XI Congreso de Ingenieria del Transporte (CIT 2014)

Cluster-analytical-creation of a typology of young adults’ travel behavior in Germany

Rico Wittwer*

Technische Universität Dresden (TUD), Faculty of Transportation and Traffic Science, Chair of Transport and Infrastructure Planning
01062 Dresden, Germany

Abstract

During the last decades a solid knowledge base has been created for understanding travel behavior. Different scientific areas like psychology, economics and engineering are researching explanations of human behavior at different levels, with different methods and partially with different measures. Most approaches seek on an individual level, to identify factors that provide information on the variability of travel behavior and help explain these variations. If major factors influencing travel behavior are able to be identified and quantified, planning deficits and chances could be revealed and new policies could be designed. In light of the different stages in life, for instance impending or completed professional education, moving to a new city, into a shared or own flat, starting (one's own) family, reorientation to a routine working life among others, young adults are intuitively a very heterogeneous group. The methodological focus of the paper to be presented is on the creation of a typology of young adults’ travel behavior using multivariate statistical methods. The main applied data source for this work is the German Mobility Panel. The sample was generated from young adults between the ages of 18 and 35 who were selected out of first participants in one panel wave between 2002 and 2007. As a result, a cluster-analytical typology on young adults based on travel behavior will be presented. This typology contains six different groups, which can be described very clearly through socio-demographics and variables of land-use.

© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/). Peer-review under responsibility of CIT 2014.

Keywords: Typology; young adults; factor analysis; cluster analysis; discriminant analysis; logistic regression; household survey; travel behavior

* Rico Wittwer. Tel.: +49-351-463-34132; fax: +49-351-463- 372 64.
E-mail address: rico.wittwer@tu-dresden.de

1. Introduction

1.1. Starting point and objective

Since the end of the 1990s, researchers in Germany have concentrated increasing attention on the mobility of young adults. Trend developments identified in current studies of young adults’ travel behaviour, for example studies on the basis of data of the German Mobility Panel (MOP), are presented in Kuhnimhof et al. (2012) and Kuhnimhof, Wirtz, & Manz (2012), among others. In the Anglo-American academic community, similar mention can be made of the recent work relating to young adults by Blumenberg et al. (2012), Davis, Dutzik, & Baxandall (2012) and Mans et al. (2012). One factor common to the American studies is their background: According to the results of National Household Travel Surveys (NHTS), there was a slight decline in the overall mobility of young adults in the USA over the period from 2001 to 2009, but at the same time a remarkable decrease in the distances travelled (16–34 age group: −23 %). Knowledge of the causes and impact of the behaviour displayed by the population is a basic prerequisite for the elaboration of target-group-specific measures to influence behaviour. The methodological focus of the present paper is to be placed on the creation of a typology of young adults’ travel behaviour using multivariate statistical methods. A multi-stage approach is intended to separate statistical problems into independent subtasks and thus to be able to offer other researchers facing similar issues adequate solutions to statistics-related problem aspects.

1.2. Study subject and definitions

The subject of the present paper is the everyday travel behaviour of ‘young adults’. This is understood to refer to those mobility decisions which arise from the organisational needs and routines of daily life, periodically recurring out-of-home activities and other single circumstances which, although a periodical, nevertheless occur within the course of a week, but not to day excursions or multi-day journeys on non-everyday occasions.

Both national and international studies usually define the target group by way of biological age limits, though it is intuitively apparent that this is not an unconditionally suitable means to distinguish life phases. The lower age limit is here 18 years (legal age of majority). Various authors from different disciplines have specified 34 years as the upper age limit (Khattab & Fenton, 2009; Kuhnimhof, Wirtz & Manz, 2012; Davis, Dutzik, & Baxandall, 2012). There are others, however, who go beyond this age. For the present study, an upper age limit of 35 years has been chosen.

