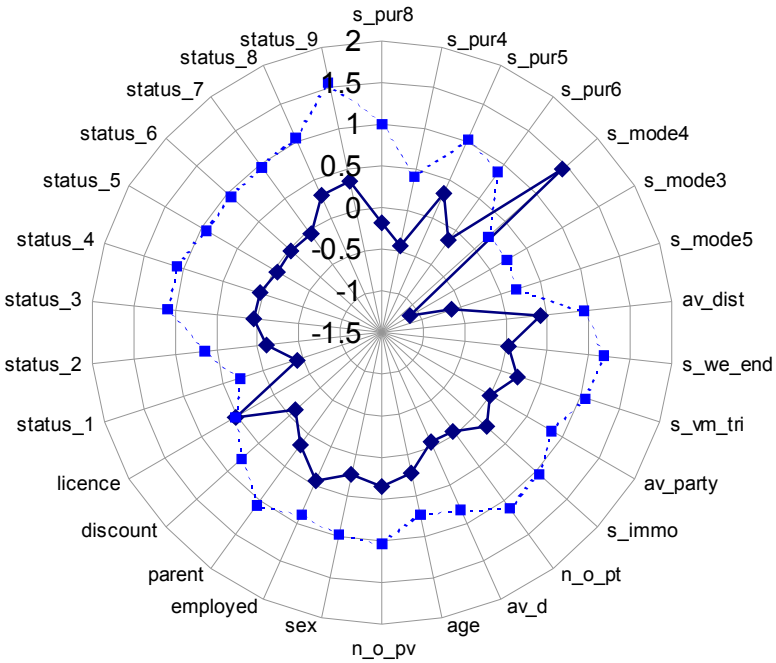
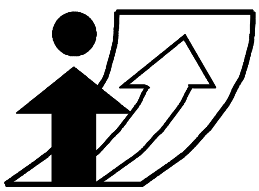
Cluster 1



Homogenous groups of travellers

Robert Schlich , IVT; ETH Zürich

Conference paper
Session 2.1 Biographies and life styles



Moving through nets:
The physical and social dimensions of travel
10th International Conference on Travel Behaviour Research
Lucerne, 10-15. August 2003

Homogenous groups of travellers

R Schlich
Institute for Transport Planning and Systems (IVT)
Swiss Federal Institute of Technology (ETHZ)
CH-8093 Zürich
Switzerland

Telephone:+41-1-633 30 85
Telefax: +41-1-633 10 57
schlich@ivt.baug.ethz.ch

Abstract

The segmentation of the population into groups of people with homogenous travel behaviour has been an important issue to travel behaviour analysis for a long time. Still the classifications commonly used are far from satisfactory because they only explain a small amount of variability within the groups. This is mainly due to two different obstacles: The first obstacle is the lack of suitable longitudinal data, the second is the gap how similarity is measured and how the order of activities is considered in the measurement.

Both obstacles shall be addressed in this paper. Based on a six week travel diary two different segmentation approaches are compared: The first uses traditional measurements but a rich set of variables and the complete set of days, the second is based on the multidimensional sequence alignment method with a smaller number of variables and a subset of random days. The resulting cluster solutions are compared and briefly introduced.

Keywords

Sequence alignment, optimal matching, homogenous groups, travel behaviour, levenshtein distance, International Conference on Travel Behaviour Research, IATBR

Preferred citation

Schlich, R (2003) Homogenous groups of travellers, paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003.

1. Introduction

The segmentation of the population into groups of people with homogenous travel behaviour has been an important issue to travel behaviour analysis for a long time. It can be understood as a very basic response to the need to bring order into a complex process. Aim of this classification is to identify groups of people who are very similar to each other concerning their travel behaviour but clearly distinct from the members of other groups.

Despite the long tradition to construct behavioural homogenous groups, there has not been much progress in the last 20 years. Some older classifications (Kutter, 1972, Pas, 1983; Schmiedel, 1984, Huff and Hanson, 1986, 1988) are still the state of the art. This is even more surprising as those classifications are far from satisfactory because they only explain a small amount of variability within the groups.

The major obstacle in addressing the first issue is the absence of suitable data. The data is insufficient insofar as earlier classifications are based on cross sectional data. This means that some aspects of behaviour (e.g. the question of intrapersonal variability) could not be addressed by these classifications at all. Moreover, Huff and Hanson (1988) showed, that the chance of misclassification of a person is much higher if cross sectional data is used – and thus the source of variability within the clusters.

Travel behaviour can be described by many indicators such as the number of trips per day, the mean trip distance or the order of activities. Different indicators are looked at in different studies analysing behaviour and no consensus has been achieved concerning the question which one is relevant. The question of how to measure similarity in behaviour is connected to the context of travel. According to Hägerstrand (1970) human behaviour can be viewed as a sequence of interdependent actions in time and space. However, daily behaviour is treated in travel behaviour research mostly as if it consisted of a chain of independent activities. Thus the sequential order or duration of activities is often neglected. But in order to forecast activities correctly it is much more important to look at activities a person already did, than at any other item. This neglect is even more obvious as the sequential order is usually collected in time budget studies or travel diaries and thus could be available. Abbotts statement ”We assume intercase independence even while our theories focus on interaction” (1995, p. 94) about the social science is as well true for travel behaviour research. According to Wilson (1998b) the main reason for ignoring the order of activities was the lack of a suitably powerful tool to analyse it.

Both obstacles (insufficient data and measurements that cannot consider the order of activities) shall be addressed in this paper. In the following section, the data obtained from a longitudinal study is briefly presented which allows for the measurement of intrapersonal variability. In the third section the sequence alignment method as a major improvement to measure behavioural similarity is introduced. The forth section describes the methodology used for the clustering process and the reasons for choosing two different approaches as a similarity measure. The resulting groups of similar behaviour are then described in terms of their behavioural variability as well as their sociodemographic structure for both approaches. Finally the clustering results are discussed and further methodological gaps are discussed.

2. Data

The following analyses are based on a dataset which is to a large extent unique, especially in terms of the length of the reporting period and of the completeness of the dataset. It is the result of a six week travel diary conducted for the research project Mobi*drive*. Funded by the German Federal Ministry for Education and Research, in spring and autumn 1999 in the cities Karlsruhe and Halle/Salle 361 persons were interviewed. The project consortium consisted of the PTV AG (Karlsruhe), the Institut für Stadtbauwesen at RWTH Aachen and the Institute of Transport, Traffic, Highway and Railway Engineering (IVT) at ETH Zurich. A discussion of sampling procedures, the survey instruments and data quality is provided by Axhausen, Zimmermann, Schönfelder, Rindsfüser and Haupt (2001), frequencies of the characteristics of all variables are documented by Schönfelder, Schlich, König, Aschwanden, Kaufmann, Horisberger and Axhausen(2002)[1]. Although there are some variations in the average number of reported trips , there is no general decline with increasing length of the reporting period, neither by fatigue, nor by seasonal influences.

The participating respondents showed only small differences in terms of their sociodemographics compared to those who refused to take part in the survey. The recruited households have a higher income, more cars and more working members. Compared to other representative studies in those cities (based on single-day travel diaries), little difference in the general indicators of travel behaviour could be found – given the massive methodological differences between those studies (Axhausen *et al.*, 2000). Although this comparison does not ensure that the composition of the sample does not bias the results, there is no indication that it does.

[1] Both papers are available at http://www.ivt.baug.ethz.ch/vrp/arbeitsberichte_d.html

3. Theory: Measuring similarity with the sequence alignment method

The major problem of similarity measurement is the lack of a generally accepted procedure to identify similarity of activity/travel patterns over long periods. Usual behaviour indicators such as the number of trips per day, mean trip distance, or mean trip duration neither consider the temporal dimension of the activity chains, nor the complexity of behaviour and are thus unsuitable. Several complex measurement methods which differ substantially concerning their theoretical background and their level of complexity have been suggested in the past. Schlich and Axhausen (2003) provide a comparison using the Mobi*drive* data and a literature review.

It is particularly controversial which attributes to examine, how to classify and to weight them, and with which algorithm the values of the attributes should be compared. Thus, the measures lead to different results for the same data (Burnett and Hanson, 1982). Hanson and Huff (1988) generally notice that the more detailed a measuring procedure is and the more attributes it covers, the smaller are the observed similarities.

Some years ago Wilson (1998a, 1998b) introduced the sequence alignment to travel behaviour research – a new measurement that includes the order of actives. Since several advances were made with this measurement approach (see below), it seems reasonable to look at the method in greater detail before other methodological steps of this analysis are introduced.

