatter-plots of all the different two-dimensional combinations of the variables to get a first idea of the data, but to systematically look for patterns in higher dimensions, we have to make use of some statistical tools. To help revealing such structures, there are several statistical methods, which we will group into cluster analysis methods, dimension reduction techniques and some more basic methods, such as analysis of variance (ANOVA). Cluster analysis directly tries to find distinct groups in the data. Dimension reduction techniques try to find projections of the data cloud onto lower-dimensional sub-spaces containing maximal information on patterns potentially present in the data. The third group of methods is based on group-wise comparison of the basic statistics for several variables and suitable visualizations and descriptions thereof.

For our purposes, i.e. to identify useful groups, it is enough to only use explorative statistical methods without any inferential framework. Thus, strictly speaking, we make statements on the sample only and cannot make any statistically sound inferences or tests on the presence of the groups for the population as a whole. For cluster analysis, such an inferential framework is generally not present anyway, and the techniques do not rely on any assumptions with regard to underlying models (Everitt, 1993). We will also present the dimension reduction techniques without reference to inferential methods. As far as the basic descriptive methods are concerned, however, it is worth dealing with the existing inferential methods and models, for example to be able to assess the significance of different means for different groups.

2.2.1 Some decisions that have to be taken before starting with an analysis

Before we can start with any analysis, we have to decide which variables from the data set to include and if they need some preparation to be used.

The choice of the variables necessarily has a big influence on the possible outcomes and it is of great importance to do it in the best possible way. Often the data will not give much information which variables might be crucial to be included in an analysis. In this case, we are forced to define the key variables using personal preference and experience and information from other studies[2]. The reasons which variables got chosen should be laid open to discussion and be well-founded. In addition, a correlation analysis may help to avoid including sets of highly correlated variables that could lower the performance of some methods.

The preparation of the variables could involve a normalisation of the values if continuous variables measured in different units are present, or defining a set of k binary variables to represent a categorical variable that can take k unordered levels, or to translate such a variable into an ordinal one by means of some criteria.

An example from the NSS data is the variable 'primary source of energy for lighting', which can take the values 1 (kerosene), 2 (other oil), 3 (gas), 4 (candle), 5 (electricity), 8 (others), 9 (no lighting arrangement). This categorical variable can be translated into 7 binary variables. On the other hand, the researcher could impose some ranking: 'electricity > gas > kerosene or oil' (with weights, e.g. 'electricity = 1, gas= 0.25, kerosene and oil=0.0') to code them in one variable, and define a second variable including all the other types, since they are not frequently reported anyway. Such a ranking and weighting does not reflect any intrinsic properties of the data or this variable alone. It has to be seen in combination with some opinion and judgement of the researcher (take for example the case that she ranks the energy sources with respect to their correlation with some notion of poverty). This does not necessarily include all the relevant issues (as an example: depending on the geographic area, the access to electricity may depend more on the presence of a grid at a certain location and less on how wealthy a certain household is).

It also has to be decided, if some observations have to be excluded from the analysis. This could be necessary due to missing values or because they are outliers. Before deleting observations, however, it should have been verified that this does not result in any bias because of affecting some groups of observations more than others.

Further on, a decision has to be taken with regard to the relative importance of the variables retained, i.e. weights for the variables have to be chosen explicitly. One has to be aware of the fact that not to address this issue explicitly results in an implicit choice of equal weights for all the variables, which by no means need be the best solution.

Finally, a measure of distance or similarity between different observations has to be chosen. In the case of a linear space, this defines the metric in the space spanned by all the variables and wherein all the observations will be represented. Choosing the

[2] Another possibility to find the relevant variables is to employ a regression analysis. Choosing a variable of interest as the dependent one, the significance of several other variables in explaining it can be estimated. The variable of interest and the functional form of the model have to be chosen by the researcher, but the data itself will then point out the significant variables. Thus, this method may let the data reveal its structure without too much input from the researcher. However, it should not be forgotten that a regression analysis is based on quite a lot of assumptions. Nevertheless, it can guide us in selecting a set of relevant variables. An example is given in (Pachauri 2001), with 'per capita energy requirement' as the dependent variable.

distance measure also includes a weighting of the variables (take as an example the Euclidean distance: $d(\vec{h}_r,\vec{h}_s)=\sqrt{\sum_{i=1}^{p} w_i(h_{ri}-h_{si})^2}$, where w_i equals 1 in the ordinary case, but can be chosen to be different from 1 to give different weights to the different variables), but we have mentioned this latter issue separately in the preceding paragraph to point out its importance.

