Modelling travel behaviour using pseudo panel data – first results

Claude Weis, IVT, ETH Zürich

Conference paper STRC 2008

STRC

8th Swiss Transport Research Conference
Monte Verità / Ascona, October 15-17, 2008

Modelling travel behaviour using pseudo panel data – first results

Claude Weis
IVT, ETH Zürich
Zurich

Phone: +41446333952
Fax: +41446331057
email: weis@ivt.baug.ethz.ch

October 2008

Abstract

The paper deals with the testing of different assumptions about the dynamics of transport demand. The general hypothesis to be tested is that travel behaviour reacts to changes in generalized costs for activity, resp. travel, participation. Individuals can adapt their travel behaviour on several levels.

Given the large number of existing studies dealing with the destination, mode and route choice dimensions, the analysis focuses on the demand generation side, i.e. the demand for out-of-home activities and number of trips undertaken, as well as their durations.

The data used for the study are the Swiss Microcensus datasets from 1974 to 2005, which were enriched with generalized travel cost data from various sources, and a detailed database of Swiss municipalities since 1950, including spatial, welfare and accessibility data.

Modelling was done using a *pseudo panel* approach, i.e. the individuals from the surveys were aggregated into cohorts with an assumed fixed membership over time. As much variance as possible was maintained when creating the pseudo panel datasets, meaning that the analysis units were chosen on as much a disaggregate level as possible. The first models test the hypotheses separately, i.e. using univariate regression models.

These first models form the basis for the formulation of a *structural equations* model, which tests all the hypotheses simultaneously for all dimensions. It provides the parameters for the individual dimensions, which are corrected for the influence of the other variables, as well as the corresponding error covariances.

Keywords

Travel behaviour modelling; trip generation; pseudo panel data; structural equations modelling

1. Introduction and motivation

The objective of the work, which is embedded in a current research project, is to test various hypotheses of how individuals' travel behaviour reacts to changes in generalized costs of participating in activities. Individuals can adapt their travel behaviour on several levels:

- the decision to leave home, i.e. to participate in out-of-home activities;
- the number and duration of out-of-home activities;
- the combination of out-of-home activities and trips into trip chains, or tours (successions of trips starting and ending at home);
- the scheduling of activities;
- the choice of locations for carrying out the activities (destination choice);
- the choice of an origin-destination connection (mode and route choice).

Given the large number of existing studies dealing with the latter two dimensions (which effectively constitute the second to fourth steps in the classic model), the analysis focuses on the upper levels, i.e. on the demand generation side. Here, the hypotheses to be tested are that, as a reaction to a reduction in generalized costs:

- the share of days with out-of-home activities will increase;
- the number and duration of out-of-home activities will increase;
- the demand for transport services (distances travelled and trip durations) will increase;
- the number of trips per tour will decrease, as the return to the home location becomes cheaper (in terms of generalized costs).

The observed effects are expected to be non-linear and exhibit hysteresis, i.e. there will be a time lag between the cause (the changes in generalized costs) and the effects. The applied modelling framework is able to capture such lagged effects.

The work draws on a number of existing and new data sources and yields information about long-term trends in transport demand as a function of structural changes of population, welfare and generalized costs of activity participation.

Modelling is done using a *pseudo panel* approach, i.e. the individuals from the surveys are aggregated into cohorts with an assumed fixed membership over time. The averages within the cohorts are then treated as individual observations in a panel. As much variance as possible was maintained when creating the pseudo panel datasets, meaning that the analysis units were chosen on as much a disaggregate level as possible. However, care had to be taken not to push the disaggregation too far, so that the number of observations constituting each cohort was sufficiently large to provide unbiased estimates of the cohort averages. A series of basic models test the above hypotheses separately, i.e. using a univariate general linear modelling framework.

These first models will form the basis for the formulation of a *structural equations* model, which tests all the hypotheses simultaneously for all dimensions. It will provide the parameters for the individual dimensions, which are corrected for the influence of the other variables, as well as the corresponding error covariances.

The paper is structured as follows. The next section describes the construction of the pseudo panel dataset and the variables it contains. The subsequent section is an overview of the descriptive characteristics of the pseudo panel cohorts and their variation over time. The model formulation and estimation steps will then be described, followed by a brief conclusion and an outlook on upcoming work.

2. Construction of the pseudo panel dataset

The concept of pseudo panel data was first introduced by Deaton (1985). It is based on grouping individuals from cross sectional observations into cohorts, the averages of which are then treated as individual observations in a panel. These data can be used in the absence of "real" panel data to simulate the following up of virtual persons (created by the aggregation into cohorts; Mason and Wolfinger, 2004) over long time periods and test for generation membership effects. Examples for the application of the method in the transport planning field are Bush (2003), a study aiming to forecast future travel demand of older adults; Dargay (2002, 2007) and Huang (2007), where the substantial influence of cohort effects on household car ownership is modelled.

The pseudo panel dataset was constructed using the Swiss Microcensus (the Swiss equivalent to the U.S. National Household Travel Survey) data. The survey has been carried out approximately every 5 years since 1974. Over the course of time, the survey methods have been varied several times, which made the comparison of the resulting data somewhat difficult. A brief overview of the survey methods used is given in Table 1 below (Simma, 2003).