1.3. Theoretical background

Many authors date the start of modern travel behaviour research to the late 1960s or early 1970s (Hensher & Button, 2000; Scheiner, 2006). The work of Kutter (1972), in particular, heralded intensive studies in activity space research to investigate the explanatory power of socio-demographic factors with regard to travel demand. Foundations can be identified above all in time geography (Hägerstrand, 1970). Over the years, the concept of action spaces was adopted by the primarily engineering-oriented transport research community, and designations such as “activity-based approach” are today essentially standard terminology (Scheiner, 1998). The research approaches display considerable diversity and substantial variation in terms of modelling form, precision and depth. Overviews of different facets in activity-based analysis of space-time behaviour are given by Axhausen & Gärling (1992), Kitamura (1997), McNally (2000), Timmermans, Arentze, & Joh (2002) or Schönfelder & Axhausen (2010).

A paradigm shift towards focus on the individual – in other words the active travel participant – and his characteristics, constraints and decision backgrounds led in some respects to a wholly new perspective. This step was based firstly on realisation that out-of-home activities were the original cause for the need to change location and that the individual was thus the decisive reference variable from which travel or travel demand – as the effect – could be derived (Simma & Axhausen, 2001). The core idea of the activity-based approach is to view travel behaviour as a single overarching construct wherein sequences of travel-related actions are combined inseparably in accordance with their temporal, spatial and factual correlations and restrictions (McNally, 2000). It is undisputed that pattern-oriented (activity-based) approaches achieve the greatest modelling precision, albeit at the expense of
exponentially increasing complexity. The explanation that attitudes, as psychosocial factors, represent a key component, and consequently their incorporation as predictors improves the quality of explanatory models, is demonstrated repeatedly in studies. Many empirical studies of travel behaviour support this conceptual framework or else build upon it. Over the years, models such as the norm activation model (NAM) according to Schwartz (1977) and the theory of planned behaviour (TPB) according to Ajzen (1991) have emerged and are acknowledged not only among social psychologists.

1.4. Sample and population

The data base comprises survey data from the ‘German Mobility Panel – MOP’. The reasons for this decision were (1) the access to more complex information on mobility, as data are collected over a period of a full week, (2) the resulting statistical benefits from the more favourable variable properties (distribution properties), and (3) the fact that such data are representative for the whole country. When determining the sample, it must be ensured that the requirement of independent residuals, as a prerequisite for a parametric method, remains satisfied. The selection was thus limited not only by age (18–35 years), but also by the year of first MOP participation. In view of the relatively small panel sample, it was from the beginning planned to merge several years. The number of years had to remain relatively small, however, so as to minimise the risk of blurring possible effects of behaviour change over time. In the end, first-time participants from the years 2002 to 2007 were extracted as the sample.

The MOP is effectively a quota sample intended to model the whole population. Since no specific group characteristics of young adults serve as sampling control criteria, there is little to contradict the assumption of random selection with regard to the persons in this group. For the present study, therefore, it could be assumed that first-time participants between 18 and 35 years from the panel years 2002–2007 were included in the sample by chance and that the group can be idealised as an unrestricted random sample. The summarising of seven MOP survey days into person-weeks and the exclusive consideration of the year of first participation bring several mathematical-statistical benefits: (1) The independence of the residuals (and consequently of the measured data) remains guaranteed; (2) The variable properties, in particular distribution properties, are more favourable for the use of parametric methods; (3) Longitudinal information (e.g. to model routines and habits) can still be utilised by way of additional behaviour indicators; (4) The use of panel data as a pseudo cross-section avoids the statistical problems associated with longitudinal surveys.

The ensuing analysis data set contains aggregated mobility information for 932 young adults for a full survey week, along with attributes describing their socio-demographic situation and the regional-structural settings in which they live. The representative nature of the attributes relative to the overall population was tested and only minor deviations were found.