In general, a distance measure d_{rs} is a function $S\times S\rightarrow[0,\infty)$: $(\vec{h}_r,\vec{h}_s)\mapsto d(\vec{h}_r,\vec{h}_s)\equiv d_{rs}$ subject to the following conditions: $d_{rr}=0$ and $d_{rs}=d_{sr}$.

There are many different possibilities of distance measures; examples are the Euclidean distance given above, the city-block distance $d(\vec{h}_r,\vec{h}_s)=\sum_{i=1}^{p} w_i\left|h_{ri}-h_{si}\right|$ or the matching coefficient for p binary variables, each with values 0 and 1: $d(\vec{h}_r,\vec{h}_s)=(a+d)/p$, where a is the number of 1-1-correspondences between $\vec{h}_r$ and $\vec{h}_s$ and d the respective number of 0-0-correspondences. These are examples for the inclusion of variables of only one type, numeric or character, at once. But often both types are present and should be included in the analysis simultaneously, which can be done by combining the respective distance measures.

After having processed the variables in this way, the data set has again the same general form as above (cf. section 2.1), but possibly a different number of variables and observations.

These issues are a main source of subjectivity in the analysis, since the choice of the weights and distance measure, and also the decision on the importance of outliers[3], is left to the researchers discretion and often cannot be justified by sound arguments derived from properties of the data alone. We will discuss some problems related to that in section 2.3.3 and 2.4.

2.3 Cluster Analysis

Based on different notions of distance between data points and the additional input of the clustering criterion (which says what properties define a good cluster), the different clustering procedures try to find clusters in the data cloud (cf. Figure 2). The success of these methods strongly depends on the form of the clusters, if there are any (most methods are biased towards spherical or elliptic clusters), and on the suitability of the distance measure and clustering criterion chosen.

In the following, we will shortly discuss some clustering methods more formally. They can be grouped into hierarchical and non-hierarchical methods. For more information and further references on these and other methods, we refer the reader to the literature (Everitt, 1993; Webb, 1999; SAS-Institute, 2000; Falk, Marohn et al., 2002).

2.3.1 Hierarchical Methods

The basic idea is either to start with each observation defining a cluster and subsequently joining the observations next to each other to form new clusters and then to join the closest clusters to build bigger ones, until one arrives at one cluster, containing every observation (agglomerative approach).

[3] If outliers are excluded from the analysis, it has to be made sure that they do not contain any important information, which is not always an easy task.

Alternatively, one can proceed the other way round by starting with one cluster containing all observations and subsequently splitting it into sub-clusters as different as possible from each other, until one arrives at the situation, where each observation defines one cluster (divisive approach).

A useful visualization of the output of the hierarchical methods is given by a tree graph, which indicates on which level (for example measured by the average distance between clusters) the clusters are joined or divided, respectively (cf. Figure 7).

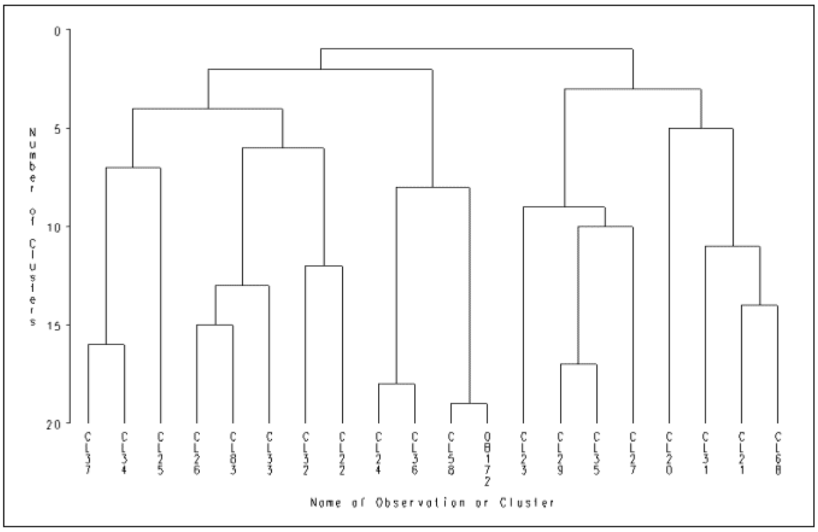

Figure 7: Tree diagram of the first 20 clusters built from households in Gujarat and Andhra Pradesh by means of the Ward method, using the variables 'per capita expenditure on rice' and 'per capita expenditure on wheat'. The first division roughly separates the states, i.e. households preliminarily consuming rice and wheat, respectively.

The crucial input for both these approaches are the chosen measure of dissimilarity or distance between clusters and the criterion to be maximised to find the partition in each step. Both approaches do not provide the researcher with any information on how many clusters between 1 and n may comprise a good solution. However, there are some methods that can help with this decision (see section 2.3.3).