The sequence alignment method is a promising approach for measuring behaviour. The method was originally developed in molecular biology to compare DNS or protein strings (Sankoff and Kruskal, 1983). The idea of comparing strings consisting of a sequence of different elements was also adopted in other scientific fields as well as in applied science (e.g. speech recognition). Wilson (1998a, 1998b) was the first to introduce sequence alignment to travel behaviour research, although the method has been adopted by social scientists for some time under the name ”optimal matching” (e.g. Billari, 1999; Schaeper, 1999; Erzberger and Prein, 1997 or Abbot and Tsay, 2000 for an overview). Since Wilsons’ work, important theoretical improvements have been made (Joh, Arentze and Timmermans, 2001a, 2001b, 2001c, 2002) and several empirical applications have been undertaken in travel behaviour research (e.g. Bargeman, Joh and Timmermans, 2002; Berger 2000a, 2000b; Hertkorn and Kracht, 2002; Rindsfüser and Doherty, 2000; Schlich, 2001; Wilson, 2001).

The measurement of similarity is usually based on different attributes such as activity type, transport mode, starting time, trip or activity duration or trip destination. Each of these attributes of an observation is compared to the corresponding attributes of another observation. This may be a single trip or activity, a whole day or a sequence of trips. Two sequences of trips can be shown as $\boldsymbol{s}=\boldsymbol{s}[s_1.....s_m]$ and $\boldsymbol{g}=\boldsymbol{g}[g_1.....g_m]$ with n and m showing the total numbers of trips per sequence. Mostly the attributes are compared for single elements of this sequence – e.g. the second trip duration is compared to the second trip duration of another day. The similarity is than calculated as a sum of the Euclidean distances of the attributes.

This methods lack the possibility to incorporate the sequential order of activities. Imagine the following sequences **s** (**s**ource) and **g** (tar**g**et) displayed in Figure 1 which represent activities in 15 minute intervals (with each letter representing a different activity).

Figure 1: Pairwise comparison of two sequences representing activities

Example of two sequences

s: *WW* WW TS ST HH HH TL LL LT HH

g: WW TS ST HH HH TL LL LT HH *HH*

Calculation of similarity:

$d(s,g)=$ and $f(x)= 1$ if $si \neq gi$

$f(x)= 0$ if $si = gi$

Activities:

(W: working; T: travel; S: shopping, L: leisure, H: home)

Source: Schlich (2001)

If the distance between both chains would be measured pairwise with the score of a one for a mismatch and a zero for a match, the distance between both sequences would be measured as 12 (for a string of 20 elements) although in both sequences the same activities are performed in the same order and for the same duration. The only difference is, that in the second sequence all activities after work start half an hour (or two intervals) earlier and that the first

string possesses two intervals of working at the beginning instead of staying at home in the other one at the end. Thus they are very similar in an intuitive way.

Improvements to this simple way to calculate Euclidean distance were introduced by Pas (1983) and Hanson and Huff (1986) who added for example different weights and the serial dependence of different attributes to the similarity function. Clarke and Jones (1988) analysed behaviour classified in time interval of 15 minutes duration. Nonetheless their similarity functions ignore the sequential order of activities and their interdependencies. Schlich and Axhausen (2003) showed that the observed variability depends strongly on the chosen measurement type. Thus it is necessary to improve similarity measurement not only to incorporate order and duration but also to establish a common standard.

The idea of the sequence alignment method is to look at the two sequences **s** and **g** and to equalise them by different operations. This idea of measuring a quantitative distance for qualitative data does not seem to be intuitive at all (Wilson 1998b). The possible operations are substitutions, insertions and deletions. Insertions and substitutions are sometimes called indels. The implied effort of each operation can be accounted for by different weights. Mostly, the weight of one is assigned to the operations deletion and insertions. The weight for substitutions can be understood as the sum of the consecutive operations of a deletion and an insertion and is thus the value of two. This can be written as follows:

- Insertion: $w_i(s_i, g_i)=1$
- Deletion: $w_d(s_i, g_i)=1$
- Substitution: $w_s(s_i, g_i)=2$

As there is usually more than one way to change the sequence **s** into **g** by substituting, deleting and inserting characters into the strings, the smallest sum of the weighted operations is called the Levenshtein distance (Levenshtein, 1968) and each way of equalising the sequences is called an alignment. An example is given in Figure 2.

Figure 2: Two possible alignments for the sequences **s** and **g**

Sequences:

s=CAMBRIDGE
g=CAMPING

Distance Sequence alignment :

1) substitute s_4(B:P), s_5(R:I), s_6(I:N), s_7(D:G) delete s_8(G), s_9(E) => d=10

2) substitute s_4(B:P), delete s_5(R), substitute s_6(D:N), delete s_8 (E) => d=6

Source: Schlich (2001)

The advantage over conventional measurements becomes clear, if one imagines the sequences ***s***=***s****[ABCDEFGH]* and ***g1***=***g1****[ADEBFGCH],* respectively ***g2***=***g2****[AFGBDECH]* (Joh *et al.*, 2002). For the pairwise comparison the distances had a score of 6 units – with the sequence alignment it is $d(s,g1) = 4$ respectively $d(s,g1) = 6$ units. According to Joh *et al*. (2002) the sequence alignment distinguishes between the ”wrong position but the same order” and ”wrong positions and different orders”. This approach can be used both to calculate a similarity between two strings as well as a distance between them. To calculate the similarity a reward is given for each match and penalty for each mismatch in the string, while the distance is calculated as the Levenshtein distance. A comprehensive description of the method can be found in Joh *et al*. (2002).[2]

The adoption of a method to a completely different field is connected with many problems. In our case the major problem is, that travel behaviour cannot be represented by a single attribute. Instead it has to be characterised by multiple attributes such as trip purpose, trip destination, travel mode or trip departure time. Unfortunately all these attributes have different measurement scales so that the methods for multidimensional alignment used in molecular biology (see McClure, Vasi and Fitch, 1994) cannot be adopted.

[2] The same fact can be expressed by the term distance (Joh *et al*. 2002) or similarity (Wilson, 1998). Wilson (1998) states that none of the two expression has clear advantages in opposite to the other. In this paper the terminology of Joh is used.

Wilson (1998b) suggests to construct separate variables for each combination of values of different attributes, which can than be compared. Main disadvantage of this method is that it cannot discover, if single values of some of the attributes were equal. Furthermore the measured similarity would get very small with increasing numbers of attributes and combinations. Thus this method seems inappropriate.

If all variables were independent from each other then the distances for k attributes could be calculated separately. The distances could then be weighted with the weighting factors β_k according to their importance and summed up. This method is called uni-dimensional sequence analysis (UDSAM) and can be written as follows:

$$d(s,g) = \sum_{k=1}^{K} \beta_k (s_k, g_k)$$

In reality the different attributes of an activity or trip depend on each other – for instance, the choice of a travel mode is influenced by the chosen activity. If all attributes were connected to each other in the same way, it would be sufficient to calculate the distance as the distance of the attribute which is given the maximum weight. With this measurement the distance would be smaller than measured with UDSAM. However, both treating the different attributes of an activities as totally dependent or independent is not justified in most cases.

Mostly, the different attributes are partially dependent. If each attribute is represented by a single sequence, then for those attributes which are connected, the same operation has to be performed at the same position in different sequences – for those which are independent from each other the operations will differ. Elements in the sequence which can be aligned simultaneously without calculating the cost twice because the same operations are performed across attributes are called segments (Joh *et al.*, 2002).

Identifying segments leads to a reduction in the total alignment costs and is thus a major task for the calculation of the Levenshtein distance for sequences with different attributes – this method is called multidimensional sequence alignment (MDSAM). The only way to get the optimal result, is to calculate all possible alignments of each attribute and compare all possible combinations of the alignment across all attributes and identify the minimal costs.

As a multiple alignment of large sample of trips is at present not possible due to the enormous computing time of this effort, Joh *et al.* (2002, 2001b) developed an heuristic approach, called optimal trajectory multidimensional sequence alignment method (OTMSAM), which approximates the complete multidimensional approach. They proposed, that not all alternatives of alignment have to be calculated across all attributes, but just those who scored

the minimum distance for each attribute. It could be shown that this reduces the required computing time substantially. Although the resulting distances correlate strongly with the sum of the unidimensional approach (r=0.95) this is in important development for the application of sequence alignment in travel behaviour research.