2. Methodology

The decisive difference compared to studies to characterise persons lies above all in the modelling depth of complex space-time behaviour. Space-time behaviour can generally be described by (1) individual behaviour indicators, (2) a set of behaviour indicators, and (3) behaviour patterns (sequences of activities). The type of survey question determines the demands regarding the modelling depth. If the survey specifically targets cause-effect correlations, i.e. the effect of one or more predictor on an analysis parameter (e.g. the probability of selecting a particular transport mode alternative), then it is common to observe single parameters or a set of space-time behaviour indicators. This extraction is naturally accompanied by a loss of specific pattern information. For a travel-related sociological analysis of fundamental differences in the behaviour of ‘young adults’ and their causes, a set of indicators is deemed to represent an adequately precise abstraction of behaviour. Anable (2005) presents a meaningful procedure for the segmentation of travel behaviour, namely a multi-stage method based on a set of behaviour indicators. A similar approach is also favoured for the present study, as it permits the separation and subsequently the adequate processing of individual statistical problems. Basically, the creation of a typology is an exploratory question.
Fig. 1 (a) shows the schematic process for cluster-analytical determination of a behaviour typology. The steps of the classification process can be characterised as follows. After defining the scope of analysis, the second step is a priori selection of a set of objects and variables as a basis for grouping. Thereafter, outliers must be diagnosed and the conditions for application of the intended parametric methods must be tested. If certain conditions are not met, it is possible to reduce the deviations from the desired variable properties by way of transformation. Even a theoretically founded selection of indicators is not free of the risk of unwanted self-weighting in individual dimensions of behaviour. Therefore, factor analysis methods are used to identify and estimate essentially independent (latent) factors. Before the actual classification, the fusion algorithm and a measure of similarity or dissimilarity are selected. After performing the cluster analysis, it is next necessary to determine a meaningful number of clusters and to describe the content of each cluster. If meaningful interpretation of the clustering solution is possible, further tests must be performed. This refers above all to the validity and reliability of the solution. To this end, an attempt is undertaken to model the exploratory clustering solution on the basis of confirmatory methods (validation). If this model test is successful, the stability of the clustering solution can be verified by applying further methods and/or measures of similarity or dissimilarity (reliability). The clustering solution is formally valid if the internal homogeneity of the groups is greater than in the original data and the clusters display a high degree of heterogeneity (Everitt et al., 2011). The final step should be a content validation test, which is intended to provide a discriminating description of the clustering solution by way of further variables (socio-demographic and regional-structural factors).

3. Creation of a typology of Young adults’ travel behavior

3.1. Set of indicators, diagnostic of outliers, transformation

In the author's opinion, there is still today no generally recognised and practically applicable method available for the calculation of behaviour similarities. Therefore, travel behaviour is to be broken down into individual constituent aspects within the framework of an action-based approach. This serves essentially to permit simplifying abstraction of the highly complex construct. Damm (1983) describes indicators which can here be taken as an initial basis. The indicator set was expanded for the present study and at first comprised 31 variables: Type of activity (journeys per week by purpose – 7 indicators), activity timing (journeys by days of the week and time of the day – 5 indicators), activity duration (mean duration per journey – 1 indicator), activity location (distance per journey by purpose – 7 indicators), mode choice (journeys per week per transport mode – 4 indicators), inherent constraints (journeys per week as passenger, number of days without travel, chained activities per tour by days of the week – 4 indicators), routines/habits (dichotomous variables on monomodality: Bicycle, public transport, car/motorcycle and situational circumstances influencing habits: Illness, vacation, car in repair shop, other peculiarities – 7 indicators).

One important step in the preparation of a variable set is the determination of outliers. The latter influence the robustness of the methods and statistics used to a considerable degree (Field, 2009).
Visual data inspection (histograms, box plots) permitted univariate identification, while hierarchical cluster analysis (single-linkage method) served as an instrument for multivariate outlier diagnosis. A total of 27 cases (2.9%) were identified and removed from the data. It can be assumed that the exclusion of these cases has no negative effect (selectivity) on the representative nature of the analysis sample. It cannot be ignored that the indicators in part violate essential prerequisites for the use of parametric methods (normal distribution of the residuals and homogeneity of variance). Further conditions such as interval-scaled properties and independence of the residuals, on the other hand, are adequately satisfied. Scale transformations (Log10(x+1), SQRT(x), 1/(x+1), x_max–x) often achieve homoscedasticity and/or straighten skew distributions into normal distributions and avoid negative correlations (Field, 2009). Where Cronbach’s alpha, as the measure of reliability, was only $\alpha = 0.455$ before transformation, it was thereafter increased to $\alpha = 0.716$. A value from 0.7 is generally considered to represent acceptable scale reliability. The very homogeneous distribution of the seven dichotomous indicators for mono/multimodality and situational circumstances resulted in reduced scale reliability. These variables were thus excluded, leaving 28 variables for the further analysis steps.