As examples, we mention the agglomerative single linkage and the Ward algorithm. Like most of the other algorithms, these give the clustering criterion in form of a distance measure between clusters and the prescription to join the two closest clusters in each step.

The single linkage algorithm uses the following definition for the distance between two clusters K and L: $D_{KL}^{single} = \min_{i \in C_K} \min_{j \in C_L} d(\vec{h}_i, \vec{h}_j)$, where $C_K \subset \{1,...,n\}$ denotes the indices of the observations in cluster K. Thus, the distance between two clusters is defined as the minimal distance between two members, one out of each cluster. This method has some desirable theoretical properties but is biased for elongate and irregular clusters and scores worse in presence of more compact ones. The Ward algorithm exploits the following distance measure for inter-cluster distance: $D_{KL}^{Ward} = \left\| \vec{H}_K - \vec{H}_L \right\|^2 \Big/ (N_K^{-1} + N_L^{-1})$, where $\vec{H}_K$ is the mean vector of cluster K, and

N_K is the number of observations in this cluster. Thus, for each step, the Ward method minimizes the within-cluster sum of squares over all partitions obtainable by merging two clusters, i.e. it merges these two clusters, whose fusion increases the within cluster variance the least. The method tends to produce clusters of roughly equal size and is very sensitive to outliers.

2.3.2 Non-hierarchical Methods

Non-hierarchical methods use quite a different approach. Most of them start with a given number c of observations, which are used as 'cluster-seeds', chosen arbitrarily or according to some algorithm. According to the distance measure and the clustering criterion, all the other observations are assigned to such a seed and thus c clusters emerge. Then the group mean or some other measure for the 'center' of each cluster is calculated and taken as a new seed. The observations are reassigned to clusters with respect to these new seeds. These steps are repeated until there are no further changes in the assignments anymore.

An example is the k-means method, which uses Euclidean distance and thus assigns the observations to the centers and defines the new cluster center by means of least squares estimation. Each iteration then reduces the least squares criterion until a minimal value is achieved. An example of a method to find the seeds is the so-called Leader-algorithm: an observation is chosen by chance as the first seed. The nearest observation that is further away than a given distance T comprises the second seed. This procedure continues until c seeds are found.

2.3.3 Further Remarks and Potential Problems with Cluster Analysis

Generally, the literature urges the researcher to be circumspect with the use of clustering methods and states that these tend to find groupings in the data even when there is no intrinsic grouping present (e.g. Everitt (1993), Webb (1999). For formal approaches to generally assess the reliability and goodness of cluster solutions we refer to this literature as well. A basic approach is to check a solution by using different methods for the clustering and maybe even different measures of distance and different weights and choices of the variables, or by randomly choosing a subset of the observations and checking if the same clustering emerges and is stable with respect to changes in the parameters involved.

Of special importance is the problem, which choice for the number of clusters represents an appropriate solution. In both hierarchical and non-hierarchical approaches this number has to be chosen somehow, either to define a 'stopping rule' in the hierarchy, giving the best level of clustering, or to define the number of clusters to be looked for right at the beginning by choosing the number of cluster seeds. There are some formal methods to deal with this problem, but none of these give good results in every case. An example is the so-called 'elbow-criterion', which formalizes the prescription to take as the number of clusters this value c, where to go from c to $c+1$ clusters results in a large drop of some notion of diversity for the partition of the data. Information on this and other criteria can be found in the literature cited above.

Depending on the specific method, there are further complications, for example the fact, that many methods are biased for spherical or elliptical clusters, i.e. they can only perform in a satisfying manner if possibly present clusters are of this form. Such issues have to be paid attention to while choosing the clustering method to be used.

2.3.4 Examples of Applied Clustering

We will shortly present four examples of applied clustering illustrating different aspects of these techniques. Further examples can be found in the literature.

The first, Anderhalden (2001), deals with the grouping of 620 municipalities in the southern Swiss alps. Based on approximately 20 socio-economic and some infrastructure variables 10 types of municipalities have been identified (main distinguishing features: 1. based on tourism, 2. high number of commuters, 3. work centers, 4. agricultural municipalities, 5. residential municipalities, 6. peripheral location, 7. towns, 8. tourist centers, 9. and 10. two outliers). The clustering has been done with the Euclidean distance measure and the Ward algorithm. Two weighting schemes have been employed, one imposing equal weights on all the variables and another with equal weights on groups of variables and thus different weights for the single variables, depending on the number of variables per group. Depending on the interpretability of the outcome, the authors decided to finally use the equal weighting variant. This whole procedure reflects how many decisions have to be taken by the researcher and that an assessment of the 'goodness' of a 'good' clustering may rely on how 'reasonable' it is in the view of the researcher, i.e. on quite a subjective criterion.