Table 1 Key data of Swiss Microcensus surveys since 1974

Year	Survey method	Sample size
1974	Time use surveys, combination of pen-and-paper and personal interview	2'114 households
1979		2'000 households
1984	Trip based diary, pen-and-paper survey	3'513 households
1989		20'472 households
1994		16'570 households
2000	Stage based diary, CATI	28'054 households
2005		31'950 households

Source: Simma (2003)

The various household, person and travel datasets were submitted to a thorough reformatting procedure in order to obtain a standardised data format for all persons over the different years and their relevant sociodemographic characteristics and key mobility figures. The unequal survey methods led to certain discrepancies in the data. For example, in the 2000 and 2005 datasets, trips to activities lasting less than 1 hour were aggregated, leading to underestimations of the number of conducted trips as well as overestimations of trip

durations, as the durations of the suppressed activities were added to the aggregated trips. This drawback had to be corrected by recompiling the trip dataset from the underlying stage records.

Furthermore, a severe decrease in reported mobility was discovered for the 1989 dataset. This discrepancy seemed not to be explicable by seasonal fluctuations, but rather related to an underreporting of trips in the corresponding diary. These effects, which were clearly due to methodological issues, were considered and corrected for in the model estimations that will be discussed below.

The cohorts for the pseudo panel dataset ought to be constructed based on characteristics that are (or can reasonably be assumed to be) time invariant. The most obvious example for such a discriminating variable would be the year of birth cohort (which has been used in multiple studies, such as Dargay, 2002 and Huang, 2007). Other criteria, such as gender, education level, or spatial characteristics, can also be considered as cohort grouping variables.

When constructing the pseudo panel, two contrarious aims are important: on the one hand, the cohorts should be chosen in a way that provides a sufficient variability in the data, i.e. the subdivision should be on as detailed a level as possible. On the other hand though, when the disaggregation level is too detailed, the number of observations in certain cohorts and for certain time periods will become small, leading to greater potential errors in estimating the cohort averages and to biased estimates of the population means (Huang, 2007).

As a compromise between sufficient cohort size and level of detail, a cohort subdivision according to four criteria was chosen:

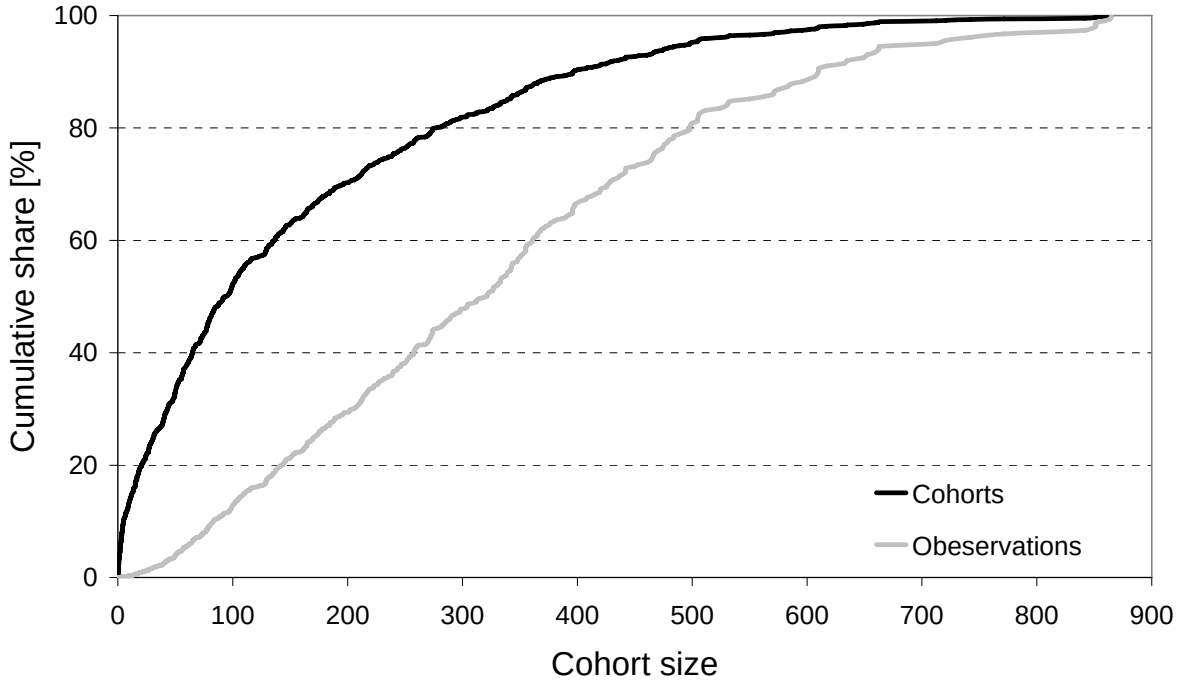
- year of birth (in 10 year bands)
- gender
- spatial region (one out of 7 in Switzerland, equivalent to the EU NUTS 2 regions; see Eurostat, 2008)

The resulting pseudo panel dataset contained 838 virtual observations over the 7 considered time periods. The distribution of the resulting cohort sizes is displayed in Figure 1.

