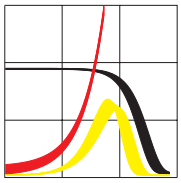


Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Doberaner Strasse 114 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2002-006
FEBRUARY 2002

How important are household demographic characteristics to explain private car use patterns? A multilevel approach to Austrian data

Riccardo Borgoni (borgoni@demogr.mpg.de)
Ulf-Christian Ewert (ewert@econhist.de)
Alexia Fürnkranz-Prskawetz (fuernkranz@demogr.mpg.de)

How important are household demographic characteristics to explain private car use patterns? A multilevel approach to Austrian data

R. Borgoni, U.C. Ewert and A. Prskawetz[1]
Max Planck Institute for Demographic Research
Rostock, Germany

Abstract

Private car use is one of the major contributors to pollution in industrialised countries. It is therefore important to understand the factors that determine the demand for car use. In explaining the variability in car use, it is important to take into account household demographic characteristics and local and regional differences in infrastructure, in addition to the economic variables commonly used in the prevailing literature on the topic. The appropriate tool to explain car ownership and car use is, therefore, a multilevel statistical approach. An Austrian household survey from 1997 finds that household characteristics such as age, gender, education and employment of the household head, household size and housing quality can effect the variability of car ownership and car use. The same survey also gives a clear indication of regional heterogeneity. This heterogeneity persists when we controlled for the variability of regional economic welfare and infrastructure as indicated by population density.

1. Introduction

One of the major polluting activities in industrialised countries is the use of private cars. Cars effect the environment in two adverse ways (Canzler and Knie, 1994, pp. 16-23; Lorbeer, 1996). The first is the intensive use of land demanded by the requisite construction of a traffic infrastructure and the second is air and noise pollution resulting from the actual use of cars.[2] Our demand for spatial mobility and our reliance on the use of private cars to achieve that mobility can thus be defined as environmental misbehaviour, a mode of behaviour with negative consequences for the natural environment.

In Austria an increasing proportion of households use private cars, and during the 1980's they were used to a higher degree in general. The total distance travelled by private cars has increased from 28.3 billion km (with 2.25 million cars) in 1980 to 43.9 billion km (with 3.1 million cars) in 1991. This is an increase in total distance travelled of about 55% and an intensity shift from 12,578 to 14,161 km driven annually per car (Umweltbundesamt, 1998, pp. 100-101, Tables 30, 31).

[1] The authors are grateful for comments and suggestions by Vladimir Shkolnikov and Karsten Hank. Editorial support from Susann Backer is also gratefully acknowledged. The views expressed in this paper are the authors' own views and do not necessarily represent those of the Max Planck Institute for Demographic Research.

[2] Air pollution is produced directly through the emission of carbon-monoxide, carbon-dioxide, sulphur-dioxide, nitrogen-oxides, non-methanol-volatile-organic-compounds and, in the case of diesel fuel, soot. Because these chemical substances are ingredients in a reaction that produces ozone, there also exists an indirect one, namely the production of ozone (cf. Adlmannseder 1993).

To better anticipate the trends in air pollution, as well as better predict the demands for increased infrastructure, it is important to understand the factors that influence the demand for private car use. While most of the research on this topic focuses on economic factors in explaining car use patterns in industrialised countries (e.g. Dahl and Sterner, 1991 provide evidence that private travel demand rises with income), we approach the explanation of car use patterns from a demographic perspective.

In many studies it has been shown that travel patterns differ depending on a person's income. The focus on income, however, obscures other equally important variables. Because income varies between men and women and during the course of the life cycle, it is important to consider the effects of demographic characteristics such as age and gender when seeking an explanation for travel patterns. Focusing on Sweden, Carlsson-Kanyama and Linden (1999) found that women and the elderly – generally persons, who tend to earn a lower income than men – generally travel less than men and the middle-aged – persons who generally earn a higher income than women and the elderly. Similar studies have been conducted for travel patterns and energy use in the US (cf. Pucher et al. 1998, O'Neill and Chen 2001). In a study of Switzerland, Franzen (1987) shows that car ownership and the total distance driven during a one-year period depends on various sociodemographic characteristics. For instance, an OLS regression with distance driven as the dependent variable shows that women and older people spend less time driving, while household income positively influences car distance driven. In order to obtain a more complete representation of the social embeddedness of an individual it is necessary to add further household characteristics like the household's size, the family type and the household's head marital status. All of these factors contribute to the decision of a household to own a private car, or not (Mikl-Horke and Leuker, 1978). These factors also influence how long households intend to hold onto a vehicle (Yamamoto and Kitamura, 2000). The demographic characteristics can also affect the number of cars per household, the car's brand (Krause, 1997; Wellner, 2000), and the specific technical features of the car. Two studies that have focused on the lifecycle concept in explaining the variation in travel demands are Greening and Jeng 1994 and Greening et al. 1997. Several, more recent studies go beyond the explanation of cross-sectional variation in private car use demands and focus on the role of demographic characteristics such as household size, age and sex-specific cohort effects to explain past and future changes of private transportation demands (Buettner and Grubler 1995, O'Neill and Chen 2001, Prskawetz et al. 2001, Spain 1997).

A further issue that seems to be partly neglected in the research on car use is the role of regional variations in geographic, economic and institutional conditions, variations that may influence the pattern of car use in addition to household demographic characteristics. Moreover, demographic characteristics may be regionally clustered. Evidence for regional variations in demographic as well as socio-economic characteristics for Austria is provided in Ewert and Prskawetz (2000,2001).

In order to take into account household level characteristics as well as the regional heterogeneity of car use patterns, we propose a multilevel statistical approach. The main feature of our approach is to distinguish between the relative importance of regional

macro-level variations and variations at the household level. This distinction helps us to better understand the variation in car ownership and car use patterns in Austria. In the cases where we find regional variations in car ownership, or car use patterns, we investigate whether these regional differences can be explained by regional-level variables such as population density, which functions as a proxy for various structural characteristics of regions.

2. Data

2.1 Data Set

The present study is based on the Austrian micro-census of June 1997 (ÖSTAT 1998). The micro-census is a representative household survey of 1% of all Austrian dwellings conducted quarterly. It provides information on household demographic characteristics such as total household size, number of children, age, gender, marital status, education and working status of the household head and housing conditions of the household. The sample size is in the order of 30,000 dwellings, but each quarter an eighth of all addresses are replaced by new addresses. In the particular case of the June 1997 micro-census, the survey contained 22,648 unweighted valid cases (for a more detailed description of the survey see ÖSTAT, 1998, 3-8).

The basic household demographic program of the survey is accompanied by alternating supplementary blocks of questions addressing specific topics of interest. The special program on Energy Use in Households, which was part of the questionnaire in June 1997, provides additional information on the households' private car use during the year preceding the survey, fuel consumption, degree of newness and the technical features of cars (the type of engine, the presence of a catalytic converter).

These data give a very general impression of private households' travel behaviour during a one-year period running from June 1996 to Mai 1997, the month before the interviews were conducted. Unfortunately, these data do not include information on specific activities for which cars were used such that concrete travel patterns cannot be reconstructed. But the possibility of combining high quality information on household characteristics with data on car ownership and car use makes this data set attractive in analysing the interrelation of household characteristics and car use patterns.