The second example, Chaturvedi, Green et al. (2001), deals with the k-modes method, which is an adaptation of the k-means method that can be used with categorical variables, applied to data on usage and accessibility of personal computers to a sample of 2000 households in the USA. Based on 8 categorical variables, the algorithm revealed a best solution of three clusters: 1. 'PC-novices', 2. 'Use and like PC', 3. 'Use PC only at work'. For a comparison, the clustering has been done with some other methods as well, and it is found that the k-modes method performs best.
An inspection of the clusters suggests that, within some high correspondence, these clusters could also have been defined directly by utilisation of the two binary variables 'Use PC at home' (or 'Own a PC') and 'Use PC at work' and the resulting four combinations of their values.

The third example is taken from computational chemistry. Bravi, Gancia et al. (1997) use cluster analysis to investigate the similarity of the different conformations of two peptides, i.e. different spatial configurations of the constituting amino acids, and thus to sensibly group the vast amount of possible conformations. The angles between different parts of the peptides and interatomic distances comprise the basic variables for the clustering. The similarity between two objects is given by the requirement that the difference between the respective values for each variable lie within some chosen thresholds. The clustering criterion is defined on the basis of the expectation that good clusters are characterised by similar energy levels for all members (a further example of such an approach is Hamprecht, Peter et al. (2001)) or by a similar spatial disposition of several important chemical functionalities. The paper describes clustering software adapted to this situation and compares the results for different choices of variables and thresholds. Furthermore, the paper critically discusses the limits of this procedure and of cluster analysis in general, which makes it an exemplary report of the application of these methods.

The last example, Brown and Glennon (2000), is taken from economics. Roughly 600 commercial banks are grouped by means of cluster analysis (without reference to the exact method) using the respective percentages of assets in different activities as variables. The resulting groups, however, are not so clear-cut and only mean values for all the variables are given for each cluster without reporting some measure of spread. Thus, it is difficult to assess the results and it might have been better just to group by cutting the data along some threshold values.

2.4 Dimension Reduction Techniques

2.4.1 Methods

In this section, we describe the methods of a second field, which could be summarised under ‘dimension reduction techniques’. The aim is to ‘look at the data from several directions’ to find ‘interesting projections’, i.e. to find lower dimensional spaces that reveal some clustering or otherwise interesting structure (cf. Figure 3). To choose the criteria to identify ‘interesting projections’ is the main input in addition to the choice of the variables and the distance measure. In general, the procedure then tries to identify the projection that maximises or minimizes this ‘criterion of interest’.

‘Looking for interesting projections’ can be understood in different ways. Formally, it always involves some transformation of the data cloud or the reference coordinate system. A simple example is given by a rotation (see Figure 8). This rotation can also be understood as the choice of new variables more appropriate to code the data set at hand. In its most simple form, the new variables are a linear combination of the old ones, $\vec{v}' = \mathrm{R}\vec{v}$, where R is the transformation-matrix for the rotation.

There are quite a lot of techniques that serve to reduce the number of dimensions present in the data. We briefly present principal component analysis, multidimensional scaling and factor analysis and refer to the literature for more information on these and other methods (Huber, 1985; Jones and Sibson, 1987; Flury and Riedwyl, 1990; Jackson, 1991; Jambu, 1991; Gower and Hand, 1996; Falk, Marohn et al., 2002)

Principal Component Analysis (PCA): Given the data and a distance measure, PCA takes the direction of maximal variance as the first principal component. The second and further ones are found using the same criterion under the restriction that they have to be orthogonal to the components already identified. Thus the general procedure results in a rotation of the coordinate system. The rationale behind this approach is the fact, that for elongated ellipsoidal data clouds the principal components define a coordinate system along the main axes of this ellipsoid and thus a more natural system for such a structure (cf. Figure 8, projecting along the first principal component reveals two clusters that would remain undiscovered in other projections).