4. Methodology

4.1 Application of the sequence analysis

The sequence alignment technique is a new method which can be called established by now. There are still some controversies about its use and correct application. In this section those controversies will be briefly introduced and how they were addressed in this analysis.

According to Dollase, Hammerich and Tokarski (2000) there are different forms of sequence alignment, depending on the incorporation of the duration of activities. If travel behaviour is observed with a travel diary the beginning and end time of an activity is known. It is then possible to classify a day into time intervals and to assign a main activity to each interval. Usual interval lengths used in previous applications were 5 to 15 minutes. The duration of the intervals will influence the results, because long intervals will neglect short activities. Furthermore the duration dominates the calculation of the similarity.

One further problem in this context is the question of how to deal with night hours. As they are normally equal in terms of all attributes this can bias the results. Wilson (1998b) showed, that the measured distance is smaller with shorter intervals and longer sequences. Thus the duration of activities is given a higher weight if short time intervals are chosen compared to considering the duration as a normal attribute in sequences with one letter for each activity. As this focus on the duration is not an aim of this work the sequences are not specified by time intervals in this analysis.

A second problem is the choice of the weighting parameters and the costs or penalties for different operations (deletion, insertion and substitution) which lack a theoretical foundation. At present there is little information about the weights for the different attributes. It is common practice in the social sciences (Schaeper, 1999), that the weight of an insertion or a deletion is fixed as half a substitution. For the absolute values of the operations there is no unambiguous criteria. According to Abbott and Tsay (2000) or Abbott (2000) they must be fixed differently depending on the subject of each analysis. Intuitively the costs for the operations should differ for different values of the attributes. Otherwise the quantitative

differences of the attributes of two sequences are lost in an alignment, as continuos variables are transformed into discrete classes. But this problem is true both for qualitative and quantitative variables – for example a substitution of a trip made by motorbike with a car trip seems to be less expensive than to substitute the car trip with a trip by foot. Unfortunately there is currently no theoretical framework for the determination of these costs. The choice of different costs for different values for each analysis is criticised vehemently by Wu (2000) and Levine (2000). They point out that the choice of different weights by the researcher without a common theory is too subjective. This is intransparent at best and makes it impossible to test results and at worst it makes them arbitrary. Due to this critique no weighting scheme for different values of attributes is used. The costs for all insertions and deletions are fixed at one and for substitutions at two in this analysis.

Sequence alignment is moreover criticised because it neglects the content of the analysed subjects. Wu (2000) points out that in reality the substitution of a value *a* by a value *b* is not the same as the reverse substitution. He illustrates this by an example from social science: the analysis of life cycles where each letter represents a working status. In reality changing from being employed to unemployed is much easier than vice versa. Nonetheless the sequence alignment measures the same alignment costs.

Lesnard (2002) looks into detail at the operation deletions and insertion. He criticises that the real order of activities is broken by those operations. If the use of deletions and insertions would be avoided by higher costs compared to substitutions than the alignment would be a common matching procedure.

> „Therefore indel which are costs too small in comparison with substitution costs lead to the vanishing of the temporal shifts between sequences. Consequently Andrew Abbot‘s recommendation is to minimise the use of indel operations in favour of substitutions. As a matter of fact, when the main goal is not to detect pattern of consecutive events then the indel operations are useless. But if only substitution operation are used then there is no more an optimal matching method but simply a matching procedure or a sequence comparison“ (Lesnard, 2002, S. 8)

Levine (2000) argues along the same line by saying that the deletion and insertion operations had a meaning for its original application in molecular biology. This meaning is lost by transferring the method to behavioural research so that causal connections cannot be detected.

Abbott (2000) replies to both arguments that it is true because the sequence alignment is a descriptive method. It is important not to be confused about its possibilities: Aim of the sequence alignment is the detection of similar patterns and not to depict processes in real life by transition probabilities to change one state into another.

A last decision was necessary concerning the software to use. At present there are several programs available. (DANA (C.H.Joh@bwk.tue.nl)), ClustalG (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX), TDA (http://steinhaus.stat.ruhr-uni-bochum.de/tda.html), Optimize (http://www.spc.uchicago.edu/users/abbot/om.html#optimize)). To my knowledge, DANA[3]. is the only software which allows the multidimensional sequence alignment. As this addresses one main problem of sequence alignment, the optimal trajectory multidimensional sequence alignment method (OTMSAM), which is a compromise between exactness and computing time was chosen.

4.2 General framework for the application

A traditional clustering approach consists of three different steps: the choice of relevant variables, the choice of a similarity measure to calculate the similarity between different persons and the choice of a fusion algorithm to merge similar persons into one cluster.

These steps are interconnected. Due to the advantages of the sequence alignment technique over common similarity measures this similarity measure is chosen here. This has implications for the further proceeding.

Comparison of daily programmes

Usually similarity is measured between persons and their behaviour indicated by some variables over the whole observation period. e.g. the mean number of trips per day and persons. These variables are merged into one similarity measure and a similarity matrix between each of the persons in the survey (361 in this case). The sequence alignment instead looks at different daily programs and calculates the Levenshtein distance as a similarity measure between each day of a person to all other days. This would mean that 15'162 days (361 persons with 42 observation days each) would have to be compared with each other. The resulting 115 million comparisons would take too much time. On a PC[4] the calculations with the software DANA using four attributes to describe the activities and the OTMSA method

[3] Dissimilarity ANalysis of Activity-travel patterns, Developed by C.H. Joh, T.A. Arentze and H.J.P. Timmermans Urban Planning Group, Faculty of Architecture, Building and Planning, Eindhoven University of Technology, http://www.bwk.tue.nl/urb, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[4] Pentium (R) III, 1.2. GHz, 512 MB RAM

took about one minute for 170 comparison – which means that a comparison of all days would take about 15 month.

For this reason one random day for each person was chosen from all workdays and each of the resulting 361 random days were compared to all other days. This resulted in a similarity matrix between random daily programmes for each of the participants.

This makes clear that the choice of the similarity measurement and the choice of variables is connected. Due to choosing a complex measurement the number of variables which are considered is limited and further restrictions next to the random day sample were made:

The comparisons are based on sequences which do not take the duration of the activities into account. The distance was then calculated with the proposed method of multidimensional sequence alignment of Joh *et al.* (2002) for the attributes trip purpose, trip mode, trip distance and departure time, with all attributes weighted equally. Days without trips or days when the interviewed persons stayed outside their home town were coded separately and were not excluded from the analysis, because they are essential to the question of how variable the behaviour is. Sequences of different lengths were not treated differently as a different number of trips is a crucial difference in this case. Thus the costs for the insertions to the shorter sequence is one for each trip.

Comparison of persons

The procedure to compare the daily programs of a random day has the disadvantage that it omits one main advantage of the available data: the possibility to consider the intrapersonal variability as one dimension of travel behaviour. It would be desirable to compare all days to each other instead of a random day per person - due to time constraints this was not possible here. For this reason and to check the robustness of the solution a second clustering solution was performed. For this approach the chosen variables were calculated for each person over the whole observation period. All variables were then merged into a similarity matrix with the squared Euclidean distance as similarity measure. The dimensions of travel which were considered and the variables chosen are shown in the following Table 1.

Table 1 Dimensions of travel behaviour and chosen variable

Dimension	Variables chosen
Trip purpose	Share of leisure, school, work, shopping [%]
Timing	Share of trips in the morning [%] Share of trips at weekends [%]
Duration	Mean duration / trip[min]
Distance	Mean distance / trip[min]
Trip Mode	Share non-motorised, public transport, private motorised transport [%]
Frequency of trips and immobile days	Number trips/ day [N] Share of immobile days [%]
Intrapersonal variability	Levenshtein distance
Coupling constraints	Number of accompanying persons [N]

Due to the long duration period it was possible to look at the variables over time. Instead of the number of trips with a particular mode it was possible to calculate the share of one mode over the six weeks. One advantage of these variables is, that the behaviour of a person is described more precisely because outlier are more frequently balanced out. Furthermore all variables are scaled metrically which is a prerequisite for the calculations of the Euclidean distance and some of the tested cluster algorithm. In order to calculate the squared Euclidean distance as a measure of distance between each person these variables were standardised by the sample mean.