In addition to the household level data we also include macro-level data in our analysis. These data are provided for various geographic units by Statistics Austria. In this paper we use the European Union's spatial classification system NUTS (*Nomenclature of territorial units for statistics*). Following the hierarchy of this classification, Austria can either be subdivided into the three regions of eastern, south-eastern and western Austria (NUTS-1), the nine federal states (*Bundesländer*) (NUTS-2), or 35 local areas (NUTS-3). Although it is the NUTS-2 level that is the EU’s official focus for regional planning (Heigl and Mai 1998, p. 294; van der Gaag et al., 2000, p. 2), NUTS-3 areas seem to be the more appropriate units for the analysis because some of the Austrian federal states show considerable internal differences in geography and climate. Furthermore, since

NUTS-3 areas are relatively small – the spatial extent of NUTS-3 areas ranges from 415 to 4,614 square km – their internal heterogeneity as regards living conditions should also be relatively small when compared with the much larger NUTS-2 regions. Our analysis of regional variation is thus restricted to the NUTS-3 spatial classification level. This is done in order to balance the need for regional disaggregation, for reasons of data availability and to maintain conformity of focused regional units to a Europe-wide classification system.[3]

Car ownership and average car use - and the environmental behaviour that goes with these practices (cf. section 2.2) - vary tremendously among NUTS-3 areas. The proportion of car owners is lowest in Vienna with 55.7% of the population owning cars and highest in the Lungau area with 92.3% owning cars. The distances travelled with the first two cars per household during a one-year period in 1996/97 are lowest in the westernmost area of Rheintal-Bodenseegebiet, with 14,250 km travelled per household (calculated on the basis of households with at least one car) and highest, with 21,564 km travelled per household in the area of Südburgenland. Interregional variations of selected socio-demographic variables are clearly evident. In 1997 the average age for heads of households within NUTS-3 areas ranged from 49.41 to 55.20 years old. The average household size in 1997 ranged from 2.02 to 3.49 persons per household and the average educational level, measured on a scale ranging from 0 (no school leaving qualification) to 8 (university degree), ranged from 1.89 to 3.19. Further, regional affluence and regional housing patterns show a large variation: per-capita income as of 1995 ranges from 153,000 ATS (southern Burgenland) up to 433,100 ATS (Vienna). The proportion of households that own the dwelling they are living in ranges from 18% (Vienna) to 89% (Weinviertel).

2.2 Variable definition

We will analyse car ownership and actual car use separately. Together, these different aspects form a latent variable that describes the phenomenon of "car use". Presumably, car ownership and actual car use may be explained by different determinants.[4] Owning a car is generally a long-term decision, having long-term implications for the household. This means that we should expect factors of considerable stability to have the major impact on the decision as to own a car or not. Such "stable" factors could be specific household demographic measures (e.g. education or living arrangement) as well as

[3] The smaller local areas in Austria, called political districts (*Politische Bezirke*), do not meet the last two of the described conditions. Most of the regional-level data cannot be gathered for this type of area, and this specific Austrian classification system does not fit completely into the European NUTS classification scheme.

[4] Our approach to model car ownership and car use separately is not only justified by the fact that both decisions presumably depend on different determinants. It is also based on the fact that our sample includes two different groups of households: those with a car and information on either car use and car technology; and those without a car and hence without any information on car use or the car's technical equipment. A third subsample comprising households declaring car ownership but refusing to provide any information on car use and car technology, is not really a true subsample as is the case with the other two groups. This third group of households can be included in the analysis of car ownership but has to be discarded from the analysis of actual car use.

regional level variables, such as the population density of the region in which the household is located. In contrast, we should not expect actual car use to be determined exclusively by demographic variables. The intensity of actual car use is presumably affected by economic variables (e.g. fuel prices, costs for using alternative means of transportation). Though this set of variables is not included in our data, we anticipate a different underlying model for actual car use and therefore examine both decisions separately (car ownership and car use). Besides demographic variables, we also refer to variables measuring a car's technology and the regional level variables that may influence the decision to purchase a car as well as the actual use of the car.

The dependent variables (car ownership and car use) are defined as follows:

- *Car ownership* is coded as a binary variable, "1" for those households who own a car, "0" for households without a car.[5]

- *Car use* is measured in terms of the distance driven with the first two cars during the one-year period from June 1996 to May 1997.

We include three sets of control variables. The first group of variables consists of car characteristics.

- The *number of cars* adds to the measurement of car ownership. This variable is censored from above at a value of 3 cars per household, meaning that we arrive at four categories (0 = no car; 1 = one car, 2 = two cars; 3 = more than two cars). We have censored the number of cars because detailed information is only available for the first two cars. Note, that only 3.9% of the households in general, and 5.8% of all car-owning households, own more than two cars.

- The cars' degree of newness is a continuous variable that refers to how long the household has already owned the car(s). This measure uses the information for the first two cars of a household. The calculation for each household is as follows:

  $$\textit{Index of Newness} = a_1 p_1 + a_2 p_2$$

  where a_i is "0" (car *i* bought before June 1996 or car *i* not existent), "(13-j)/12" (car *i* bought j=1, ...12 month prior to the interview date) and p_i denotes the share of total distances per household for which car *i* was used during the one-year period. The values of the variable range from "0" (indicating that all cars had been bought before June 1996, i.e. the start of the period under consideration) up to "1" (indicating that all cars had been bought in Mai 1997, the last month of the one-year period).

- The *engine technology* of cars is a continuous variable indicating the use of diesel engines. This measure also uses the information for the first two cars of a household. The calculation for each household is as follows:

[5] In 37 cases households record a positive distance but do not own a car. The most plausible explanation for this answering pattern is that these households used hire-cars. We treat these cases as if they were car owners. This can be done because we want to explain availability of cars to the household – which can be reached either by owning or by renting a car – rather than ownership of cars in a legal sense.

$$Index\ of\ Diesel\ Engines = b_1 p_1 + b_2 p_2$$

where p_i is again the share of total distances per household for which car *i* was used during the one-year period under consideration. The b_i's are dummy-variables with value "0", in case either car *i* is equipped with a petrol engine or car *i* is not existent, and value "1" is used whenever car *i* has a diesel engine. The value of the variable ranges from "0" (indicating that all cars are petrol engine cars) up to "1" (indicating that all cars are diesel engine cars).

- The *presence of environmental technology* is a continuous variable indicating to what degree a household uses cars that are equipped with catalytic converters. This measure again uses the information for the first two cars of a household and is calculated in a similar manner to engine technology:

$$Index\ of\ Catalytic\ Converter = c_1 p_1 + c_2 p_2$$

where p_i is again the share of total distances per household for which car i was used during the one-year period under consideration. The ci's are dummy-variables with value "0", in case either car i is not equipped with a catalytic converter or car i is not existent, and value "1" whenever car i has a catalytic converter. The value of the variable ranges from "0" (indicating that no car has a catalytic converter) up to "1" (indicating that all cars have catalytic converters).

The second group of variables consists of household characteristics. We use the age (years and months in June 1997), gender (female vs. male), nationality (Austrian vs. non-Austrian) and education (measured on a 9-items scale with "0" {no school-leaving certificate} and "8" {university degree}) of the household head.