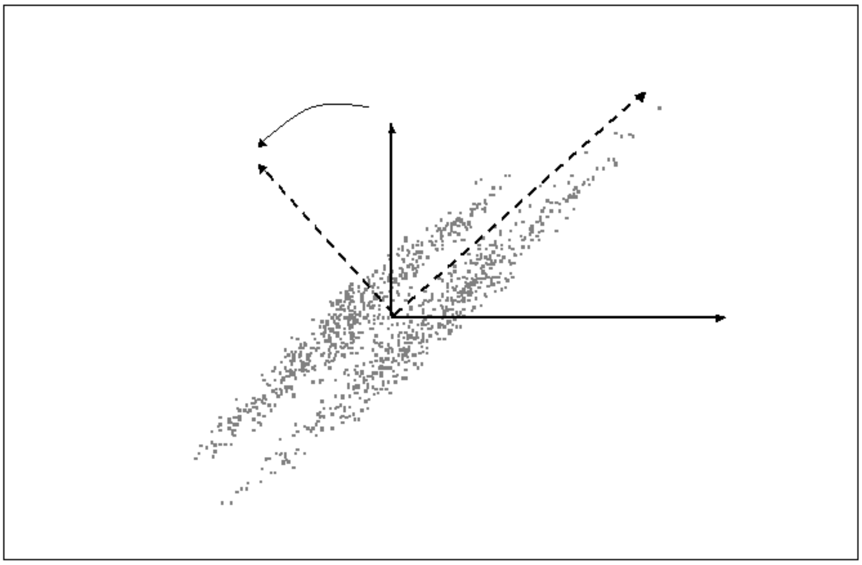

Figure 8: Two sets of transformed normal-distributed random data (generated by the author), rotation of the original coordinate system to define variables more suited to describe the data: the first and the second principal component.

Multidimensional Scaling (MDS): This technique aims to project a high-dimensional data cloud onto a lower dimensional Euclidean space in such a way that all the respective distances between the observations are retained in the lower-dimensional space as well. This is metric MDS. Non-metric MDS only aims at preserving the ranking between the distances.

Factor Analysis (FA): FA is built on the assumption that the values p variables take for n observations, coded in the $n \times p-$matrix X, are better explained by some k other, underlying variables, the 'latent factors' $f_j, j = 1,...,k$.

The general factor-model is given by the following equation: $\vec{h}_i = \overline{h} + \mathrm{L}F_i + \varepsilon, i = 1,...n$, where $\vec{h}_i$ is the i-th observation, $\overline{h}$ the mean vector of all n observations, $F_i \equiv \left(f_1(\vec{h}_i),...,f_k(\vec{h}_i)\right)^T$ give the 'factor scores' for the i-th observation, giving the values this observation takes for the respective factors, L is a $p \times k-$matrix showing the so-called 'factor loadings', the association between the values the original variables take for the i-th observation and its factor scores and ε is an error term. An example of this method will be given in section 2.4.2.

Since for all orthogonal $k \times k-$matrices $\mathrm{A}, \mathrm{A}^{-1} = \mathrm{A}^T$, we have $\mathrm{L}F_i = \mathrm{LAA}^{-1}F_i \equiv \widetilde{\mathrm{L}}\widetilde{F}_i$. Thus one solution of the problem yields an infinite amount of solutions related to the first by a rotation. Thus, to choose a solution, further criteria have to be taken into account. An example is to look for a solution with either factor scores near 1 or near 0, thus giving a direct association between the variables and the factors.

Factor analysis performs best when there are a considerable number of correlated variables present in the data, since only then the basic assumption can be expected to be fulfilled.

2.4.2 Examples of Applied Dimension Reduction

As a first example, we present the paper of Lelli (2001), which contains an application of factor analysis to well-being measurement. The paper works with Sen's functioning and capability formalism. The data have been taken from a panel study on Belgian households and comprises 3800 households and more than 800 variables, of which 54 have been utilized. These can be subdivided into the following groups: "social interactions", "cultural activities", "economic status", ”health", "psychological distress", "working conditions" and "sheltering". The factor analysis with seven factors suggests a similar picture: the first factor has high loadings in variables related to "psychological stress", the second collects some variables related to leisure activities with social interaction, etc. Thus, it may be useful to base further investigation on these seven factors, since they reflect some intrinsic relationship among the observations and provide the researcher with a set of variables suitable for an analysis of well-being based on this data set.

As a second example may serve Agrafiotis, Rassokhin et al. (2001), wherein a library of 90'000 closely related chemicals characterised by 166 binary variables coding the presence or absence of certain structural features at certain places in the molecules is investigated. Three different distance measures and two MDS-algorithms are employed leading to roughly the same four clusters in a two-dimensional projection, thus providing evidence for their intrinsic presence and pointing out the combinations of variables suited best to describe these different groups.

2.4.3 Further Remarks

For all the dimension reduction techniques, the question how many dimensions, or new variables, have to be retained to describe the data appropriately, has to be answered somehow. As with the number of clusters, there are no generally applicable methods yielding good results, but criteria such as the interpretability of the new variables or the total variance explained by a certain number of them, may be applied. For further criteria, refer to the literature.