5. Implementation

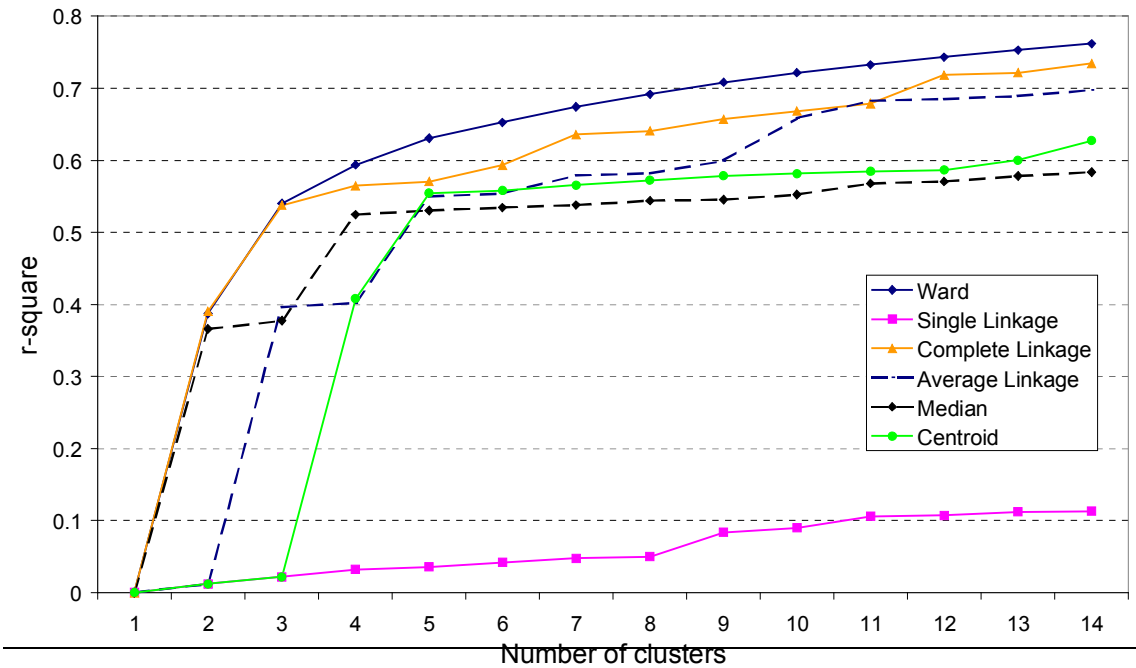
Before starting the segmentation it is useful to point out, that a classification cannot be wrong or right, but useful or useless:

> „In modern research on classification, especially in cluster analysis, it may be useful to remember occasionally that, in a sense, all classification is more or less arbitrary, that its boundaries are fuzzy (.....), and that we probably should ask more often in how much (not: whether or not) a given classification scheme matches our observations.“ (Hampel, 2002, 3)

For the segmentation of the sample based on personal characteristics over the whole observation period the clustering algorithm Ward's minimum variance cluster algorithm was chosen (Ward, 1963). This method joins clusters from the previous generation by minimising the sum of squares over all partitions. This algorithm obtains clear partitions between groups if the variables are uncorrelated which was tested before for the chosen variables. Other hierarchical partitioning cluster algorithms (centroid, median, single linkage, complete linkage, average linkage) were tested in order to test how robust the solutions are with respect to the chosen algorithm.

The main problem of all cluster analysis is to fix the number of clusters which are chosen as optimal. This number was selected here by considering the sum of squares for all clusters, the pseudo *F*-statistic, and the overall proportion of variance accounted for by the clusters (r^2). Using these criteria the number of clusters was fixed as five. An overview over the explained variance for the different clustering algorithms with different numbers of clusters is given in Figure 3.

Figure 3 Number of clusters for the comparison of persons: r-square

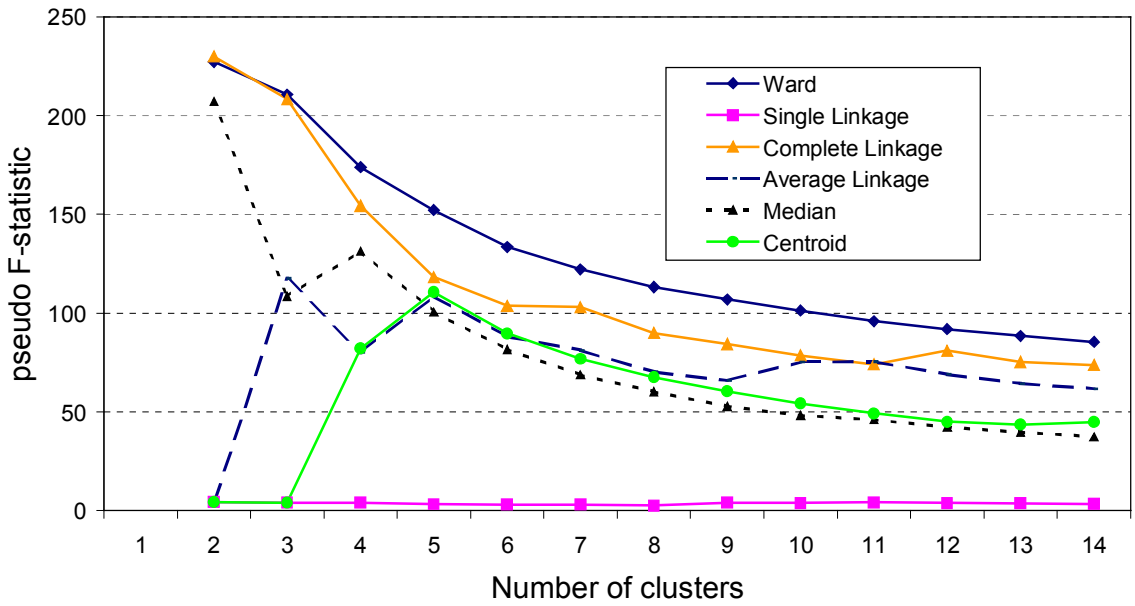


The analysis showed that there is a strong decrease in the sum of squares below five clusters for most clustering methods; the variability within the clusters is increasing strongly for a smaller number of clusters. At this stage 63% of the variance can be explained by these clusters with the Ward algorithm while a larger number of clusters results only in a very small

increase in explanatory power (r^2= 0.65 for six clusters). Although the difference in explained variance between for and five clusters is smaller for Wards algorithm than for other algorithms the number was chosen because for every other number the difference was also small except for a partition with two clusters –which would not explain much.

The number is also confirmed by the pseudo *F*-statistic which measures the separation among all clusters at each generation of clusters. This statistic shows a continuous decline for five clusters which indicates that no other partition is more appealing. Another argument in favour of the five cluster solution is that the number of persons in each cluster is similar in all clusters.

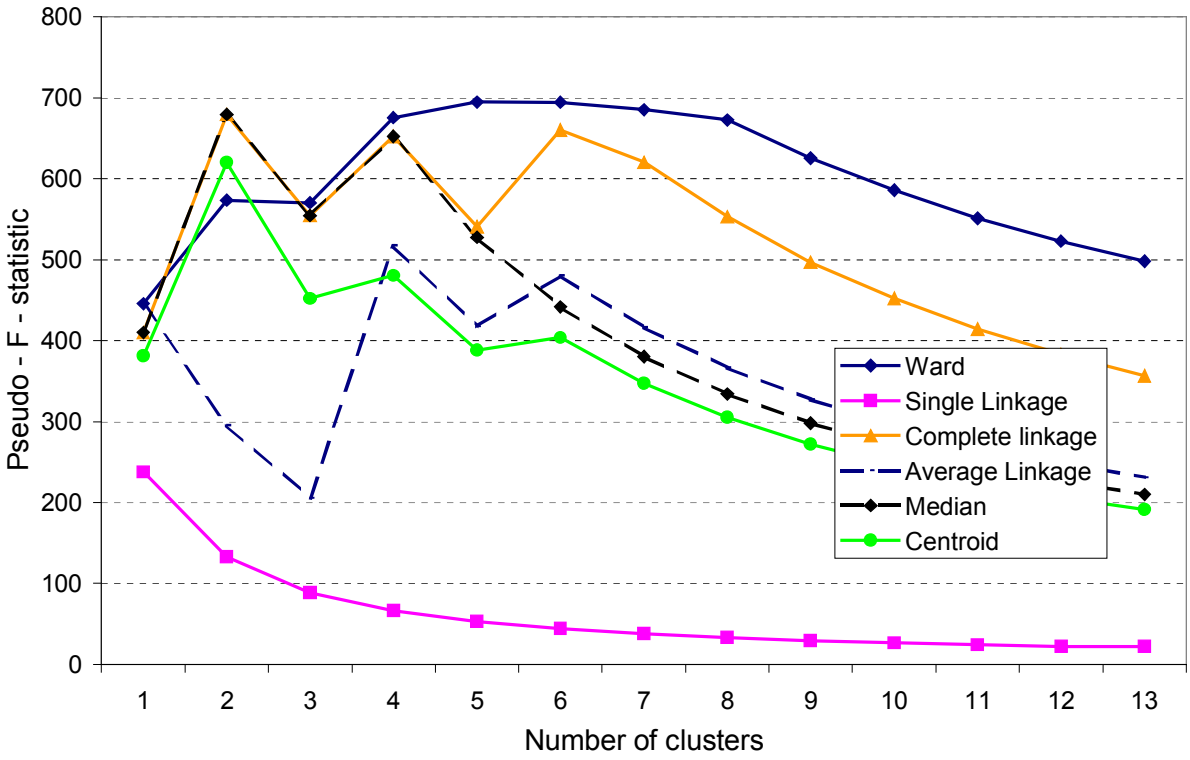
Figure 4 Number of clusters for the comparison of daily programmes : pseudo F statistic