As can be seen, a large portion of the cohorts are quite small (50% of them have a cohort size of 100 or below). However, these small cohorts contain relatively few of the total observations, approximately 85% of the individual observations being in cohorts of sizes above 100 (the size threshold below which the exclusion of the observations from the modelling process is recommended in Huang, 2007). Thus, even weighting by cohort size or

eliminating those cohorts that are under a certain size threshold, a large portion of the underlying observations will still be considered in the analysis, thus leading to reliable modelling results.

Figure 1 Distribution of cohort sizes



Based on the cohort variables, the averages for those variables expected to have an impact on the mobility indicators to be modelled were computed:

- Age
- Household size
- Employment status (as percentage of employed in cohort)
- Monthly household income (in Swiss Francs of 2005)
- Car and motorcycle driving license ownership (as percentage of owners in cohort)
- Number of cars, motorcycles and bicycles in household

The indicators for travel behaviour that are to be modelled based on the independent variables mentioned above are:

- Out-of-home activity (as percentage of mobiles in cohort)
- Number of trips per day
- Number of journeys (sequences of trips departing from and ending at the home location) per day
- Number of trips per journey
- Total duration of out-of-home activities
- Total daily trip duration
- Estimated daily trip distance

Furthermore, the dataset was enriched with several variables that can be used as a proxy for generalized costs of mobility tool ownership, resp. travel:

- Car purchase costs (Frei, 2005)
- Price indices for individual travel (Abay, 2000)
- Fuel costs
- Accessibility measures (Tschopp *et al.*, 2005)
- Average daily travel speeds, computed from travel times and distances

The next section provides an overview of the qualitative characteristics of the pseudo panel cohorts and their variation over time.

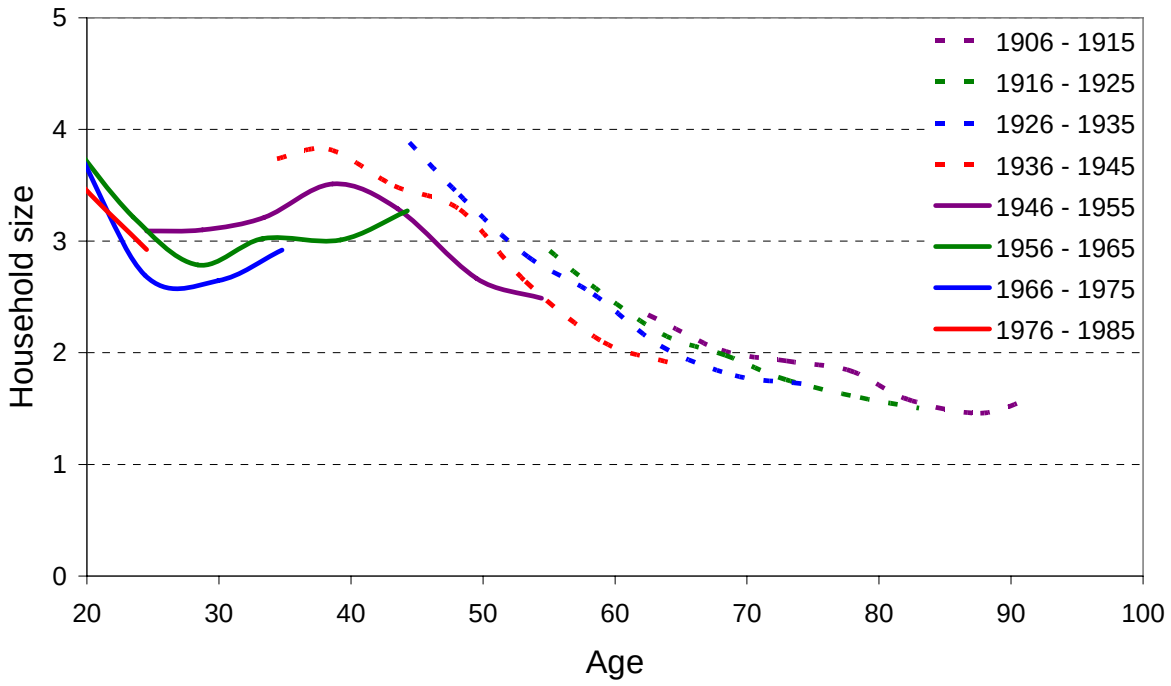
3. Explorative analysis

This section deals with the characteristics of the pseudo panel cohorts and their variation over time, and shows the generation and life cycle effects of the representative indicators as well as the undesired effects of the survey methods on reported mobility that have been mentioned above.

Figure 2 displays the average household sizes for members of the respective year of birth cohorts and their life cycle evolution. Here, both a life cycle and a generation effect can be made out. The life cycle effect for all cohorts shows the expected trends. Young adults tend to live in their parents' homes and thus in large households. As individuals approach their mid 20's, averages household size decreases as a consequence of moving out of the family home and founding own households. Then, after turning 30, the trend again turns to an increase of household sizes, as the individuals settle down and found their own families. As the mid 40's pass, household sizes decrease again as an effect of children moving out, and later on of spouses passing away.