Compared to the clustering techniques, the dimension reduction methods bear less danger of imposing some structure not present in the data itself. They simply provide the researcher with directions worth looking at the data from. The success of these techniques in revealing interesting structure strongly relies on the choice of the 'criterion of interest' and how well it allows for capturing potentially present structure in the data at hand – and, naturally, if there is some structure at all. Clearly, the distance measure chosen has an influence on how the data looks in the projected sub-space, and thus has some influence on potential patterns.

Dimension reduction is often performed as a preliminary step to the application of some clustering technique. This bears computational and methodological advantages as it may help to identify promising groups of variables in advance.

2.5 Basic Descriptive Methods

In this section, we present some more basic approaches to group the data than the ones described above. If the clustering techniques do not reveal any sensible and reliable groups and if the dimension reduction methods do not reveal any interesting structure, it may be best to collect some simple properties of the data in a way that facilitates further investigations. To look at basic statistics is clearly the most common tool in data analysis. We want to emphasize that it is not only the most basic thing to do but also a valuable means of analysis if more complex approaches fail, and that an analysis based on it need not be of lower quality or meaningfulness than one based on more elaborate methods.

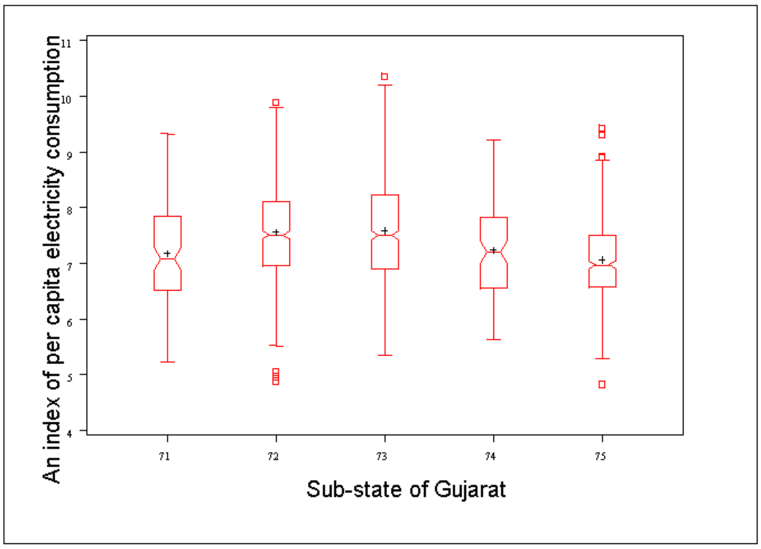

Figure 9: Boxplots for an index of electricity consumption in the five sub-states of Gujarat. These plots contain information on the mean (little cross), median (horizontal line in the box), 1. and 3. quartiles (lower, upper edge of the box), spread (whiskers, indicating the lowest and highest values within the 1. Quartile minus 1.5 times the IQR[4] and the 3. Quartile plus 1.5 times the IQR), extreme values (single data points) and a test for the significance of different medians (if the notches do not overlap they are significantly different at the 0.05 level).

The household data we have at hand contains a lot of categorical variables – thus a trivial grouping can be done with respect to the several values these variables can take. Calculating basic statistics as means, medians, quartiles and standard deviations for continuous variables for each group and drawing boxplots may give a good overview on the data (see Figure 9). Some additional information may be gained by calculating correlations between different variables or between linear combinations of them.

The advantage of groups defined by means of discrete variables is that they are defined in a clear, lucid and reproducible way. Since all the different categorical variables can be utilized, this approach offers quite a flexible and broad way to look at the data. There are also minimal prerequisites concerning some subjective choice of the researcher: she only has to choose the variables. Admittedly, such a grouping may be trivial – which need not be the case for the insight gained by a comparison of means for different groups – but it is more natural than, for example, a grouping according to some quantiles of a continuous variable.

However, to arrive at sound comparisons of means, etc. of different groups, not only on the level of the sample, but for the whole population, some assumptions concerning the probability distributions involved have to be made and critically analyzed. To test for the significance of different means for different groups, etc. at

[4] IQR: Interquartile range: the difference between the 3. and the 1. quartile.

the population level, analysis of variance (Falk, Marohn et al., 2002) and other methods may then be applied.

3 Concluding Remarks

In this paper, we have presented three approaches that could help to find groups in a large data set, e.g. in household expenditure data. We pursued the mainly didactical aim of giving a critical overview of these methods, which are not yet regularly applied in economics. They are available with most statistical packages and, due to ever increasing computational power, readily applicable even to large data sets. Thus, they may be used more and more and it is important to point out their limitations and caution potential users. In many cases, it is not at all straightforward, and might even not be possible at all, to make promising use of them.