Similar results concerning the number of clusters were obtained if the cluster process was based on the similarity matrix for random daily programmes calculated with the Levenshtein distance. In general all clustering algorithms could explain a higher proportion of the variability between the clusters for this input matrix then the one calculated with the person matrix. Again the Ward algorithm had the highest amount of explained variability compared to all other algorithms. Although there is a steady increase in the total variance accounted for with growing number of clusters the number of five was again assessed as useful for two reasons. This was because the Pseudo-F-statistic showed a local maximum value for five

clusters (Figure 5) calculated with the Ward algorithm. In addition, this number allows an easier comparison of the results between the two approaches. The number of persons per clusters for this solution gives clusters with a different numbers of persons.

Figure 5 Number of clusters for the comparison of daily programmes: pseudo F statistic

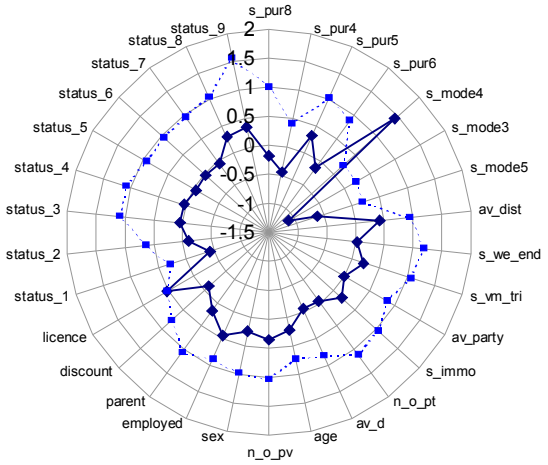


6. Comparison of results

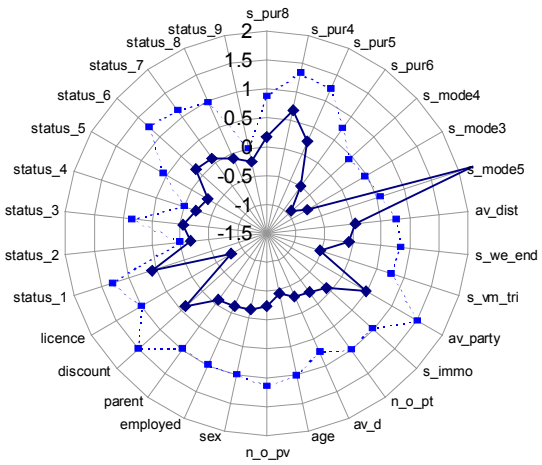
The results of the two partitions are quite different with respect to travel behaviour and socio-demographic structure of each cluster. For the traditional clustering of persons, some of the resulting clusters can be described quite well – they are clearly distinct from each other both in terms of sociodemographic and their behavioural variables. While Figure 6 gives an overview over standardised values to identify peculiarities in single clusters quickly, a comprehensive table with all mean values to describe the different clusters is given in the following Table 2 and Table 3. The thick line in the figure gives the mean value of the variables in each cluster, the broken lines gives the standard deviation.

Figure 6 Sociodemographic and behavioural description of all person-clusters

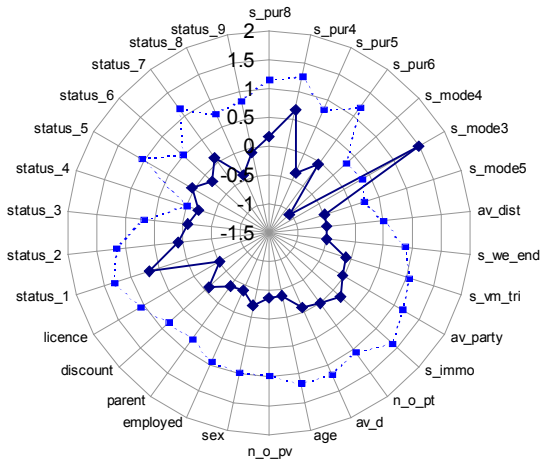
Cluster 1



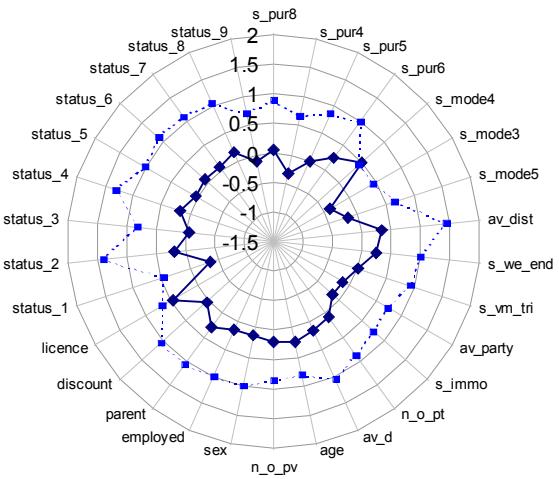
Cluster 2



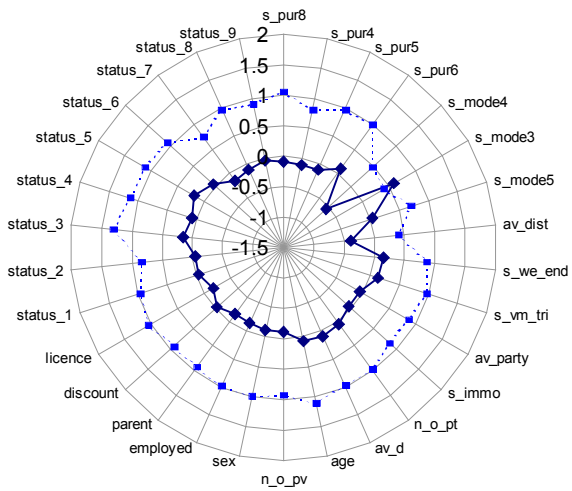
Cluster 3



Cluster 4



Cluster 5



s_pur8: % of leisure trips
s_pur4: % of school trips
s_pur5: % of working trips
s_pur6: % of shopping trip
s_mode4:% of car trips
s_mode3: % of unmotorised trips
s_mode5: % of pt trips
av_dist: mean distance/trip
s_we_end: % of weekend trips
s_vm_tri. % share of am trips
av_party: size of group
s_immo: % of immobile dayd
n_o_pt. num trips/day
av_d: intrapersonal variability
age:

n_o_pv: num. Of pvs
sex
employed
parent
discount
licence
status_1: pupil
status_2: student
status_3: Apprentence
status_4 Housemaker
status_5 Retiree
status_6 Unemployed
status_7 Parttime
status_8 Fulltime
status_9': Self-employed

The biggest differences occur for the behavioural variables, especially for the different modes used, but also for the different trip purposes. For the number of daily trips, the share of immobile days in the survey and the level of intrapersonal variably the differences are surprisingly small.