Concerning the household structure, we use two variables: the size of the household (which is censored from above at the value of 5 persons) and the living arrangement of the household head (which can be derived from information about the marital status and the number of children living in the household). For the latter variable we distinguish between single persons, couples, couples with child(ren) and single parent households. The household size and the living arrangement of the household are closely related to one another (Cramer's V = .668 for all 22.648 cases). This strong relationship causes so-called structural zeros, for instance, it is impossible to observe other living arrangements than single person households that have a household size of one. Moreover, household size is closely correlated to living arrangements, i.e. we may expect households with children to be households of a larger size. For this reason it would not be advisable to use both variables as regressors together in one regression. In order to overcome the interpretation problems resulting from these structural zeros, we constructed a new variable that combines household size and household composition. With regards to households with up to three persons living together, we distinguish among various living arrangements within each household size-category, distinguishing mainly between adult only households and households with children. For larger households we only note the size and not the specific living arrangement. To be specific, the new variable consists of 7 categories: (1) single person (household size = 1 person); (2) "adult only" households, which can be couples or two single persons declaring to live together in one household (household size = 2 persons), (3) a single parent with one child (household size = 2

persons); (4) "adult only" households, for example couples without children living together with a third person, maybe a parent, but also three single persons declaring to live together in one household (household size = 3 persons), (5) couples with one child, or a single parent with two children (household size = 3 persons), (6) all households with size of 4 persons; (7) all households with 5 or more persons. This categorisation is conformed by a classification tree analysis (CART) for the distances driven (cf. Appendix A).

We are also able to measure housing conditions on the household level. Thus, out of a number of variables we have created two continuous measures by applying a factor analysis. The first measure (factor) describes the contrast between rented flats with small usable floor space located in high-rise building (low value) and owner-occupied flats or houses with large usable floor space (high value). The second measure describes the contrast between old (low value) and newly built flats (high value).[6] Originally, both factors were not correlated to each other, but in order to improve the overall fit, both coordinate axes have been rotated in such a way that a non-orthogonal coordinate system was obtained. This results in both factors being correlated to one another (.248, $p \leq 0.01$). Because most of the variance was observed along the factor 1 axis, and as a consequence of the significant correlation between the two factors (after having rotated them), we decided to use only the first of the two measures describing housing conditions of households in our statistical analysis.

The third group of explanatory variables is made up of regional level variables. Out of a potentially larger set of various macro-level variables we have chosen to include only *population density* in our analysis. Population density is calculated as the ratio of persons living in a region in 1991 to the number of square km of populated plots of land in the region. (The figures of populated plots of land for each Austrian NUTS-3 region can be derived from Kautz et al. 1999.) We use population density as the only regional level variable since population density is correlated to several other regional level measures. Population density is a valid indicator for the regional amount of affluence, the structure of the economy, the local labour market (employment opportunities) and the necessity to commute between working and living place. In Austria, regions of high population density are characterised by high per capita income ($r = .617$, $p \leq 0.01$), a predominance of employment in the tertiary sector, and a high net-inflow of daily commuters from NUTS-3 areas (-.429, $p \leq 0.01$).

3. Method

According to the design of the micro-census, dwellings were sampled throughout the federal states of Austria (NUTS-2). Because of this, there is the potential for a kind of

[6] The factor analysis is based on the following variables describing the housing conditions in which the household lives in: (a) number of rooms and (b) usable floor space of the flat, (c) legal ownership condition (rented or owner-occupied flat), (d) number of flats in the building and (e) location of the flat in the building (ground floor, first floor, etc.), (f) sanitary installations in the flat and (g) time period when the building was built. Variables (a)-(e) are correlated with the first factor and variables (f) and (g) are related to the second factor.

clustering to be present in the data – since people living in the same area tend to have similar patterns of car-use, due to the nature of local infrastructure, or the similarity in demographic characteristics within regions. A multilevel approach is therefore appropriate in treating the present data set in a statistically correct manner.

3.1 The multilevel approach

Data sets from several contexts often take on a hierarchical, or clustered structure. In many cases such a structure arises naturally; for instance when the analysis concerns the offspring of different families. Clustering can also be the result of less strongly individual characteristics. Clustering can further be the result of particular sampling schemes (e.g. children belonging to different schools or patients belonging to different clinics or treated by different treatments and so on).

When we deal with social, epidemiological or environmental research questions the existence of hierarchies should be taken into account in a proper way. Quite often these hierarchies mirror different behaviours, or social attitudes. Simply belonging to different groups can modify individual behaviour.

Two obvious examples of a clustered structure are data temporally, or spatially, collected. In the former, clustering is the result of measuring a unit on several discrete occasions (for instance in repeated measure analysis, or panel data), or measuring spells spent in a given state in a continuous time context (like in event history or survival analysis). Clustering can also occur when units are sampled or measured according to some spatial feature, for instance in censuses or some survey data. Spatial clustering is a well known problem in epidemiology (Wakefield et al., 2000), labour market studies (Fahrmair and Lang 2001) and demography (Congdon 2000, Steele et al. 1996).

The multilevel approach provides a convenient framework for studying hierarchically structured data.

In the terminology of multilevel analysis, clustering consists of different, (usually) nested levels. For example, we call political districts units of level 2, while individual items sampled in each of the political districts are units of level 1. Of course we can have more complicated structures. For instance, if each item in a district has been measured more than once over time, each occasion represents a level 1 unit, while items and districts would be level 2 and level 3 units respectively (sometimes we can find a reverse order in naming levels in the literature). A comprehensive overview of the theory and various applications of the multilevel approach are given by Goldstein (1995) and DiPrete and Forristal (1994).

Statistically, these clusters in the data result in a correlation among intra-group observations. Ignoring this phenomenon may render invalid many of the usual statistical techniques. One of the major advantages of multilevel modelling is its ability to take into account the biases in statistical procedures stemming from clustering, i. e. providing correct standard errors for estimates, confidence intervals and significance tests.

Other advantages of a multilevel framework are the proper systematic analysis of the effect of covariates (measured at different levels) on the response variable, and how interactions between different level variables may affect the response. This contrasts with the many analyses that focus on the influence of the macro level on the micro level. The clustered structure of the data naturally defines different components of the variability of a given phenomena in relation to the different levels in the data. Usually, both individual variables (like the age of a unit) and group variables (like the population density of a given area) combine in defining the profile of each item. Sometimes the statistician is interested in understanding and predicting the unobserved, between-level variability. If this is the case, it is the fundamental task of the statistician to determine whether higher-level variables play a role in reducing the unobserved heterogeneity. In our study we set out to determine whether contextual variables at the NUTS-3 level can explain the unobserved variability in regional NUTS-3 levels. If they do explain the unobserved variability, then regional influences need to be taken into account. The most relevant influences would need to be included in the larger model. As we have already indicated in the previous section, the NUTS-3 variables in our data are highly correlated. Because of this we decided to test only for the significance of the contextual variable that measures population density at the NUTS-3 level.

3.2 The specified models

To analyse car ownership we use a *multilevel logit model with random effects* at the NUTS-3 level:

$$\log \frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)} = X_{ij}^{T}\beta + U_i, \qquad (1)$$

where Y_{ij} is a Bernoulli variable with parameter $\Pr(Y_{ij} = 1)$ representing the decision to own ($Y_{ij} = 1$), or respectively not to own ($Y_{ij} = 0$), a car for household j in area i, X_{ij} is a qx1 vector (q-1 is the number of predictors in the model) observed on household j in area i, β is a qx1 vector of parameters (including the intercept) and the U_i's are random effects for the NUTS-3 areas $i = 1,...,35$ and are assumed to be independent and normally distributed with zero mean and unknown variance. As predictors for the decision to run a car we use the set of variables that characterise the demographic structure of the household. We estimate the model for the whole sample including those respondents who indicated car ownership but for whom we lack the information on actual car use, or the car's technical features.

A broad review of multilevel modelling for binary data has been recently given by Guo and Zhao (2000).