The first approach, cluster analysis, designed just for the task to find groups, urges the researcher to take several decisions with respect to the variables to be included and their importance, the distance measure, the number of clusters and the concrete method to be applied. These choices most often cannot be justified by objective reasons or properties of the data and thus the method bears the danger to provoke the researcher to implicitly impose some structure not present at all. This need not be the case but a general advice is to use cluster analysis with great care and to have a critical attitude to its results.

Secondly, the dimension reduction techniques offer some methods to find interesting projections of the data. Thus they can help to find a subset of variables, with respect to which the data may reveal some structure. The main influence of the researcher is given by his choice of the variables to include in the analysis and the distance measure to be utilized. These methods are less prone to generate structure not present in the data, but their performance depends on some basically subjective choices as well.

In case these two approaches do not reveal any structure – be it not present or be it not detectable –, or if the structure revealed does not seem reliable and it is suspected that the choice of the different inputs has strongly influenced or even generated it, we suggest to take a step back and look at basic statistics such as means of some interesting variables only. The values for these for different groups generated by the different levels of categorical variables can then be compared. The groups generated in this manner are trivial, but as this approach can be applied to all the categorical variables and even to quantiles of continuous ones, it offers quite a broad and flexible overview of the data. In addition, the groups thus generated are lucidly defined and always reproducible.

4 References

Agrafiotis, D. K., Rassokhin, D. N., et al. (2001). Multidimensional Scaling and Visualizing of LArge Molecular Similarity Tables. Journal of Computational Chemistry **22**(5): 488-500.

Anderhalden, S. (2001). Gemeindetypisierung des Südalpenraums. Diplomarbeit, ETH Zürich, Schweiz.

Bravi, G., Gancia, E., et al. (1997). SONHICA (Simple Optimized Non-HIerarchical Cluster Analysis): A New Tool for Analysis of Molecular Conformations. Journal of Computational Chemistry **18**(10): 1295.

Brown, J. A. and Glennon, D. C. (2000). Cost structures of bank grouped by strategic conduct. Applied Economics **32**: 1591-1605.

Chaturvedi, A., Green, P. E., et al. (2001). K-modes Clustering. Journal of Classification **18**: 35-55.

Everitt, B. S. (1993). Cluster Analysis 3rd ed., Edward Arnold.

Falk, M., Marohn, F., et al. (2002). Foundations of Statistical Analyses and Applications with SAS. Basel, Birkhäuser.

Flury, B. and Riedwyl, H. (1990). Multivariate Statistics: A practical approach, Chapman & Hall.

Gower, J. V. and Hand, D. J. (1996). Biplots, Chapman & Hall.

Hamprecht, F. A., Peter, C., et al. (2001). A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering, and visualisation. Journal of Chemical Physics **114**(5): 2079.

Huber, P. J. (1985). Projection Pursuit. The Annals of Statistics **13**(2): 435-475.

Huttin, C. (2000). A cluster analysis on income elasticity variations and US pharmaceutical expenditures. Applied Economics **32**: 1241-1247.

Jackson, J. E. (1991). A User's Guide to Principal Components, Wiley.

Jambu, M. (1991). Exploratory and Multivariate Data Analysis, Academic Press.

Jones, M. C. and Sibson, R. (1987). What is Projection Pursuit. J.R. Statist. Soc. A **150**(part 1): 1-36.

Lelli, S. (2001). Factor Analysis vs. Fuzzy Sets theory: Assessing the influence of different techniques on sen's functioning approach. Working paper, Center for Economic Studies, K.U. Leuven.

NSSO (1998). Unit level data from the 50th and other Rounds of Household schedule 1.0 Consumer Expenditure, National Sample Survey Organisation, Department of Statistics, Government of India, New Delhi.: http://mospi.nic.in/nsso.htm.

Pachauri, S. (2001). An econometric analysis of cross sectional variations in total household energy requirements. Annual SAEE research conference on applied energy economics and policy, Zürich, Switzerland, CEPE, Centre for Energy Policy and Economics.

R-project (2001). The R project for statistical computing. http://www.r-project.org/.

SAS-Institute (2000). SAS OnlineDoc, Version 8. http://www.sas.com/service/library/onlinedoc/.

SAS-Institute (2001). SAS - service and support. http://www.sas.com/service/index.html.

Webb, A. (1999). Statistical Pattern Recognition, Arnold.

CEPE Reports

Aebischer, B., Veränderung der Elektrizitätskennzahlen im Dienstleistungssektor in der Stadt Zürich und im Kanton Genf. CEPE Report Nr. 1, Zürich, November 1999.