3.2. Dimension reduction: Prior factor analysis

The prior application of principal component analysis (PCA) reduces the risk of systematic, selective self-weighting from the a priori variable definition. A further benefit of the factor analysis approach is that the regression-based estimation of latent factors (factor scores) minimises the influence of any surviving outliers. The Pearson's correlation matrix reveals medium and strong correlations, which are an argument for PCA. Furthermore, these correlations lend sufficient support to the assumption of linear correlations. MacCallum et al. (1999) state that, under particularly unfavourable conditions, samples of considerably more than 500 cases are to be recommended. This recommendation is also met by the data base used. Possible criteria for the extraction of a suitable number of factors are Kaiser’s criterion (eigenvalue greater than 1), a scree plot (best number of factors = elbow–1), the explained cumulative variance of the factors (in the human sciences, a value of 50–60% is often already considered reasonable) and so-called parallel analysis (Williams, Onsman, & Brown, 2010). For larger samples (>500 cases) and overall relatively high communalities (all above 0.7 or mean value >0.6), Kaiser’s criterion is an adequate indicator for factor extraction. Application of Kaiser’s criterion identified eight factors which already explained almost 70% of the variable variance, and could subsequently be extracted and inspected with IBM SPSS 20. Three variables displayed a communality of less than 0.4, which means that either these parameters were not connected with the others, or else an additional factor needs to be considered (Costello & Osborne, 2005). The three variables were excluded and the PCA calculated anew. The determined factors were now able to explain almost 75% of the variance in the unrotated data. The mean communality was 0.744. When the so-called Kaiser-Meyer-Olkin criterion is calculated, the determined value exceeds 0.5 (KMO = 0.577) and thus lies in an acceptable range (Williams, Onsman, & Brown, 2010). The anti-image matrix displays almost exclusively values close to zero. Bartlett’s test, as the third quality criterion, returns a highly significant result ($p < 0.0001$, chi² (300) = 14.994). An oblique method was used for rotation. The advantages of this method are the better distribution of the factor loadings between the factors (Field, 2009; Williams, Onsman, & Brown, 2010) and also the fact that a certain mutual dependence seems realistic when considering the behaviour dimensions. With orthogonal rotation, correlations between the factors would be excluded a priori. Given the determined simple structure, the eight extracted factors can be interpreted without major difficulties. The latent factors are designated: (1) Education needs (obligations) ($\alpha = 0.853$); (2) Need for economic independence ($\alpha = 0.393$); (3) Activity space and time budget ($\alpha = 0.452$); (4) Mode orientation ($\alpha = {-}0.780$); (5) Work obligations ($\alpha = 0.599$); (6) Necessary errands ($\alpha = 0.591$); (7) Leisure preference/in-house recreation ($\alpha = 0.401$); (8) Care obligations ($\alpha = 0.765$). The scale reliability is not particularly high for some factors. On the other hand, values around 0.4 for a few indicators per factor are by all means still acceptable (Zinnbauer & Eberl, 2004). The eight latent factors were subsequently estimated by way of multiple regression and saved. Contrary to other possible estimation methods (Bartlett, Anderson-Rubin), the factors to be determined may correlate with each other, and that was already explicitly permitted through the oblique rotation. The factor values obtained are available in z-standardised form for further analyses.
3.3. Classification: Cluster analysis

The latent construct revealed by the PCA can subsequently be passed on for classification. With 905 persons, the data set is relatively large. When hierarchical methods are applied to larger samples (approx. n > 250), it is potentially problematic that “…in early phases of the merging, ties may arise (two or more cluster pairs display the same similarity) with the ability to exert a strong influence on subsequent results…” (Bacher, Pöge, & Wenzig, 2010, own translation). For the present case, therefore, the following procedure is favoured: (1) Hierarchical cluster analysis to determine the optimum number of clusters; (2) Calculation of the cluster centres for the determined clustering solution; (3) Cluster centre analysis (k-means), taking into account the calculated cluster centres as starting values (initial solution). As k-means clustering at the same time involves Ward’s method and the (squared) Euclidean distance, this combination is also used for calculations in the prior cluster analysis. The optimum number of clusters is subsequently determined by comparing two criteria. Firstly, the so-called dendrogram was considered. The clearest classification was found to be a six-cluster solution. The scree plot (see Fig. 1 (b)) also suggested a solution with six clusters. For the subsequent cluster centre analysis, the k-means method was calculated with the previously calculated cluster centres as the initial solution. A clustering solution is only useful if it permits meaningful interpretation on the basis of the underlying indicators. The next step is therefore content description and cross-variable interpretation of the clusters.