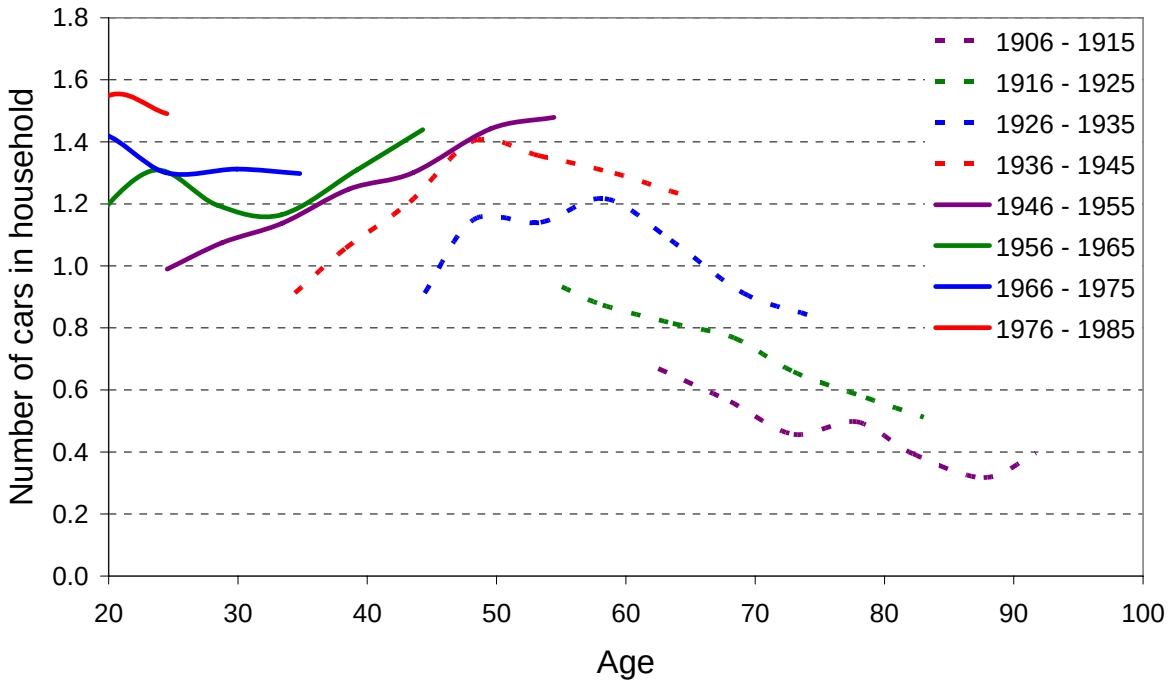
As for the generation effect, it can be seen that younger cohorts tend to live in smaller households. This can be explained by the larger share own single person households (especially for young adults) as well as by the decreasing birth rates. Also, elderly people today tend to live on their own rather than moving back in with their families or committing themselves to nursing homes.

Figure 2 Household size by age for different cohorts



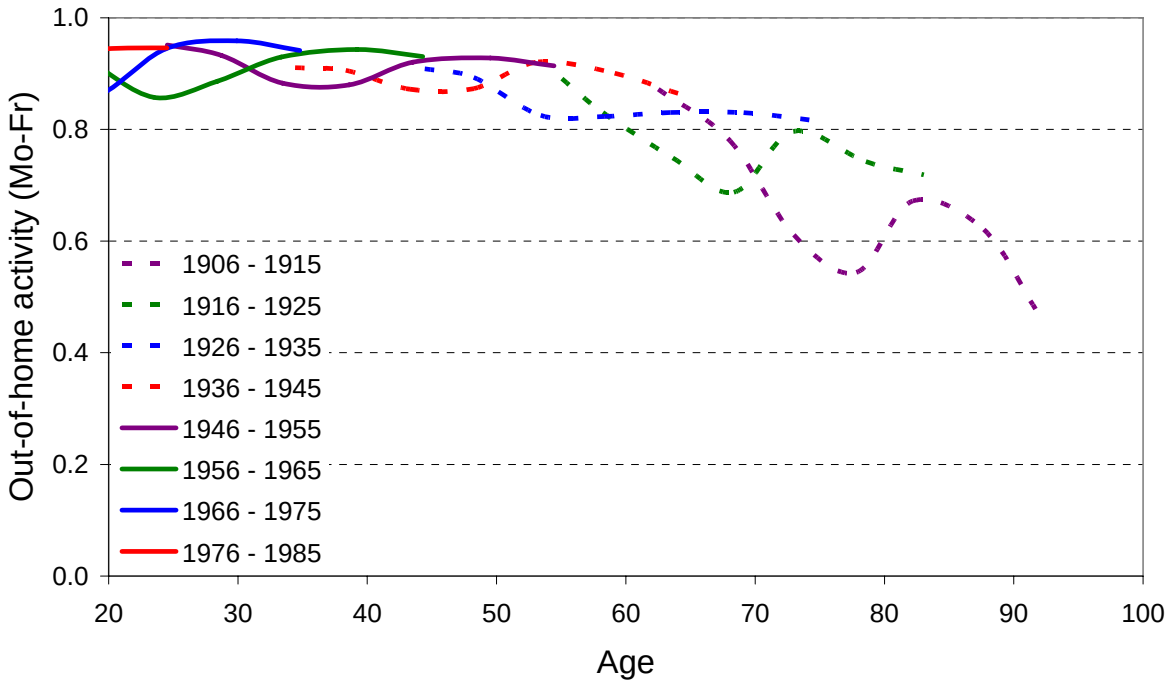
The cohort and age effects for car ownership are depicted in Figure 3. The life cycle effects that are seen here are analogous to those of household size described above. In fact, young adults who live in their parent's home tend to have access to more cars (around the same amount as the middle aged) than those having just moved out from home and founded their own households. Car ownership logically decreases with age. However, this is much less so the case for the younger cohorts, suggesting that these are more mobile and also capable of driving for longer times than the older cohorts. This is a clear indication for a health effect, elderly people today being much more mobile than they were 20 or 30 years ago. Overall, the generation effect clearly tends towards higher car ownership in younger cohorts, again pointing to an increased general availability of the mobility tool over time.