To model the actual car use, we use a *random effect linear model* of the following type

$$y_{ij} = X_{ij}^{T}\beta + U_i + \varepsilon_{ij} \qquad (2)$$

where y_{ij} is the logarithmic transformation of the distance for household j in area i (see Appendix B for the motivation to use a logarithmic transformation), X_{ij} is the $p\times1$ vector

(p-1 is the number of predictors in the model) observed on household j in area i, β is a px1 vector of parameters (including the intercept), the U_i's are random effects for area i = 1,...,35 assumed to be independent and normally distributed with zero mean and variance σ_u and ε_{ij}'s are random independent draws from a normal distribution with zero mean and unknown variance σ_ε.

In analysing the actual car use we consider only those households in the sample who own a car. This results in a somewhat smaller sample of 15028 items out of the original 22648 observations. We also discard some further items in terms of the analysis. In particular, we omit those households from the original sample who indicated a distance of less than 500 km travelled, since these observations seem very unreliable and in particular so since these observations mostly refer to households that have owned their car for at least one year. Among the remaining observations we also omitted those households who owned a car but who did not provide any information about car usage, or car technology. The data set was thus reduced to 14985 records (we also discarded a few observations that contained missing values for the predictors of interest).

As level 1 predictor variables we use the demographic characteristics of the household as well as car technology variables. In case the NUTS-3 variance turns out to be significant, we include population density as a contextual variable and check whether it controls for regional heterogeneity. In this particular case model (2) specifies in (3)

$$y_{ij} = X_{ij}^T \beta + Z_i \gamma + U_i + \varepsilon_{ij} \qquad (3)$$

where Z_i stands for the population density of NUTS-3 region i. A likely result could be that the random effect of regional heterogeneity disappears, which would indicate that all heterogeneity can be explained by the contextual variable. Similar to equation (3) we will also test whether population density is an appropriate contextual variable for car ownership. In this case equation (1) will be extended to yield equation (4):

$$log \frac{Pr(Y_{ij} = 1)}{Pr(Y_{ij} = 0)} = X_{ij}^T \beta + Z_i \gamma + U_i, \qquad (4)$$

The previous models can be estimated using different statistical approaches. We have chosen to use iterative generalised least squares (IGLS). This approach is well known in the statistical literature and it is fully described in a multilevel set up (see Goldstein, 1995). The basic idea is to begin by estimating the fixed parameters in an initial, ordinary least square regression, and then to use the obtained residuals to compute the dispersion matrix of the response. After this, an iterative procedure starts. The first step consists of a generalised least square regression. The second step uses the obtained residuals to compute their matrix product. Stacking the columns, one on top of the other, we obtain a vector, making it possible to compute the variance of the random coefficients using an appropriate design matrix for those random coefficients. The two steps are then iterated. The procedure gives maximum likelihood estimators under the normality assumption. Concerning multilevel modelling of binary data, marginal quasi-likelihood and penalised quasi-likelihood, there are two prevailing approaches in approximating the maximum likelihood estimates. Both rely on the Taylor expansion to arrive at the approximation.

Different versions of these algorithms exist depending on whether first order terms, or second order terms, are included in the expansion. An extensive comparison of different approaches, and the risks connected with them, is presented more fully in Guo and Zhao (2000). However, we arrived at the model using different algorithms - as described in the preceding paragraph - resulting in minor changes only in the estimated values. A full discussion of the advantages and features of alternative approaches is beyond the scope of this paper and the interested reader should refer to the afore-mentioned literature.

4. Results

In Table 1 and Table 2 we summarise the results for car ownership and car use patterns in Austria. We apply the specified models outlined in section 3.2. For the estimation of the considered models, we used Mlwin 3.1 (Rasbash et al. 2000). For categorical predictors, the reference class is a female-headed, single person household where the household head is unemployed and of Austrian nationality.

To model the decision to use a car or not, we estimate the multilevel logit model with random effects as given by equation (1) (cf. Table 1). As predictor variables we choose the following set of household characteristics: age, gender, nationality, education and employment of the household head, the combination of household size and household composition (see section 2.2) and the measure of housing quality, which reflects the contrast between rented flats and owner-occupied flats or houses (see section 2.2).

A model that only includes the constant term and the random effect (Model M0) may be contrasted with a model that includes various household characteristics (Model M1). By inclusion of household characteristics the unobserved heterogeneity across NUTS-3 levels is strongly reduced, the variance estimate decreases for almost 48%. The parameter estimates for all household characteristics are significant. Car ownership is lower for households headed by older and non-naturalised persons and is higher for male, more educated and employed household heads, as well as for households that own their apartment or house. Compared to the reference class (single person households), all other living arrangements have higher car ownership rates. Three person adult-only households have the highest rate of car ownership while single parents with one child have the second lowest rate of car ownership. Car ownership rates are higher for larger households (3 or more persons). Taking into account their standard error, households of size 4 and 5+ show similar car use patterns.

If we consider that car ownership is determined by the necessity as well as the affordability of owning a car, the estimated demographic effects are intuitive. We expect higher incomes for Austrians, male, more educated and employed household heads and households with owner-occupied flats or houses. For these groups of persons it may therefore be easier to afford an own car. Moreover, the necessity to own a car and ease commuting may be higher for younger and employed persons and households that own their flat or house. Hence, for these latter groups of households the need for spatial mobility may be the determinant that influences car ownership the most. The fact that car ownership increases with the household size may be explained by two facts. First, larger

sized households are more likely to consist of more than one adult and hence the number of potential car owners is higher in these households. Secondly, larger sized households are also more likely to be households with children (more generally dependent persons) and the presence of children (dependent persons) may increase the demand for daily trips (e.g. to the kindergarten, school in case of children) for these households.

Table 1: Explaining car ownership patterns in Austria (logit model results)

	M2		M1		M0	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
Constant	0.026	0.129	0.433	0.130	1.104	0.054
Age	-0.032	0.002	-0.032	0.002		
Gender	0.863	0.049	0.859	0.049		
Nationality	-0.829	0.094	-0.825	0.094		
Education	0.211	0.013	0.206	0.013		
Employment	0.928	0.057	0.919	0.057		
Adult-only households of size 2	1.313	0.056	1.298	0.056		
Single parent with 1 child	0.992	0.086	0.984	0.086		
Adult-only households of size 3	1.983	0.133	1.962	0.135		
Households of size 3 with child	1.646	0.077	1.632	0.078		
Households of size 4	1.889	0.091	1.874	0.092		
Households of size 5+	1.899	0.117	1.888	0.118		
Measure of housing quality	0.619	0.029	0.629	0.028		
Population density	-0.00012	0.00004				
Variance Components						
NUTS-3	0.026	0.011	0.045	0.016	0.086	0.024

In addition to the level 1 variables of household demographic characteristics, we include population density as a regional level variable in model M2. Since including population density helps to further reduce the level 2 variance for almost 42% in comparison to model M1, population density may be regarded as an important contextual variable. The parameter estimates on the level 1 variables are stable across models M2 and M1.