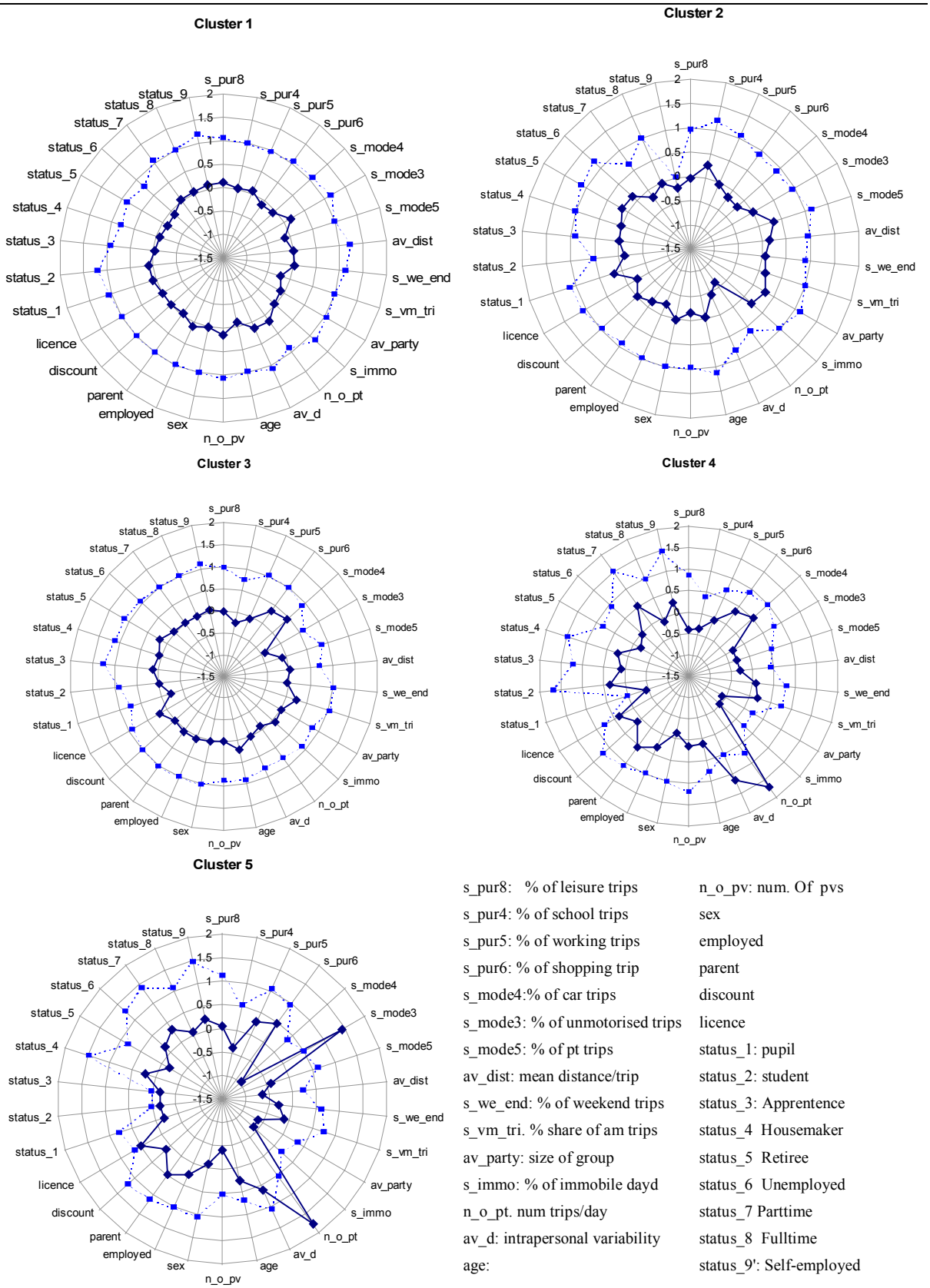
For each cluster clear variations in the sociodemographic composition of the members can be detected – different shares occur for the variables employment, sex, parent, driving licence but also for the age or the number of personal vehicles per household.

Concerning the composition of the clusters based on Levenshtein distance the results look quite different. As Figure 7 shows there are small differences in the average values of sociodemographic and behavioural variables – especially the cluster 1 to 3 look similar to each other.

In general all variables show less dispersion over the five clusters which is especially true for the sociodemographic variables. There are no typical characteristics in the different clusters. The dissimilarity for behavioural characteristics are also small, except for the number of personal trips or the average variability in behaviour.

Obviously, the construction of the clusters with the sequence alignment captures information which is not included in traditional similarity measurement and clustering. This additional information is based on the order and number of different activities of the day. As one cannot find typical characteristics in the cluster it seems plausible to assume that this vital information concerning travel behaviour is not correlated to the sociodemographic characteristics of people.

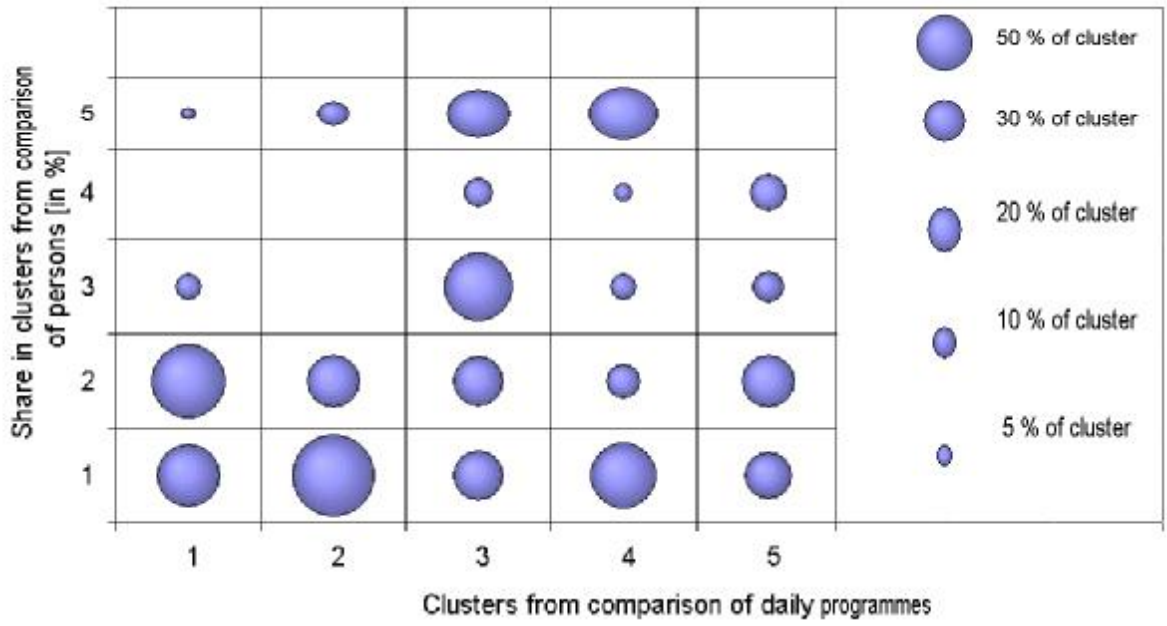
Figure 7 Sociodemographic and behavioural description of all daily programme-clusters



s_pur8: % of leisure trips
s_pur4: % of school trips
s_pur5: % of working trips
s_pur6: % of shopping trip
s_mode4:% of car trips
s_mode3: % of unmotorised trips
s_mode5: % of pt trips
av_dist: mean distance/trip
s_we_end: % of weekend trips
s_vm_tri. % share of am trips
av_party: size of group
s_immo: % of immobile dayd
n_o_pt. num trips/day
av_d: intrapersonal variability
age:
n_o_pv: num. Of pvs
sex
employed
parent
discount
licence
status_1: pupil
status_2: student
status_3: Apprentence
status_4 Housemaker
status_5 Retiree
status_6 Unemployed
status_7 Parttime
status_8 Fulltime
status_9': Self-employed

This is supported by a comparison of the cluster solutions. For each cluster of the first solution (based on the comparison of persons) it was calculated to what percentage the members in each cluster were allocated to the different clusters of the comparison of daily programmes. If a huge amount of people were allocated into one cluster with the same other people for both classification, this would be indicated by huge shares in the cross classification of cluster membership. This pattern occurs if you compare different algorithms for the clustering process for each clustering approach separately. Figure 8 shows the different shares for the cross classification of different clustering approaches and makes clear that both methods assign different persons to the clusters.

Figure 8 Cross classification of cluster membership



7. Description of the clusters from the comparison of persons

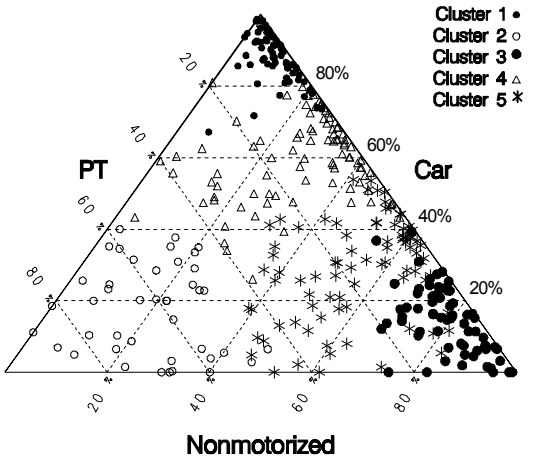
The comparison showed, that the approach of comparing daily programmes based on the Levenshtein distance captures some information that is usually not integrated in classifications. It is not surprising that those clusters cannot be described in terms of sociodemographics. If the clusters were different from each other in this regard, this would mean, that the order of activities is captured by traditional measures.