Figure 3 Number of cars per household by age for different cohorts



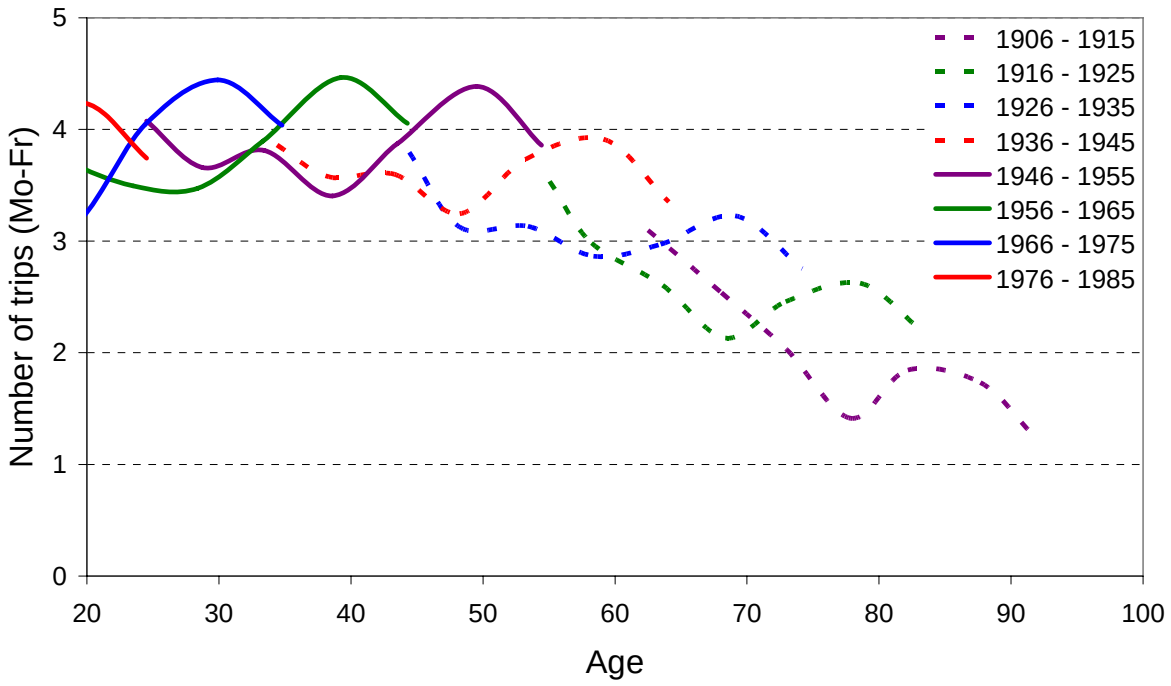
Household size and car ownership, two important characteristics of the household structure and possible indicators for mobility, exhibit the expected and consistent trends over time, over the various age groups and for the different survey periods. The key mobility figures that will be discussed in the following paragraphs and which will form the basis for the models to be estimated later on, unfortunately do not follow the same consistent scheme. In fact, they exhibit significant fluctuations for the various survey periods. As has been discussed above, these most likely are not due to seasonal effects, but are probably artefacts of the employed survey methods. As can be seen in Figure 4, weekday mobility (as a percentage of those individuals that reported at least one trip or out-of-home activity) approximately reproduces the life cycle effect that one would expect, i.e. continuously decreasing mobility with increasing age. However, for each cohort, there is a slight drop in reported mobility around the middle of the curve. Each of these decreases coincides with the 1989 Microcensus survey. No natural reason for this fluctuation being apparent, this suggests that some measurement error must be present in this study.

Figure 4 Percentage of mobiles by age for different cohorts



The same undesired effect of mobility underreporting becomes even clearer through Figure 5, displaying the average number of reported trips by individuals by age and generation. The life cycle effect can still be seen, but the decrease in the 1989 study is ever so apparent.