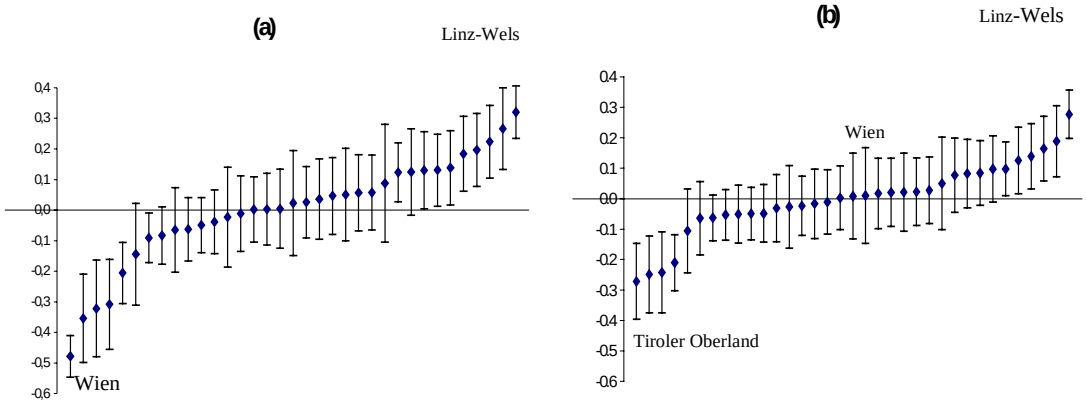
Obviously our intuition laid out in the introduction that factors of stability may impact on car ownership seems to be proven. Population density is generally a slow changing variable. Since population density is a good predictor for the prevailing local infrastructure and public transportation means it therefore explains part of the regional heterogeneity in car ownership.

To highlight the geographical differences in car ownership patterns, Figure 1.a and Figure 1.b show the plots of the estimated regional effects plus and minus their standard deviation against their rank for both model M1 and M2 of Table 1. The regions with

highest and smallest residuals U_i are reported in the picture together with Wien. It is quite evident that for some regions the effect coefficients are (either in a positive or negative way) significantly different from zero in both models, while in other regions the effect coefficients are not significant. (The estimated regional effect is regarded as significant if the coefficient differs from zero more than once in its standard deviation.) It is also evident that the number of significant regional effects decreases (from 15 to 10) and the range of their values narrows when population density is introduced as a further regressor (cf. Figure 1.b).

In Figure 2 we present the estimated regional effects on the log odds of car ownership (2.a and 2.c) and their significance (2.b and 2.d) in reference to a map of the NUTS-3 regions[7] (the cut points used to color the maps are based on the quintiles of the regional effects estimated according to model M1). The darker the area, the stronger its effect is on the log odds. It can easily be seen that the effects are smaller and less significant once we introduce population density as a further regressor variable (cf. Figure 2.c and 2.d).

Figure 1: Caterpillar plots of estimated regional effects (plus/minus standard deviation) of car ownership for model M1 (a) and model M2 (b) of Table 1.



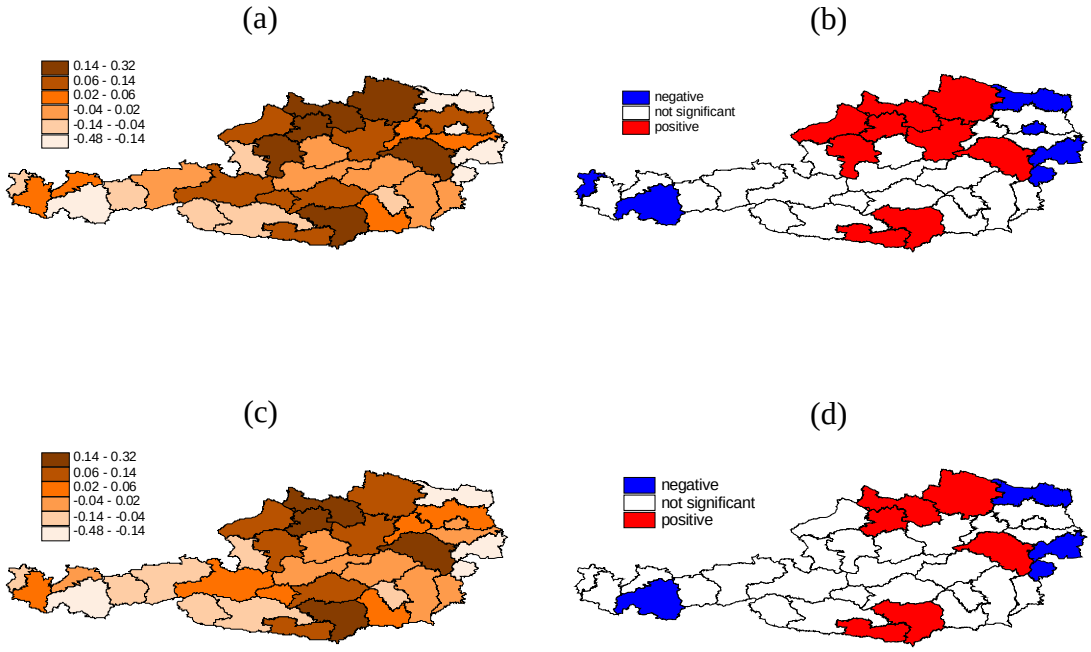
To explain car use patterns in Austria we apply the linear random effect model as outlined in the previous section, equation (2) (cf. Table 2). Figure 3 gives an overview of the nested model sequence we performed.

Starting from a simple model that only includes a constant (Model M0) and introducing the variance component at level 2 (Model M1) shows significant variations of annually driven distances across NUTS-3 regions.

The proportion of the total interregional variance σ^2_u is about 1.5% of the total variance ($\sigma^2_u + \sigma^2_\varepsilon$). We elaborate on the model M1 by adding a set of explanatory variables that refer to car technology (model M2) and by adding the set of household characteristics we introduced in Table 1 (model M3).

[7] All the maps depicted in this paper are made using Arc/View GIS Esri 1996.

Figure 2: Map of the estimated regional effects and their significance for model M1(a,b) and model M2 (c,d) of Table 1.



By introducing the car characteristics (number of cars, degree of newness, diesel technology and presence of a catalytic converter), the level 2 variance declines further. Compared to Model M1, the variance of the random component decreases by almost 78% in Model M2. Car characteristics are therefore important variables that account for regional variation of car use across the NUTS-3 regions.

According to the fixed parameter estimates, it follows that the number of cars and the use of diesel engines increases car use while the newness of the car and the presence of catalytic converters are related to a decrease in car use.[8]

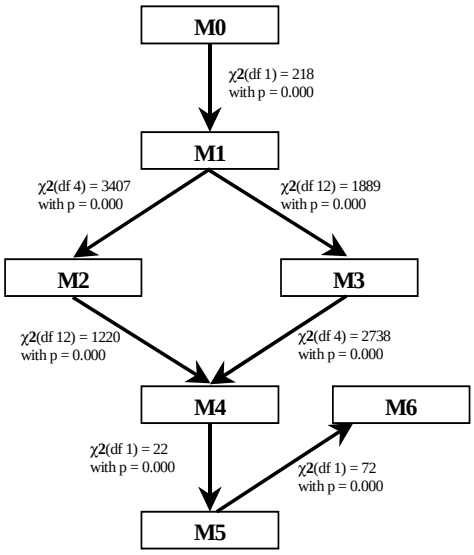
Recalling that car ownership is characterised by significant regional variability it seems quite evident, that controlling for the number of cars per household also reduces the regional variability of car use. In a very similar manner the use of diesel engines tends to be clustered in NUTS-3 regions in the north east and south east of Austria, where many households commute to their working places in Vienna, Linz, Salzburg or Graz.

[8] As far as the causal direction between car technology and car use is concerned, one may argue that causality runs the other way around. I.e. for a single household a high intensity of car use in the long-run should result in the choice of a car with diesel engine which will be cheaper in terms of fuel costs. However in our set up we are interested to explain car use patterns and we therefore just include the use of diesel engines as a control variable in order to avoid that effects of all other considered explanatory factors on car use are overestimated.