To this end, each cluster is designated and the essential differences between the individual groups (untransformed) are described by way of selected indicators (Table 1). With regard to activity frequency, considerable differences are shown between the derived clusters. While the ‘highly mobile’ group undertakes well over five journeys per person and day, their ‘low-mobile’ counterparts record less than half of this value. The especially low number of weekend journeys in the ‘low-mobile’ group (1.9 per day) is conspicuous. One important distinguishing factor for the ‘low-mobile’ group is the number of days with no journeys whatsoever: They record no out-of-home mobility on an average of 1.6 days per week. Appreciable numbers of journeys serving education purposes are recorded only by the ‘education-oriented’ group.

Table 1. Behaviour-related characterisation of the created typology ‘young adults’ (selected indicators).

<table>
<thead>
<tr>
<th>Indicator</th>
<th>1 (n = 115)</th>
<th>2 (n = 119)</th>
<th>3 (n = 185)</th>
<th>4 (n = 162)</th>
<th>5 (n = 160)</th>
<th>6 (n = 164)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journeys/pers.+day (week)</td>
<td>4.1</td>
<td>3.9</td>
<td>3.0</td>
<td>3.9</td>
<td>2.4</td>
<td>5.3</td>
</tr>
<tr>
<td>Journeys/pers.+day (Mon-Fri)</td>
<td>4.4</td>
<td>4.3</td>
<td>3.2</td>
<td>4.2</td>
<td>2.7</td>
<td>6.0</td>
</tr>
<tr>
<td>Journeys/pers.+day (weekend)</td>
<td>3.2</td>
<td>3.1</td>
<td>2.6</td>
<td>3.2</td>
<td>1.9</td>
<td>3.5</td>
</tr>
<tr>
<td>Days with no journeys</td>
<td>0.34</td>
<td>0.22</td>
<td>0.36</td>
<td>0.25</td>
<td>1.64</td>
<td>0.24</td>
</tr>
<tr>
<td>Work in %</td>
<td>24 %</td>
<td>24 %</td>
<td>42 %</td>
<td>5 %</td>
<td>8 %</td>
<td>12 %</td>
</tr>
<tr>
<td>Education in %</td>
<td>2 %</td>
<td>5 %</td>
<td>0 %</td>
<td>32 %</td>
<td>7 %</td>
<td>1 %</td>
</tr>
<tr>
<td>Shopping in %</td>
<td>17 %</td>
<td>30 %</td>
<td>23 %</td>
<td>17 %</td>
<td>38 %</td>
<td>30 %</td>
</tr>
<tr>
<td>Leisure in %</td>
<td>28 %</td>
<td>35 %</td>
<td>28 %</td>
<td>39 %</td>
<td>36 %</td>
<td>26 %</td>
</tr>
<tr>
<td>Duties in %</td>
<td>4 %</td>
<td>3 %</td>
<td>6 %</td>
<td>6 %</td>
<td>8 %</td>
<td>30 %</td>
</tr>
<tr>
<td>Business in %</td>
<td>25 %</td>
<td>3 %</td>
<td>1 %</td>
<td>1 %</td>
<td>3 %</td>
<td>2 %</td>
</tr>
<tr>
<td>Mean journey duration (min)</td>
<td>29.3</td>
<td>26.0</td>
<td>24.2</td>
<td>26.0</td>
<td>22.9</td>
<td>15.2</td>
</tr>
<tr>
<td>Mean journey distance (km)</td>
<td>20.4</td>
<td>9.9</td>
<td>15.3</td>
<td>13.2</td>
<td>9.0</td>
<td>7.3</td>
</tr>
<tr>
<td>Night-time journeys in %</td>
<td>13 %</td>
<td>12 %</td>
<td>16 %</td>
<td>18 %</td>
<td>11 %</td>
<td>7 %</td>
</tr>
<tr>
<td>Journ./pers.+week as car passenger</td>
<td>3.40</td>
<td>2.44</td>
<td>2.43</td>
<td>5.55</td>
<td>2.88</td>
<td>4.28</td>
</tr>
<tr>
<td>Chaining: Journeys/tour, Mon-Fri</td>
<td>3.04</td>
<td>2.60</td>
<td>2.19</td>
<td>2.32</td>
<td>2.18</td>
<td>2.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resulting cluster designation</th>
<th>Business-oriented optimisers with long distances</th>
<th>Multimodal persons</th>
<th>Working traditionalists</th>
<th>Education-oriented persons with choice constraints</th>
<th>Low-mobile persons without mobility constraints</th>
<th>Highly mobile persons with household obligations</th>
</tr>
</thead>
</table>
The only cluster with a mentionable share of business-related activities is the ‘business-oriented’ group (25 %). Journeys to bring and fetch other persons or in connection with care obligations or general errands (summarised under ‘Duties’) are found above all in the ‘highly mobile’ cluster. Journey durations are twice as long in the ‘business-oriented’ group when compared to the ‘highly mobile’ group. This circumstance is due mainly to the considerable deviation in the mean journey distance between the two groups. The greatest proportion of night-time activities is recorded by the ‘education-oriented’ group, the lowest proportion in the ‘highly mobile’ group. The ‘business-related’ group displays the highest tendency to chain its journeys on weekdays (3.04 activities/tour), followed by the ‘highly mobile’ group (2.72 activities/tour). The chaining of activities is least common in the ‘low-mobile’ group. The strongest affinity to daily private vehicle use is shown in the group of ‘working traditionalists’ (Fig. 2). A very high rate of private vehicle use is also revealed – as to be expected – in the ‘highly mobile’ and ‘business-oriented’ groups. The ‘education-oriented’ group is still strongly focussed on private transport, followed by the ‘low-mobile’ group. Behaviour in the ‘multimodal’ group differs in this respect: On average, a private car or motorcycle is used for only one in five journeys. The shares of public transport and bicycle use are correspondingly high in this group. In both the ‘multimodal’ and ‘low-mobile’ groups, more than one in three journeys are walking journeys. ‘Working traditionalists’, by contrast, only rarely walk. The three clusters with the most intensive use of private transport are also characterised by the highest proportions of monomodal car/motorcycle users. The ‘education-oriented’ group is quite remarkable in this respect. Despite a share of over 60 % private vehicle use over all journeys, well less than one-third of the persons in this group are monomodal car/motorcycle users.