Figure 5 Number of reported trips by age for different cohorts



The estimated models, part of which will be discussed in the following sections, account for these survey methodology effects and try to flatten them out and reproduce the "real" life cycle and generation effects.

4. Formulation and estimation of the general linear models

In this section, the individual models for the various mobility indicators based on the determining factors listed above will be described. In order to reduce the effect of the biased cohort means described above, weighted least squares (WLS) estimation was applied, the weights being the square roots of the respective cohort sizes (Huang, 2007).

Separate models were estimated for the various indicators mentioned above: out-of-home share, number of journeys, number of trips, duration of out-of-home activities, trip duration and estimated distances travelled. The independent variables used in the models are displayed in Table 2. The modelling framework is a general linear-in-parameters regression model. The general linear model is a generalization of the standard linear regression model allowing the inclusion of categorical variables. It is assumed that the various mobility indicators can be expressed as:

$$y_{i,m} = \mu + \alpha_i + \tau_m + \beta_j \cdot x_j$$

where $y_{i,m}$ are the dependent variables, and μ is a mean intercept term. x_j are the independent variables and β_j the parameters associated to them. α_i are error terms the value of which vary across the behavioural units (the birth year cohorts), yet are invariant over time for any given cohort. τ_m are error terms the value of which varies for the different survey methods, but not over the behavioural units. These error terms serve to cancel out the measurement errors inhibited by the data collection process. The components α_i and τ_m will be treated as constants rather than random variables, leading to cohort specific as well as survey method specific dummy variables incorporated in the linear model. Thus, the model is also called a *fixed effects model* (Kitamura, 2000). Table 2 summarizes the estimation results for the number of trips model, i.e. $y_{i,m}$ is the number of trips for birth year cohort i in a survey period where method m was applied. Parameter values and t statistics are provided. The categories mentioned last in each variable group serve as reference categories.

Table 2 Parameter estimates and model fit for number of trips model

Variable		β	t
Intercept		0.893	21.35
Survey method	Time budget (1974, 1979)	0.009	1.93
	Trip based diary, pen-and-paper / personal (1984, 1989)	-0.068	-24.17
	Stage based diary, CATI (1994, 2000, 2005)		
Year of birth	1896 – 1905	-0.149	-8.32
	1906 – 1915	-0.079	-5.01
	1916 – 1925	-0.011	-0.79
	1926 – 1935	0.025	2.02
	1936 – 1945	0.036	3.44
	1946 – 1955	0.034	3.77
	1956 – 1965	0.030	4.04
	1966 – 1975	0.042	6.52
	1976 – 1985	0.035	6.61
	1986 – 1995		
Gender	Male	0.047	23.36
	Female		
Age	Linear (*1/10)	0.106	11.20
	Squared (*1/100)	-0.009	-15.35
	Natural logarithm	-0.183	-9.79
Household size		0.006	2.88
Employed		0.025	4.98
Car driving license		0.056	9.28
Accessibility private transport		0.002	0.74
Accessibility public transport		0.031	8.67
			Adj. r^2 = 0.763

All variables were found to have a significant effect on cohort level trip generation. The estimated fixed effects for the survey methodologies confirm their above mentioned impact on the dependent variable. The most significant negative effect on trip reporting arises for the

trip based diary surveys in the 1980's, as it is indicated by the graphical representation in Figure 5.

Males throughout generations are slightly more mobile than females. The same holds for employed individuals as well as for car driving license owners, the latter being an indication of a direct effect of mobility tool ownership on reported mobility. Household size has a negative effect on mobility, confirming the general assumption of increased mobility for individuals in single person households.

Perhaps the most interesting effect is observed for the accessibility measures. In the present study, accessibility to population is defined as (see Tschopp *et al.*, 2005):