Obviously commuting is cheapest with diesel engine cars, because of the significant lower costs of fuel.

In contrast, the inclusion of household characteristics (Model M3), instead of car characteristics, leads only to a decrease of 33% in the level 2 variance, as compared to Model M1. While household characteristics definitely add to the explanatory power of the model, they cannot explain variations across NUTS-3 levels. The fixed parameter estimates on the household characteristics are qualitatively similar to those obtained in the case of car ownership (cf. Table 1). However, there are some noteworthy differences: while the propensity to own a car is lower for non-Austrian household heads, the pattern of car use is the same between Austrians and non-nationals, as evidenced by the corresponding insignificant parameter estimate. Another noteworthy exception is the relative magnitude of the coefficients on the variable combining household size and household composition. Single person households not only have the lowest car ownership rates but the lowest actual car use as well. Among 2 person households, however, actual car use is higher, as long as at least one child is present. The opposite result could be observed for car ownership, i.e. adult only households had higher car ownership rates for households of size 2.

Figure 3: Path of the considered models.

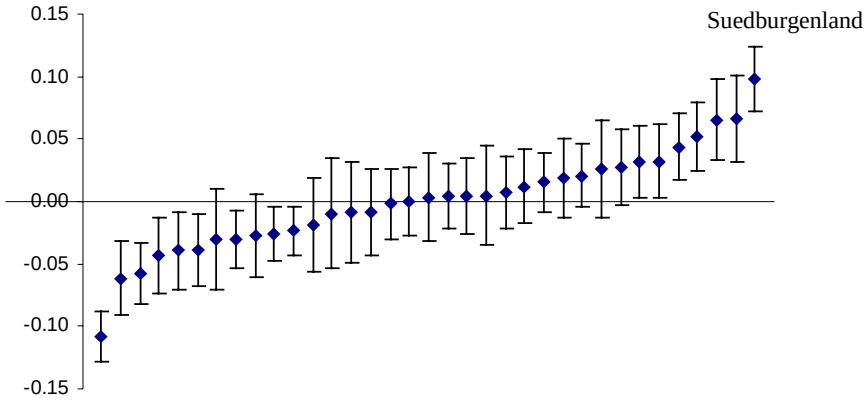


These results are not unexpected. Car ownership is closely related to one's financial means and the legal driving age, while actual car use is more closely related to the need to commute, which could itself be a result of the number of children (e.g. daily trips to the kindergarten, school, leisure time activities). Adding car technology variables and household characteristics as predictor variables in one model (M4) does not essentially improve the fit of the model, especially when compared to the model in which only the car technology variables are added (M2).

We also tested whether population density would be an appropriate contextual variable to explain the remaining level 2 variance. However, as Model M5 evidences, the inclusion of population density does not reduce the significance of the level 2 variation and therefore cannot account for the remaining heterogeneity left between NUTS-3 regions. Also, the estimated effect of population density is insignificant. As a final check of the role of population density we compare model M5 with model M6, the difference being that in M6 the random effect has been omitted. We find that when we control for population density without accounting for an unspecific random variation effect, the coefficient on population density becomes significant and therefore overshadows the regional variations. But as model M5 shows, this unspecific random variation between NUTS-3 regions does not disappear by introducing a regionally varying measure such as population density.

Figure 4 is an effort to clearly represent the geographical pattern of car use. Figure 4 plots the estimated regional effects (plus/minus their standard deviation) versus their rank for model M4, Table 2. Regional patterns of car use are less dispersed when compared with the regional pattern of car ownership. Regional variations seem therefore to have more of an effect on the individual decision to own a car than on the actual pattern of car use.

Figure 4: Caterpillar plot of estimated regional effects (plus/minus standard deviation) of car use for model M4 of Table 2.



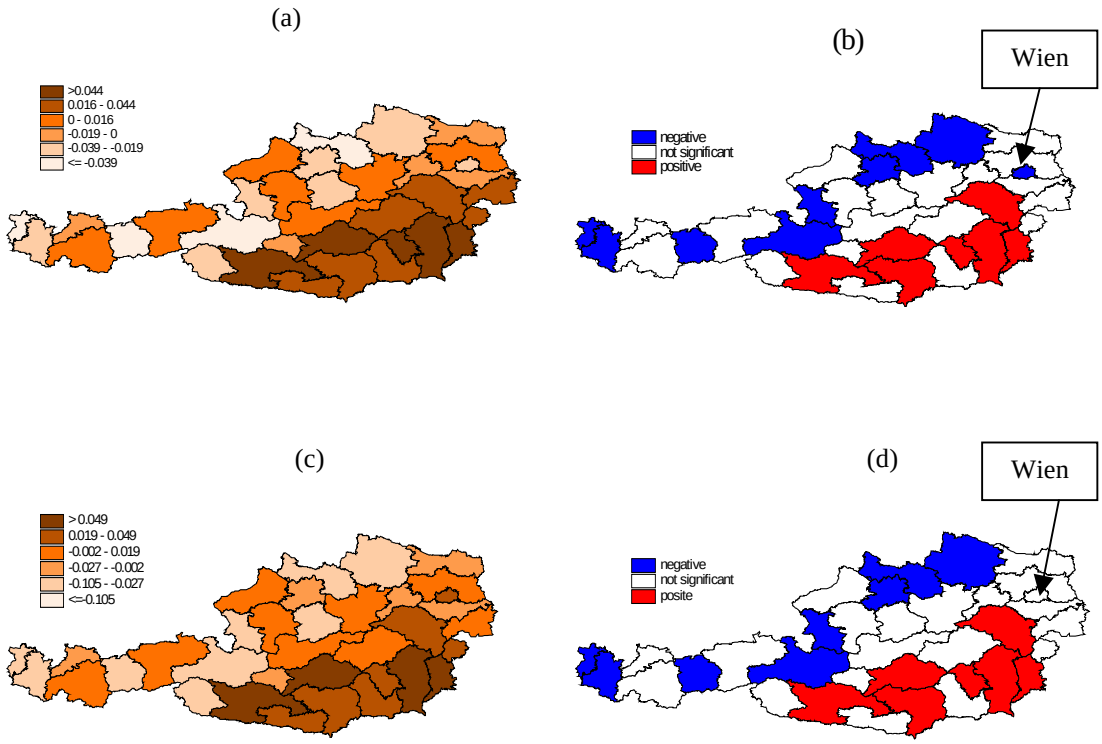
In Figure 5.a and 5.b, we have plotted the regional effects as indicated in Figure 4 and their significance in a map of NUTS-3 regions. It is quite evident that the stronger positive effects tend to cluster in the south of the country while the stronger negative effects tend to cluster in the north. This regional pattern is different from the one observed for car ownership (Figure 2). In the pattern for car ownership we can see stronger negative regional effects in the western and eastern part of Austria, while the stronger positive effects tend to cluster in the North and South of Austria.