3.4. Model test: Analysis of variance and discriminant analysis

To test the model, the first step was to perform analyses of variance for single factors. The purpose of the analyses of variance was to assess the variables used for cluster definition in respect of their contribution to the resultant grouping. The level of significance was corrected accordingly before interpretation to allow for the multiple testing (Bonferroni correction). All eight factors were shown to be highly significant (df = 5, p < 0.001) for group separation. Subsequently, discriminant analysis was performed to assess the precision of classification in the groups.

For the present study, this is both contentual and confirmatory in character. As use of the same variables could lead to accusations of a somewhat tautological approach, the 25 transformed original variables are used in place of the latent factors. To assess the prediction quality, the proportional criterion put forward by Morrison (1969) was considered. For a six-group case and the resulting group sizes, the random prediction chance is 17.1 %. Cross-validation was performed for the classifications. This is advantageous because full use of the available information enables an undistorted estimation of the prediction chance.

---

Fig. 2. Mode selection by behaviour type.
The first three discriminant functions already account for the majority of the explainable variance (78.9 %) and display relatively high canonical correlations (> 0.75). Wilks’ lambda reveals good, low values (between 0.009 and 0.114) for these three functions – in contrast to the fifth function – and for all functions a high level of significance (p < 0.001). In total, 86.9 % of the cases can be correctly predicted, which represents an extraordinarily good result. The poorest result of all groups was that in Group 6 ‘Highly mobile persons’ with 76.8 % accurate classifications.