At this stage of the work it was not possible to do a joint segmentation integrating both approaches. As it is difficult to describe the clusters based on the comparison of daily programmes in terms of sociodemographics, only the clusters based on the comparison of persons will be characterised briefly.

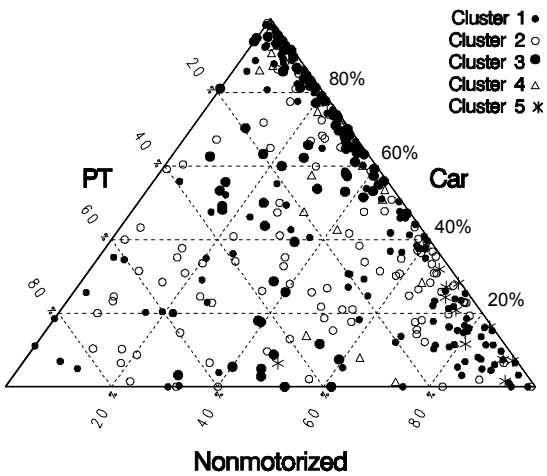
The following figures show as an example the share of different modes used over the entire survey period (public transport, private transport, non motorised) for each person. Each symbol (dots in two different sizes, circle, triangle or star) shows the individual combination of different modes used of one person over the survey period by cluster membership. At each axis, the share of this mode is given.

Figure 9 Mode choice by cluster membership

Comparison of persons



Comparison of daily programmes



The figures illustrate the differences for the modes used from members of the different clusters. Instead of showing just the means they show the individual different combinations. Members of the same cluster have similar combinations to other members of the same cluster which distinct clearly from the other clusters. In contrast to this, members of the same cluster based on the comparison of daily programmes do not have similar combinations.

The figures help together with the mean values of some behavioural and sociodemographic variables given in Table 2 and Table 3 to characterise the five different clusters .

Cluster one consists of persons, who nearly exclusively use the car as a mode. Consequently they have the highest average distance per trip of all clusters, the highest share of working trips and of trips performed in the morning. As this cluster also has the highest share of male persons (62%), employed persons (74%) and the biggest number of cars per household it is reasonable to characterise the cluster as the "working men" cluster. Surprisingly the members of this group are more often immobile on a day than members of other clusters, while the number of trips per day and the variability in their daily behaviour is average.

Cluster two is together with Cluster three the only cluster with a substantial share of school trips (11%) and the cluster with the highest share of leisure trips (17.5%). It consists of a high percentage of pupils and its members are thus in average younger than any other cluster-members, save the members of cluster tree. Both clusters have a very small share of car trips (less than 20%) but in contrast to cluster three, members of cluster two travel mostly by public transport. Nonetheless the cluster cannot be characterised as "pupil cluster" as it has a substantial share of fulltime employed person (31%). More typical seems the smallest level of intrapersonal variability and the smallest number of daily trips which characterises the cluster as the cluster "with stable behaviour".

Some characteristics of the third cluster were already mentioned. Beside, most striking is the high share of nearly 80% of unmotorised trips. Correspondingly, they have the smallest number of cars per household. The cluster consists to a high percentage of pupils and students (together 50%) and retirees (nearly 20%). Nearly two out of three members of the cluster are women. As the average distance of their trips is smaller than the distance of trips from other cluster members and their share of immobile days is high, the cluster is locally oriented– it can be called "local" cluster.

Table 2 Behavioural characteristics of the clusters

	Cluster 1 (n=85)		Cluster 2 (n=42)		Cluster 3 (n=66)		Cluster 4 (n=92)		Cluster 5 (n=92)	
Variable	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
leisure trips [in %]	14.6	8.7	17.7	7.6	17.6	10.0	16.5	7.6	15.4	9.2
School trips [in %]	0.9	3.7	11.1	12.0	11.1	11.3	2.0	5.9	4.1	7.3
Working trips [in %]	12.7	11.5	11.6	13.5	4.9	9.1	8.8	9.5	7.7	10.8
shopping trips [in %]	11.5	7.5	8.5	6.4	12.3	10.1	14.6	8.5	13.6	8.7
Car trips [in %]	88.9	6.9	17.6	12.5	14.5	9.3	60.7	12.5	28.7	13.8
Unmotorised trips [in %]	8.1	5.9	20.0	12.6	79.1	9.7	27.4	11.7	54.5	10.8
Public transport trips [in %]	2.1	3.9	61.6	11.4	5.3	5.0	11.1	13.5	16.0	14.6
Mean distance/trip [km]	11.8	7.0	8.9	5.7	5.0	3.7	11.3	10.7	5.8	3.1
Trips at weekend [in %]	22.0	7.7	21.4	5.5	18.6	5.7	23.4	6.5	22.9	5.7
Trips in the morning [in %]	44.8	14.7	34.9	10.3	40.6	14.4	41.9	12.5	43.8	13.4
Immobile days [in %]	9.1	10.3	6.3	9.6	8.9	13.6	5.6	7.6	6.8	8.5
Number of trips/day [n]	3.6	1.4	3.3	1.3	3.7	1.3	3.7	1.1	3.7	1.2
Intrapersonal variability [Levenshtein distance]	6.8	2.4	6.1	2.2	6.7	3.5	7.4	3.1	7.2	2.9

Table 3 Sociodemographic characteristics of the clusters

	Cluster 1 (n=85)		Cluster 2 (n=42)		Cluster 3 (n=66)		Cluster 4 (n=92)		Cluster 5 (n=92)	
Variable	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Age [years]	44.3	14.1	32.0	19.3	32.8	22.1	44.7	15.1	41.5	20.9
Num. of pv [n]	1.4	0.6	1.0	0.7	0.9	0.6	1.3	0.5	1.1	0.5
Sex [% of males]	62	49	43	50	39	49	55	50	45	50
Parent [%]	39	49	29	46	15	36	46	50	25	43
Discount card [%]	4	20	17	38	5	21	8	27	6	24
Licence [%]	93	25	31	47	44	50	89	31	60	49
Pupil [%]	1	11	40	50	44	50	3	18	16	37
Student [%]	1	11	0	0	5	21	7	25	2	15
Apprentice [%]	4	20	2	15	2	12	2	15	6	24
Housemaker [%]	5	22	0	0	0	0	8	27	6	24
Retiree [%]	14	35	5	22	18	39	17	38	24	43
Unemployed [%]	5	22	10	30	2	12	8	27	7	26
Part-time [%]	8	27	12	33	12	33	11	31	5	21
Fulltime [%]	49	50	31	47	15	36	42	50	31	46
Self-employed [%]	12	33	0	0	3	17	2	15	4	19

The next cluster is the cluster of "active families". It has the highest share of parents (46%) and a high percentage of employed persons (both fulltime and halftime). Members of this cluster seem to have a lot of obligations and are travelling a lot – mostly by car. They travel longer average distances for their trips than other persons (save cluster one), have the highest number of trips per day, the highest level of intrapersonal variability and the lowest share of immobile days. The highest share of shopping trips and of weekend trips indicates that those people travel for various reasons, not just for commuting.

Cluster five is difficult to describe. Its members are quite heterogeneous with respect to sociodemographic. All employment status are represented with an average share in the cluster – similarly to all other sociodemographics. The behavioural characteristics can also be described as average – except the high share of non – motorised trips and thus smaller average distances and a high level of intrapersonal variability. The cluster can be summarised as the "average" cluster.

8. Conclusion

The comparison of a clustering based on the comparison of random daily programmes (calculated with the Levenshtein distance as a measure of similarity) and a clustering based on persons provides two insights: Comparing daily programs with the sequence alignment method and the Levenshtein distance as similarity measure adds new information and modifies the results of other segmentation. While the results of the traditional approach with a large number of variables are quite independent from the chosen Ward algorithm (which means that persons were clustered into a group with the same person for other algorithms) they differ strongly from the Levenshtein solution.