Filippini, M., Wild, J., Luchsinger, C., Regulierung der Verteilnetzpreise zu Beginn der Marktöffnung; Erfahrungen in Norwegen und Schweden; Studie im Auftrag des Bundesamtes für Energie. CEPE Report Nr. 2, Zürich, 23. Juli 2001.

Aebischer, B., Huser, A., Energiedeklaration von Elektrogeräten; Studie im Auftrag des Bundesamtes für Energie. CEPE Report Nr. 3, Zürich, Januar 2002.

CEPE Working Papers

Scheller, A., Researchers' Use of Indicators. Interim Report of The Indicator Project. CEPE Working Paper Nr. 1, ETHZ, Zurich, September 1999.

Pachauri, Sh., A First Step to Constructing Energy Consumption Indicators for India. Interim Report of The Indicator Project. CEPE Working Paper Nr. 2, Zurich, September 1999.

Goldblatt, D., Northern Consumption: A Critical Review of Issues, Driving Forces, Disciplinary Approaches and Critiques. CEPE Working Paper Nr. 3, Zurich, September 1999.

Aebischer, B., Huser, A., Monatlicher Verbrauch von Heizöl extra-leicht im Dienstleistungssektor. CEPE Working Paper Nr. 4, Zürich, September 2000.

Filippini, M., Wild, J., Regional differences in electricity distribution costs and their consequences for yardstick regulation of access prices. CEPE Working Paper Nr. 5, Zurich, May 2000.

Christen, K., Jakob, M., Jochem, E., Grenzkosten bei forcierten Energiesparmassnahmen in Bereich Wohngebäude - Konzept vom 7.12.00. CEPE Working Paper Nr. 6, Zürich, Dezember 2000.

Luchsinger, C., Wild, J., Lalive, R., Do Wages Rise with Job Seniority? – The Swiss Case. CEPE Working Paper Nr. 7, Zurich, March 2001.

Filippini, M., Wild, J., Kuenzle, M., Scale and cost efficiency in the Swiss electricity distribution industry: evidence from a frontier cost approach. CEPE Working Paper Nr. 8, Zurich, June 2001.

Jakob, M., Primas A., Jochem E.,Erneuerungsverhalten im Bereich Wohngebäude – Auswertung des Umfrage-Pretest. CEPE Working Paper Nr. 9, Zürich, Oktober 2001.

Kumbaroglu, G., Madlener, R., A Description of the Hybrid Bottom-Up CGE Model SCREEN with an Application to Swiss Climate Policy Analysis. CEPE Working Paper No. 10, Zurich, November 2001.

CEPE Reports und ***CEPE Working Papers*** sind teilweise auf der CEPE-Homepage (www.cepe.ethz.ch)
erhältlich oder können bestellt werden bei: CEPE, Sekretariat, ETH Zentrum, WEC, CH-8092 Zürich.

Spreng, D. und Semadeni, M., Energie, Umwelt und die 2000 Watt Gesellschaft. Grundlage zu einem Beitrag an den Schlussbericht Schwerpunktsprogramm Umwelt (SPPU) des Schweizerischen National Fonds (SNF). CEPE Working Paper No. 11, Zürich, Dezember 2001.

Filippini M., Banfi, S., Impact of the new Swiss electricity law on the competitiveness of hydropower, CEPE Working Paper No. 12, Zurich, January 2002

Filippini M., Banfi, S., Luchsinger, C., Deregulation of the Swiss Electricity Industry: Implication for the Hydropower Sector, CEPE Working Paper No. 13, Zurich, April 2002

Filippini, M., Hrovatin, N., Zoric, J., Efficiency and Regulation of the Slovenian Electricity Distribution Companies, CEPE Working Paper No. 14, Zürich, April 2002

Spreng D., Scheller A., Schmieder B., Taormina N., Das Energiefenster, das kein Fenster ist, CEPE Working Paper No. 15, Zürich, Juni 2002

Fillippini M., Pachauri Sh., Elasticities of Electricity Demand in Urban Indian Households, CEPE Working Paper No. 16, Zurich, March 2002

Semadeni Marco, Long-Term Energy Scenarios: Information on Aspects of Sustainable Energy Supply as a Prelude to Participatory Sessions, CEPE Working Paper No. 17, Zurich, July 2002

Müller Adrian, Finding Groups in Large Data Sets, CEPE Working Paper No. 18, Zurich, October 2002

CEPE Reports und ***CEPE Working Papers*** sind teilweise auf der CEPE-Homepage (www.cepe.ethz.ch) erhältlich oder können bestellt werden bei: CEPE, Sekretariat, ETH Zentrum, WEC, CH-8092 Zürich.