3.5. Stability and formal validity: Varying cluster analysis methods and distances, F- and t-values

Testing of the stability of a clustering solution is a necessary prerequisite for its content validity. For test purposes, the k-means solution was compared to Ward’s hierarchical method with use of the squared Euclidean distance (1st comparison). As a further comparison, average-linkage calculations were performed with the same measure of similarity (2nd comparison). Although Ward’s method is usually based on the squared Euclidean distance, the literature nevertheless contains frequent references to the combination of Ward’s method with Pearson’s correlation (3rd comparison). Bacher, Pöge, & Wenzig (2010) state that, in a comparison of different methods, a coincidence of around 70 % between the clustering results can be viewed as adequately stable. The stability test performed for the present study revealed that the classifications coincided in 76 % of the cases in the first comparison, 67 % in the second comparison and 70 % in the third comparison. One essential property of successful clustering is compliance with the requirement of intra-cluster homogeneity or inter-cluster heterogeneity. For the testing of formal validity, it is usual to calculate the so-called F-value. If the F-value remains below 1 for each variable and cluster, the variations of the variables within a cluster are lower than in the entirety of the data (intra-cluster homogeneity). The clustering solution should thus return values of F < 1 for as many combinations as possible. The homogeneity criterion (F-value) is satisfied for the vast majority of the variable-group combinations. Averaged over all variables, an intra-cluster homogeneity between F = 0.51 and F = 0.85 is obtained for the individual clusters. A further formal criterion, primarily as a reference for cluster interpretation, is the so-called t-value; this aspect is not to be discussed in closer detail at this point, however.

4. Socio-demographic and regional-structural characterisation of the behaviour types

4.1. Identification of predictors: Logistic regression

To test the content validity, it is attempted to describe and explain the determined typology by way of socio-demographic and regional-structural attributes. To this end, in the sense of a confirmatory – i.e. structurally-oriented – approach, logistic regression is an expedient method. Compared to discriminant analysis, the logistic regression approach is able to provide significantly more robust estimations, as the method is tied to far fewer (strict) premises. The logistic regression model presented below was created on the basis of objectively logical contentual considerations. When creating the model, it is on the one hand desirable to realise the highest possible explanatory power, but at the same time also to use explanatory variables as sparingly as possible. Against this background, different models were tested both as main effect models and with consideration of interaction terms, but were found to be less suitable on the basis of the applied information criterion (AIC). Table 2 shows furthermore the overall adjustment of the model and the corresponding pseudo-R\(^2\) statistics. The model is found to be highly significant. According to Nagelkerke’s R\(^2\), for example, 74.1 % of the variance in respect of group allocation can be attributed to the eleven variables used by the model. The result indicates that socio-demographic factors already play a major role in explaining variances. Due to the mutual dependencies between socio-demographic and regional-structural factors, however, no mentionable improvement of the model was achieved by taking into account both descriptor ranges. The only indicator which was kept was that of an integrated residential location, which is not significant for the model as a whole, but is at least the one regional-structural factor which distinguishes the ‘multimodal’ group significantly from the remaining groups in a group-referenced comparison. To assess the overall quality of the model, finally, the classification matrix was analysed. This confirms – in the same way as the classification precision previously demonstrated by way of discriminant analysis – the cross-validated prediction accuracy of the model. Overall, 57.8 % of the cases are correctly allocated, which is significantly higher than the likelihood to be expected from random prediction (17.1 %).
Table 2. Likelihood ratio tests of the logistic regression.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>AIC of the reduced model</th>
<th>–2 log-likelihood for the reduced model</th>
<th>Chi²</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1610.464</td>
<td>1430.464</td>
<td>0</td>
<td>0</td>
<td>.</td>
</tr>
<tr>
<td>Number of cars in household</td>
<td>1640.170</td>
<td>1470.170</td>
<td>39.706</td>
<td>5</td>
<td>.000***</td>
</tr>
<tr>
<td>Gender</td>
<td>1613.153</td>
<td>1443.153</td>
<td>12.689</td>
<td>5</td>
<td>.026*</td>
</tr>
<tr>
<td>Age, in 3 groups</td>
<td>1618.559</td>
<td>1458.559</td>
<td>28.094</td>
<td>10</td>
<td>.002**</td>
</tr>
<tr>
<td>Advanced school education</td>
<td>1625.762</td>
<td>1455.762</td>
<td>25.298</td>
<td>5</td>
<td>.000***</td>
</tr>
<tr>
<td>Car availability</td>
<td>1629.378</td>
<td>1469.378</td>
<td>38.913</td>
<td>10</td>
<td>.000***</td>
</tr>
<tr>
<td>Season ticket for public transport</td>
<td>1645.306</td>
<td>1475.306</td>
<td>44.842</td>
<td>5</td>
<td>.000***</td>
</tr>
<tr>
<td>Employment, in 4 groups</td>
<td>1764.439</td>
<td>1614.439</td>
<td>183.975</td>
<td>15</td>
<td>.000***</td>
</tr>
<tr>
<td>Size of household, in 3 groups</td>
<td>1612.377</td>
<td>1452.377</td>
<td>21.912</td>
<td>10</td>
<td>.016*</td>
</tr>
<tr>
<td>Children &lt; 10 years in household</td>
<td>1633.562</td>
<td>1473.562</td>
<td>43.098</td>
<td>10</td>
<td>.000***</td>
</tr>
<tr>
<td>Use of car-sharing</td>
<td>1618.073</td>
<td>1448.073</td>
<td>17.609</td>
<td>5</td>
<td>.003**</td>
</tr>
<tr>
<td>Integrated residential location</td>
<td>1608.872</td>
<td>1438.872</td>
<td>8.408</td>
<td>5</td>
<td>.135</td>
</tr>
</tbody>
</table>