Table 2: Explaining car use patterns in Austria (Linear Regression Model)

	M6	M5	M4	M3	M2	M1	M0
Parameter	Estimates *Std. Err.*	Estimates *Std. Err.*	Estimates *Std. Err.*	Estimates *Std. Err.*	Estimates *Std. Err.*	Estimates *Std. Err.*	Estimates *Std. Err.*
Constant	8.856 *0.039*	8.864 *0.040*	8.859 *0.040*	9.183 *0.043*	8.695 *0.017*	9.478 *0.018*	9.455 *0.006*
Number of cars	0.512 *0.010*	0.507 *0.010*	0.506 *0.010*		0.552 *0.010*		
Degree of newness	-0.308 *0.028*	-0.304 *0.028*	-0.304 *0.028*		-0.261 *0.029*		
Diesel technology	0.264 *0.013*	0.259 *0.013*	0.259 *0.013*		0.281 *0.013*		
Catalytic converter	-0.175 *0.013*	-0.176 *0.013*	-0.176 *0.013*		-0.182 *0.013*		
Age	-0.009 *0.001*	-0.009 *0.001*	-0.009 *0.001*	-0.006 *0.001*			
Gender	0.142 *0.016*	0.144 *0.016*	0.144 *0.016*	0.152 *0.018*			
Nationality	0.047 *0.029*	0.065 *0.030*	0.065 *0.030*	0.014 *0.032*			
Education	0.026 *0.003*	0.027 *0.003*	0.027 *0.003*	0.032 *0.003*			
Employment	0.112 *0.016*	0.116 *0.016*	0.116 *0.016*	0.191 *0.018*			
Adult-only household of size 2	0.041 *0.020*	0.042 *0.020*	0.042 *0.020*	0.100 *0.022*			
Single parent with 1 child	0.123 *0.032*	0.125 *0.032*	0.125 *0.032*	0.181 *0.035*			
Adult-only household of size 3	0.146 *0.033*	0.141 *0.033*	0.140 *0.033*	0.295 *0.036*			
Households of size 3 with child	0.121 *0.021*	0.121 *0.021*	0.121 *0.021*	0.303 *0.023*			
Households of size 4	0.072 *0.021*	0.073 *0.021*	0.073 *0.021*	0.282 *0.023*			
Households of size 5+	0.043 *0.024*	0.045 *0.024*	0.044 *0.024*	0.331 *0.026*			
Measure of housing quality	0.028 *0.007*	0.025 *0.007*	0.026 *0.007*	0.086 *0.008*			
Population density	-0.00005 *0.000001*	-0.00005 *0.00005*					
Variance Components							
NUTS-3		0.002 *0.001*	0.002 *0.001*	0.006 *0.002*	0.002 *0.001*	0.009 *0.003*	
Residual	0.445 *0.005*	0.443 *0.005*	0.443 *0.005*	0.531 *0.006*	0.481 *0.006*	0.602 *0.007*	0.613 *0.007*
-2*Log(L)	30408	30336	30358	33096	31578	34985	35203

Introducing population density among the predictors does not affect the spatial pattern of the estimated regional effects (cf. Figure 5.c and Figure 5.d) .The only difference is the region of Wien. Once population density has been taken into account, Wien is no longer characterised by a significant negative effect. It might be noted that Wien was also one of the areas that's effect on the log odds of car ownership disappears when population density is included.

Figure 5: Map of the estimated regional effects and their significance for model M4 (a, b) and model M5 (c, d) of Table 2.



5. Conclusion

Environmental behaviour in industrialised countries is commonly explained and associated with economic variables. However, as recent literature has evidenced, it is equally important to include demographic factors in explaining environmental behaviour. In this paper we expand upon this recent literature. We also take into account the regional heterogeneity of environmental behaviour. Specifically, we apply a multilevel statistical approach in order to estimate the significance of individual level demographic characteristics (level one variables) and regional specific context variables (level two variables) in explaining car ownership and actual car use patterns in Austria.

As they regard car ownership patterns, our results show that household characteristics such as age, gender, nationality, education and employment of the household head as well as household size, family structure and housing quality are significant predictor variables. Moreover, we have shown that a significant regional heterogeneity in car ownership patterns across NUTS-3 regions in Austria exists and can be partly explained by household-level demographic characteristics. Hence, there exists some regional clustering of specific household types in Austria. In addition to these household demographic (level one) characteristics, we have demonstrated that part of the regional heterogeneity may be further reduced by including population density (a level two variable) as a further predictor.

Our results for actual car use have shown that along with household demographic variables, car technology variables are significant factors in explaining actual car use patterns. Once again, regional heterogeneity is clearly evident in patterns of car usage. In explaining the regional heterogeneity, however, only car technology variables are relevant, while demographic individual level variables and population density add very little to explaining the NUTS-3 variance. It seems as if some of the cars' technology measures are not randomly distributed over the 35 NUTS-3 regions. The number of cars per household and the degree of diesel engine usage are especially correlated to population density. In densely populated regions the number of cars both with and without diesel engines is lower than in rural regions (cf. Ewert and Prskawetz 2000, 2001) the reason presumably being that high population density also is a good indicator for a well structured public transportation system which meets many of the private mobility needs. Given this regional clustering of car characteristics, it is not surprising that population does not add anything to the explanation of the inter-regional variance, as long as we control for the cars' technical features.

In summary, our results imply that household characteristics are important predictor variables for car ownership and car use patterns. Besides the importance of household demographic variables we have shown that there exists a significant regional heterogeneity of car ownership and car use patterns across the NUTS-3 regions in Austria. This regional heterogeneity can partly be reduced by household characteristics in the case of car ownership patterns and by car technology variables in the case of actual car use. In addition to these level one variables, the contextual variable of population density helps to further reduce the regional heterogeneity in car ownership patterns.

Referring to a graphical representation of the regional effect coefficients, we have demonstrated that the regional heterogeneity of car ownership and actual car use is different. While we find strong negative regional effects on car ownership patterns in the western and eastern regions of Austria, negative regional effects for actual car use are mostly clustered in the northern regions of Austria. Contrastingly, the positive regional effects on car ownership are concentrated in the northern and southern parts of Austria while the positive regional effects on car usage are concentrated in the southern parts of Austria.

References

Adlmannseder, J., 1993. 'Raum- und Zeitverhalten von Ozon in den westlichen Stadtbezirken von Graz', *Mitteilungen des naturwissenschaftlichen Vereins der Steiermark* 123: 19-31.

Breiman, L., Friedman, J. H., Olshen, R.A. and Stone, C. J., 1984. *Classification and regression trees*. Wadsworth International Group, Belmont (CA.).

Buettner, T. and Grubler, A., 1995. 'The birth of a "green" generation? Generational dynamics of consumption patterns', *Technological Forecasting and Social Change* 50: 113-134.

Canzler, W. and Knie, A., 1994. *Das Ende des Automobils. Fakten und Trends zum Umbau der Automobilgesellschaft*. Verlag C. F. Müller, Heidelberg.

Carlsson-Kanyama, A. and Linden, A.-L., 1999. 'Travel patterns and environmental effects now and in the future: Implications of differences in energy consumption among socio-economic groups', *Ecological Economics* 30: 405-417.

Congdon, P., 2000. 'Monitoring suicide mortality: a bayesian approach', *European Journal of Population* 16: 251-84.

Chambers, J. M. and Hastie, T. J., 1991. Statistical Models in S. Chapman & Hall, New York.

Dahl, C. A. and Sterner, T., 1991. 'Analysing gasoline demand elasticities: A survey', *Energy Economics* 13 (3): 203-210.

DiPrete, T. A. and Forristal, J. D., 1994. 'Multilevel models: Methods and substance', *Annual Review of Sociology* 20: 311-57.

Ewert, U.C. and Prskawetz, A., 2000. 'Private car use in Austria by demographic structure and regional variations'. MPIDR Working Paper WP 006/2000, Rostock.

Ewert, U.C. and Prskawetz, A., 2001. 'Can regional variations in demographic structure explain regional differences in car use? A case study in Austria', *Population and Environment* 23(3): 315-345.