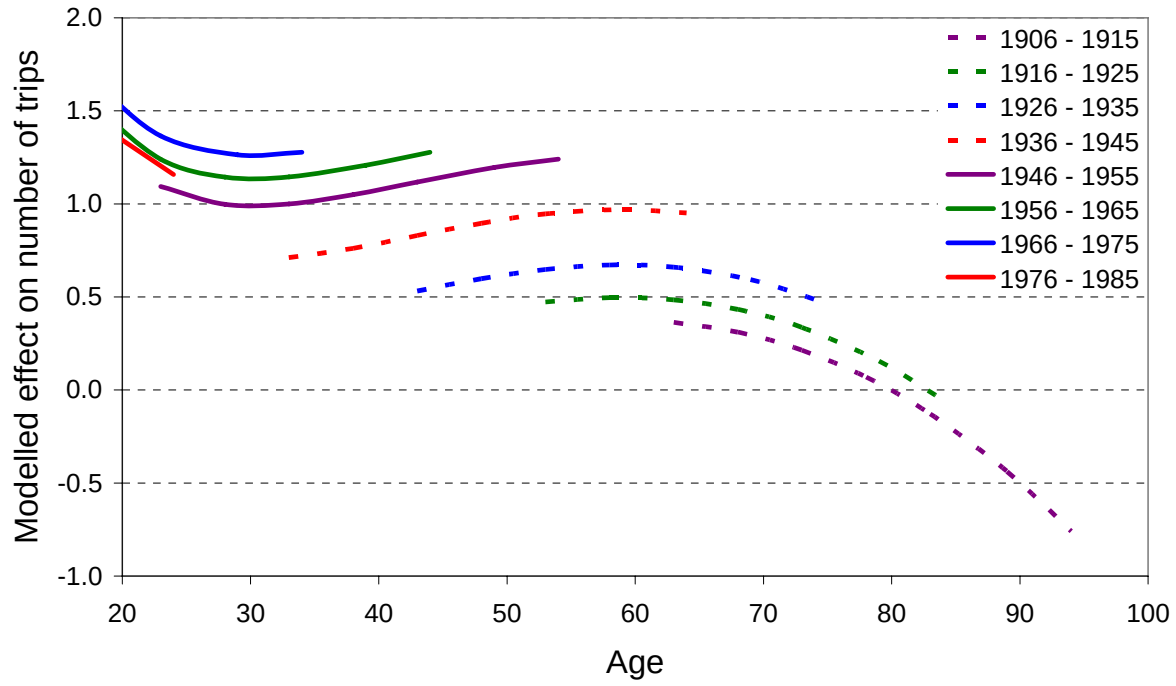
$$A_i = \sum_{j=1}^{n} X_i \cdot f\left(c_{ij}\right)$$

where A_i is the accessibility measure for spatial unit i (the spatial unit here being Swiss municipalities), X_i the number of inhabitants in spatial unit i, c_{ij} the intercentroid travel time from spatial unit i to spatial unit j (n being the total number of municipalities), and f a weighting function. Tschopp *et al.* (2005) use a negative exponential function for weighting, i.e. decreasing accessibility levels for rising travel times.

The accessibility is a proxy for generalized cost of travel and thus a direct indicator for testing the hypotheses that travel behaviour reacts to changes in generalized costs. Interestingly, all other influence factors being accounted for, accessibility still has a significant effect on trip making, suggesting that reductions in generalized costs do indeed infer an increase in travel demand.

As for the generation and life cycle effects (reproduced by the birth year cohort fixed effects as well as the linear, squared and logarithmic age terms), the bare numbers in Table 2 are somewhat difficult to interpret. Therefore, the effect of age on trip making for the different birth year cohorts as it is inferred by the model is graphically represented in Figure 6. As can be seen, the expected real cohort and generation effects are well reproduced by the model – decreasing mobility with age and more mobile recent cohorts.

Figure 6 Modelled effect on number of reported trips by age for different cohorts



Analogous models were estimated for the other mobility indicators; the models yielded the expected results, which will be discussed in more detail in the following section describing the *structural equations* model.

5. Specification of the structural equations model

The formulation and estimation of the basic models described in the previous section yielded the expected effects of the included independent variables on the various mobility indicators (exemplified by the model for the number of weekday trips). Here, a *structural equations* model (SEM) shall be described, which models the effects of the independent (exogenous) variables on the indicators (endogenous variables) simultaneously. Furthermore, the models structure allows accounting for the error correlations between the endogenous variables.

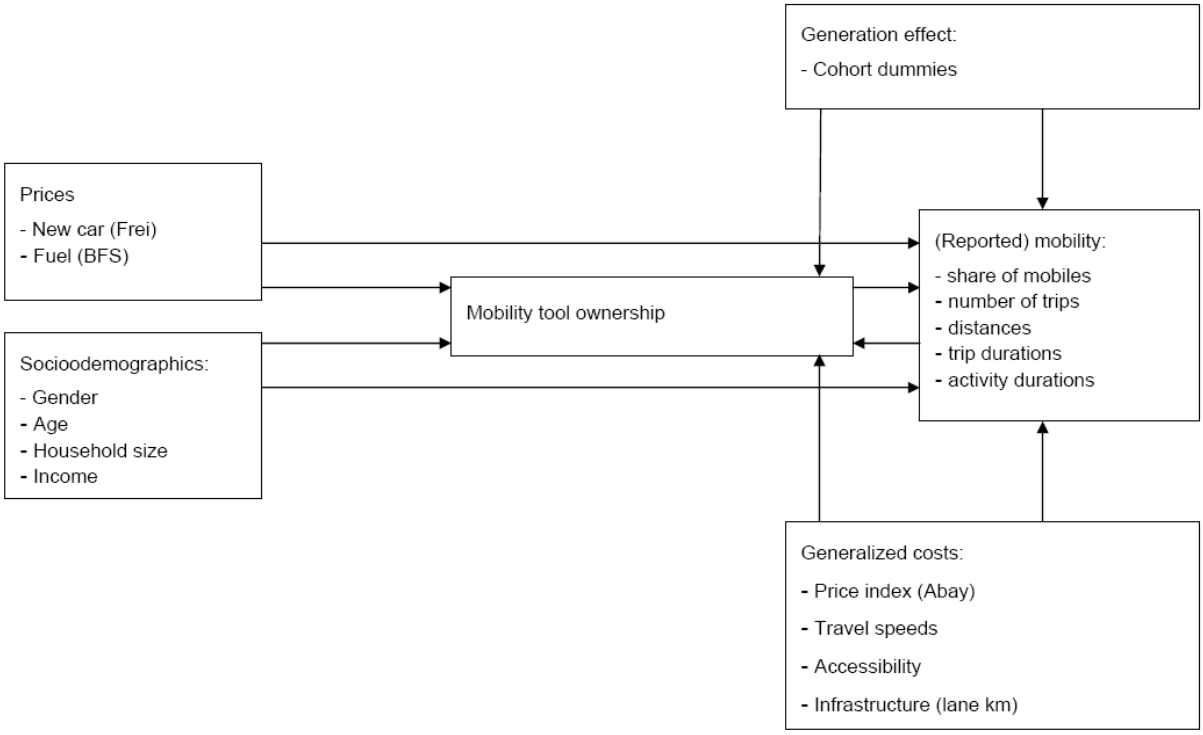
The *structural equations* approach (Bollen, 1989) is a confirmatory method for testing and estimating causal relationships between various factors. The general formulation is as follows:

$$y = \mathrm{B}y + \Gamma x + \zeta$$

where B is an *m x m* coefficient matrix, Γ an *m x n* coefficient matrix, *y* an *m x 1* vector of endogenous variables, *x* an *n x 1* vector of exogenous variables and ξ an *m x 1* vector of errors in the equations. The flowchart in Figure 7 represents the causal effects implied by the basic models. The model assumes no direct causal relationships between the dependent variables, i.e. B is equal to zero. The dependencies between the dependent variables will be captured via error correlations.

It is expected that the SEM will confirm the trends exhibited by the basic models and allow the computation of demand elasticities for all relevant dimensions simultaneously.

Figure 7 Structure of the SEM



6. Conclusion and outlook

The results shown in this paper confirm the original hypotheses. Decreases in generalized costs for travel and activity participation do appear to induce higher individual mobility, as the significant effect of the accessibility measures used in this study on mobility behaviour confirm. The induced travel effect on the upper (i.e., trip generation side) levels of travel demand generation is certainly an interesting finding.

Further work on the topic will try to confirm the trends exhibited by these first results by applying the models to alternative formulations of the generalized costs, e.g. road lane km by type, car purchase and maintenance costs, etc. Furthermore, alternative specifications for the cohorts will be tested for their robustness against the parameter estimates. It may be argued that the modelled accessibility effects on travel behaviour are due to residential self selection, i.e. the more mobile cohorts relocating to places of residence with higher accessibility. Future work will focus on formulating the cohorts in a way to better understand these self selection effects and model them separately from the actual induced demand.

7. References

Abay, G. (2000) Die Preisentwicklung im Personenverkehr 1994-1999, *Report to the Swiss Federal Office for Spatial Development*, Bern.

Bollen, K.A. (1989) *Structural equations with latent variables*, Wiley, New York.

Bush, S.B. (2005) Forecasting 65+ travel: An integration of cohort analysis and travel demand modelling, paper presented at the *84th Annual Meeting of the Transportation Research Board*, January 2005.

Dargay, J.M. (2002) Determinants of car ownership in rural and urban areas: A pseudo-panel analysis, *Transportation Research E*, **38** (5) 351-366.

Dargay, J.M. (2007) The effect of prices and income on car travel in the UK, *Transportation Research A*, **41** (10) 949-960.

Deaton, A. (1985) Panel data from time series of cross-sections, *Journal of Econometrics*, **30** (1) 109-126.

Eurostat (2008) Nomenclature of territorial units for statistics – NUTS statistical regions of Europe, http://ec.europa.eu/eurostat/ramon/nuts/home_regions_en.html, Eurostat, Luxembourg, September 2008.

Frei, A. (2005) Was hätte man 1960 für einen Sharan bezahlt? *MSc Thesis*, Institute for Transport Planning and Systems (IVT), Swiss Federal Institute for Technology, Zurich.

Huang, B. (2005) The Use of Pseudo Panel Data for Forecasting Car Ownership, *Dissertation* University of London, London.

Kitamura, R. (2005) Longitudinal Methods, in Hensher, D.A. and K.J. Button (eds.) *Handbook of Transport Modelling*, 113-129, Elsevier Science, Oxford.

Mason, W.M. and N.H Wolfinger (2001) Cohort analysis, in Smelser, N.J. and P.B. Baltes (eds.) *International Encyclopedia of Social and Behavioral Sciences*, Elsevier, Amsterdam.

Simma, A. (2003) Geschichte des Schweizerischen Mikrozensus zum Verkehrsverhalten. paper presented at the *3rd Swiss Transport Research Conference*, Ascona, February 2003.

Tschopp, M., P. Fröhlich and K.W. Axhausen (2005) Accessibility and Spatial Development in Switzerland During the Last 50 Years, in Levinson, D.M. and K.J. Krizek (eds.) *Access to Destinations*, 361-376, Elsevier, Oxford.