Note: R² = 0.719 (Cox & Snell), R² = 0.741 (Nagelkerke), R² = 0.358 (McFadden), Model: Chi² (85) = 958.1***, n = 754 (valid) cases, - p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001, * Shopping facility, bar, cinema and sports facility in immediate vicinity of home (max. 1–2 km distance or 15- to 20-minute walk)

4.2. Description

The most conspicuous findings for each of the six groups are summarised in the following: (1) **Business-oriented optimisers with long distances**: Predominantly male, majority aged 31–35 years, mainly in full-time employment, no car-sharing; (2) **Multimodal persons**: Majority without a car, high level of education, predominantly in possession of a public transport season ticket, frequently single-person household, rarely children, significant – but small – proportion of car-sharing, tendency to integrated residential locations (76.5 % in cities > 100,000 inhabitants); (3) **Working traditionalists**: Car available, less commonly advanced school education, seldom public transport season tickets, high proportion of full-time employment, less often integrated residential locations (58.4 % rural locations and small towns < 20,000 inhabitants); (4) **Education-orientated persons with choice constraints**: Predominantly very young (18–25 years), frequently in possession of a public transport season ticket, almost all in education, no children in the household, practically no car-sharing; (5) **Low-mobile persons without mobility constraints**: Majority female, predominantly not working or in education, but only few journeys for education purposes, partly integrated residential locations (66.2 % in towns and cities > 20,000 inhabitants); (6) **Highly mobile persons with household obligations**: Predominantly female, majority over 30 years of age, highest car availability, predominantly part-time employment or not working, large households, large majority with children under 10 years in the household, moderate frequency of integrated residential locations (only 24.4 % in cities > 100,000 inhabitants).

5. Conclusion and outlook

The present paper describes the behaviour-oriented creation of a typology of young adults on the basis of data collected by the German Mobility Panel. The determined groups display very typical aspects of behaviour and are significantly distinct from each other. The typology is statistically stable. The particularly high classification precision demonstrated by the discriminant analysis and the good reproducibility of the groups when determined on the basis of other methods and measures of similarity are evidence for a high degree of validity and reliability of the typology. Eleven essential influencing variables were used for the modelling of socio-demographic and regional-structural factors. These variables permit a very good substantial description of the typology, which is thus suitable as a basis for further application-oriented operationalisation.
Acknowledgements

This paper presents the results of own research conducted by the author at the Chair of Transport and Infrastructure Planning at Dresden University of Technology within the framework of an habilitation project. Responsibility for the contents remains solely with the author.

References


Blumenberg, E., Tailor, B. D., Smart, M., Ralph, K., Wander, M., & Brumbagh, S. (2012). What's Youth Got to Do with It? Exploring the Travel Behavior of Teens and Young Adults. Faculty Research, University of California Transportation Center, UC Berkeley, 138 p.