The cluster analysis for the traditional measurement results in a five cluster solution which can be interpreted in a plausible way. Both in terms of behaviour (especially the mode choice) and sociodemographics they are clearly distinct from each. It could be useful to group people based on the detected major sociodemographics into five groups. This procedure would improve the traditional classification e.g. by Schmiedel (1984) due the new longitudinal data and the richness of the used variable set, including the intrapersonal variability.

The second result is that the clusters based on the comparison of random daily programmes with the sequence alignment do not differ concerning the sociodemographic composition. This is hardly surprising – the order of activity is not dependent on characteristics like age and gender. The sequence alignment method is able to depict this issue and can help to constitute

groups which are homogeneous concerning behaviour. Unfortunately they are not easy to identify due to the absence of typical sociodemographic characteristics.

The results need to be further improved. A major obstacle is that the clusters for daily programs are based on one random day. The calculations should be repeated with more time for a bigger number of random days for each person. Another topic for further research is the question how to combine different cluster solutions.

9. Acknowledgements

I wish to thank Chang-Hyeon Joh for his software DANA and for his help in addressing special problems. His work was a great help to me. Further acknowledgements go to the Collegium Helveticum, (http://www.collegium.ethz.ch/) where scholarship enabled me to work on this subject.

10. References

Abbott, A. and A. Tsay (2000) Sequence analysis and optimal matching methods in sociology: Review and prospect, *Social Methods & Research*, **29** (1) 3-33.

Abbott, A. (2000) Reply to Levine and Wu, *Social Methods & Research*, **29** (1) 65-76.

Abbott, A. (1995) Sequence analysis: New methods for old ideas, *Annual Review of Soci*ology, **21** (1) 93-113.

Axhausen, K.W., A. Zimmermann, S. Schönfelder, G. Rindsfüser and T. Haupt (2002) Observing the rhythms of daily life: A six-week travel diary, *Transportation,* **29** (2) 95-124.

Bargeman, B., C.H. Joh and H.J.P. Timmermans (2002) Vacation behaviour using a sequence alignment method*, Annals of Tourism Research*, **29** (2) 320-337.

Berger, M. (2000a) Formation of typologies of similar changes and differences in activity behaviour - a multi-method-approach and application of the optimal-matching-technique, paper presented at the 9th International Association of Travel Behaviour Conference, Gold Coast/Queensland, July 2000.

Berger, M. (2000b) Abbbildung und Erklärung von Unterschieden zwischen Aktivitätenmustern - ein Multimethodenansatz unter Verwendung der Optimal-Matching-Technik, *Stadt Region Land,* **69**, 145-155.

Billari, F.C. (1999) Sequence analysis in demography: Changes and wishes, paper presented at a workshop on Longitudinal Research in Social Science: A Canadian Focus, London (Ontario), October 1999.

Burnett, K.O and S. Hanson (1982) The analysis of travel as an example of complex human behaviour in spatially-constraint situation: Definition and measurement issues, *Transportation Research A*, **16** (2) 87-102.

Dollase, R., K. Hammerich and W. Tokarski (2000) *Temporale Muster – die ideale Reihenfolge der Tätigkeiten,* Leske und Buderich Verlag, Opladen.

Erzberger, C. and G. Prein (1997) Optimal Matching Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe, *ZUMA-Nachrichten,* **21**, 52-81.

Hägerstrand, T. (1970) What about people in regional science?, *Papers of the Regional Science Association*, **24** (1) 7-21.

Hampel, F. (2002) Some thoughts about classification, *Research Report,* **102** , Seminar für Statistik, ETH Zürich, Zürich.

Hertkorn, G. and M. Kracht (2002) Analysis of large scale time use survey with respect to travel demand and regional aspects, paper presented at the International Association of Time Use Research (IATUR), Annual Conference, 2002, Lisbon, October.

Huff, J.O. and S. Hanson (1986) Repetition and variability in urban travel, *Geographical Analysis*, **18** (2) 97-114.

Hanson, S. and J.O. Huff (1988b) Systematic variability in repetitious travel, *Transportation*, **15** (2) 111-135.

Lesnard, L. (2002) The professional arrangements of French dual-earner couples in the 80s and 90s – A global approach based on a sequence comparison algorithm, *Serie des documents de travail du Centre de Recherche en Economie et Statistique (CREST),* **2002-45,** Institut National de la Statistique et des Etudes Economiques, Malakoff Cedex.

Levenshtein, V. (1966) Binary codes capable of correcting deletions, insertions and reversals**.** *Soviet Physics Doklady,* **6**, 707 –710.

Levine, J.H. (2000) But what have you done for us lately? Commentary on Abbott and Tsay, *Social Methods & Research*, **29** (1) 34-40.

Joh, C.H., T.A. Arentze, F. Hofman and H.J.P. Timmermans (2002) Activity pattern similarity: A multidimensional alignment method. *Transportation Research B*, **36** (5) 385-403.

Joh, C.H., T.A. Arentze and H.J.P. Timmermans (2001a) A position-sensitive sequence alignment method illustrated for space-time activity-diary data, *Environment and Planning A*, **33** (2) 313-338.

Joh, C.H., T.A. Arentze and H.J.P. Timmermans (2001b) Multidimensional sequence alignment methods for activity-travel pattern analysis – A comparison of dynamic programming and genetic algorithms, *Geographical Analysis,* **33** (3) 247-270.

Joh, C.H., T.A. Arentze and H.J.P. Timmermans (2001c) Pattern recognition in complex activity-travel patterns: a comparison of Euclidean distance, signal processing

theoretical, and multidimensional sequence alignment methods, *Transportation Research Record* **1752**, 16-22.

Joh, C.H., T.A. Arentze and H.J.P. Timmermans (2002) Activity pattern similarity: A multidimensional sequence alignment method, *Transportation Research B,* **36** (3) 385-403.

Jones, P. and M. Clarke (1988) The significance and measurement of variability in travel behaviour, *Transportation*, **15** (1) 65-87.

Kutter, E. (1972) Demographische Determinanten des städtischen Personenverkehrs, *Veröffentlichungen des Instituts für Stadtbauwesen der TU Braunschweig* **9**, TU Braunschweig, Braunschweig.

McClure, M.A., T.K. Vasi and W.M. Fitch (1994) Comparative analysis of multiple protein-sequence alignment methods, *Molecular Biology and Evolution*, **11** (4) 571-592.

Pas, E.I. (1983) A flexible and integrated methodology for analytical classification of daily travel-activity behaviour, *Transportation Science*, **17** (4) 405-429.

Rindsfüser, G. and S. Doherty (2000) Konzepte, Module und Datenerfordernisse für SMART – Simulationsmodell des Aktivitäten-(Re)Planungsprozesses*, Stadt Region Land*, **69**, 109-131.

Sankoff, D. and J.B. Kruskal (1983) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparision*, Addison-Wesley, Reading.

Schaeper, H. (1999) Erwerbsverläufe von Ausbildungsbsolventinnen und –absolventen – eine Anwendung der Optimal Matching Technik, *Sonderforschungsbereich 186 der Universität Bremen, Arbeitspapier*, **57**, Universität Bremen.

Schlich, R. (2001) Analysing intrapersonal variability of travel behaviour using the sequence alignment method, paper presented at the European Transport Research Conference, Cambridge, September 2001.

Schlich, R. and K.W. Axhausen (2003) Habitual travel behaviour - evidence from a six week travel diary, *Transportation* **30** (1) 13-36.

Schmiedel, R. (1984) Bestimmung verhaltensähnlicher Personenkreise für die Verkehrsplanung, Dissertation at the Universität Karlsruhe, Karlsruhe.

Schönfelder, S., R. Schlich, A. König, A. Aschwanden, A. Kaufmann, D. Horisberger and K.W. Axhausen (2002) Mobi*drive*: Data format guide, *Arbeitsberichte Verkehr- und Raumplanung,* **116**, IVT, ETH, Zürich.

Ward, J.H. (1963) Hierarchical grouping to optimize an objective function, *Journal of American Statistical Assocaition*, **58**, 236-244

Wilson, W.C. (2001) Activity pattern of Canadian women: An application of ClustalG sequence alignment software, paper presented at the 80th annual meeting of the Transportation Research Board, Washington, January 2001.

Wilson, W.C. (1998a) Activity pattern analysis by means of sequence alignment methods, *Environment and Planning A*, **30** (6) 1017-1038.

Wilson, W.C. (1998b) Analysis of travel behavior using sequence alignment methods, *Transportation Research Record*, **1645**, 52-59.

Wu, L.L. (2000) Some comments on sequence analysis and optimal matching methods in sociology: Review and prospect, *Social Methods & Research*, **29** (1) 41-64.