ESRI, 1996. *Using ArcView GIS*. Enviromental Systems Research Institute, Redlands (Ca.).

Fahrmair, L. and Lang, S., 2001. 'Bayesian inference for generalised additive mixed models based on random Markov field priors', *Applied Statistics* 50: 201-220.

Franzen, A., 1997. *Umweltbewusstsein und Verkehrsverhalten. Empirische Analysen zur Verkehrsmittelwahl und der Akzeptanz umweltpolitischer Massnahmen.* Rüegger, Chur, Zurich.

Goldstein, H., 1995. *Multilevel Statistical Models*. 2nd edition, Edward Arnold, London.

Grenning, L. A. and Jeng, T. H., 1994. 'Lifecycle analysis of gasoline expenditure patterns', *Energy Economics* 16(3): 217-228.

Greening, L. A. , Schipper, L., Davis R.E. and Bell, S.R., 1997. 'Prediction of household levels of greenhouse gas emissions from personal automobile transportation', *Energy* 22 (5): 449-460.

Gou, G. and Zhao, H. (2000). 'Multilevel modeling for binary data' *Annual Review of Sociology,* 26:441-62.

Heigl, A. and Mai, R., 1998. 'Demographische Alterung in den Regionen der EU', *Zeitschrift für Bevölkerungswissenschaft* 23: 293-317.

Kautz, H., Krajasits, C. and Eisenkölb, G., 1999. *Regionalbericht 1998*. Österreichisches Institut für Raumplanung (ÖIR), Vienna.

Krause, C., 1997. 'Auto-Typen. PKW-Zielgruppenforschung mit den SINUS-Milieus', *Media Spectrum* 11/97: 32-33.

Lorbeer, D. A., 1996. 'Auto und Umwelt', *Gegenwartskunde. Zeitschrift für Gesellschaft, Wirtschaft, Politik und Bildung* 45/1: 101-111.

Mikl-Horke, G. and Leuker, H., 1978. *Das Auto als Verhaltensdeterminante* (Berichte des Institutes für Allgemeine Soziologie und Wirtschaftssoziologie an der Wirtschaftsuniversität Wien 17). 2nd edition, Österreichische Gesellschaft für Wirtschaftssoziologie, Vienna.

Österreichisches Statistisches Zentralamt (ÖSTAT), 1998. *Engergieverbrauch der Haushalte 1996/97: Ergebnisse des Mikrozensus Juni 1997*, Beiträge zur österreichischen Statistik, Heft 1.279.

O'Neil, B. C. and Chen, B., 2001 'Demographic determinants of energy use in The United States' fortchcoming in a Special Supplement to Population and Development Review.

Prskawetz, A., Leiwen, J. and O'Neill, B. C., 2001. 'Demographic composition and projections of car use in Austria'. Paper presented at the IIUSP Brasil, August 2001.

Pucher, J., Evans, T. and Wenger, J., 1998. 'Socioeconomics of urban travel: Evidence from the 1995 NPTS', *Transportation Quarterly* 52(3): 15-33.

Rasbash, J., Browne, W., Goldestein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Lanfford, I. and Lewis, T., 2000. *A User's guide to MlwiN*. Institute of Education, University of London, London.

Silverman, B.W., 1985. *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Spain, D., 1997. 'Societal trends: The aging baby boom and women's increased independence'. Report prepared for the U.S. Dept. of Transporation. Order no. DTFH61-97-P-00314.

Steele, F., Diamond, I., and Amin, S., 1996. 'Immunisation uptake in rural Bangladesh: A Multivel Analysis', *Journal of the Royal Statistical Society Ser.* B 159: 289-299.

Umweltbundesamt, 1998. *Umweltsituation in Österreich.* Fünfter Umweltkontrollbericht des Bundesministers für Umwelt, Jugend und Familie an den Nationalrat, Vienna.

van der Gaag, N., van Imhoff, E. and van Wissen, L., 2000. 'Internal migration scenarios and regional population projections for the European Union', *International Journal of Population Geography* 6: 1-19.

Venables, W. N. and Ripley, B.D., 1999. Modern applied statistics with S-plus. III ed. Springer. New York.

Wakefield, J.C., Best, N.G. and Waller, L., 2000. 'Bayesian approach to disease mapping', in P. Elliott, J. C. Wakefield, N. G. Best and D. J. Brings (eds), *Spatial epidemiology: Methods and applications* (Chapter 7). Oxford University Press, Oxford.

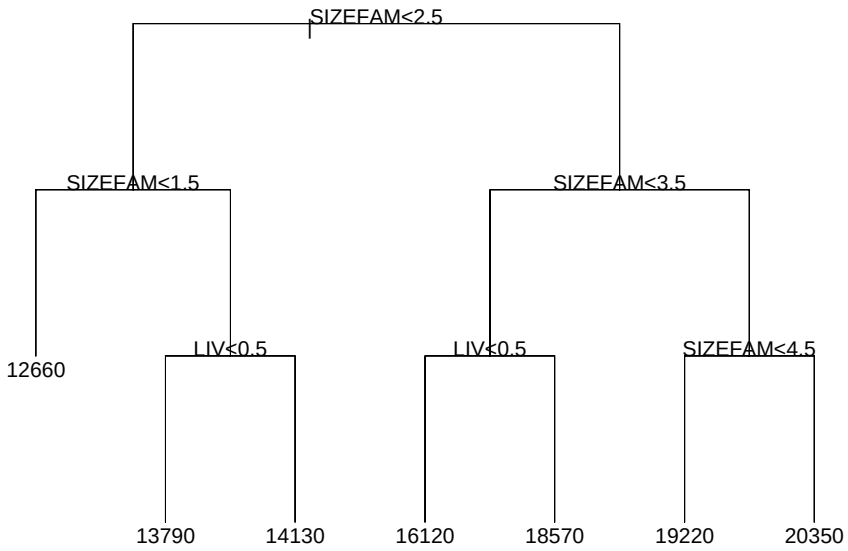
Wellner, A. S., 2000. Who is in the House?, *American Demographics (January)*, 48-51.

Yamamoto, T. and Kitamura, R., 2000. 'An analysis of household vehicle holding durations considering intended holding durations', *Transportation Research A* 34: 339-351.

Appendix A

Tree based models rely on a recursive, binary partition of the predictor space into disjunct subspaces (Breiman et al. 1984). In each, partition data are split into subgroups (called nodes). The aim is to find the partition that maximally distinguishes the response variable in the left from the right branches. This procedure is continued until the nodes are homogeneous (according to a selected index), or data are too sparse. This procedure is highly adaptive and thus special care has to be taken to avoid an overfitting of the data. A common approach is to grow a large tree and then begin simplifying it by pruning it back until an adequate model is found. Usually a good strategy is to split the data in (at least) two parts: a training set used to build up the tree, and a test set to check the tree's robustness. Other more sophisticated and computationally intensive methods are proposed in the literature (see for instance Chambers and Hastie, 1991). In our study, this procedure was not used in an inductive way to find a model. Instead, we applied the regression tree technique to test whether a newly formed variable was powerful enough to correctly distinguish between several combinations of household size and household composition.

Figure A1: The estimated classification tree to code a new living arrangement predictor.



Nevertheless, to check the performance of the new predictor we applied the regression tree technique to 60% of our data (selected at random) and checked it against the residual set. The result is the distance driven per household. After pruning (we deleted some

nodes, but this did not affect the residual deviance of the resulting model) we obtained the tree reproduced in Figure A1.

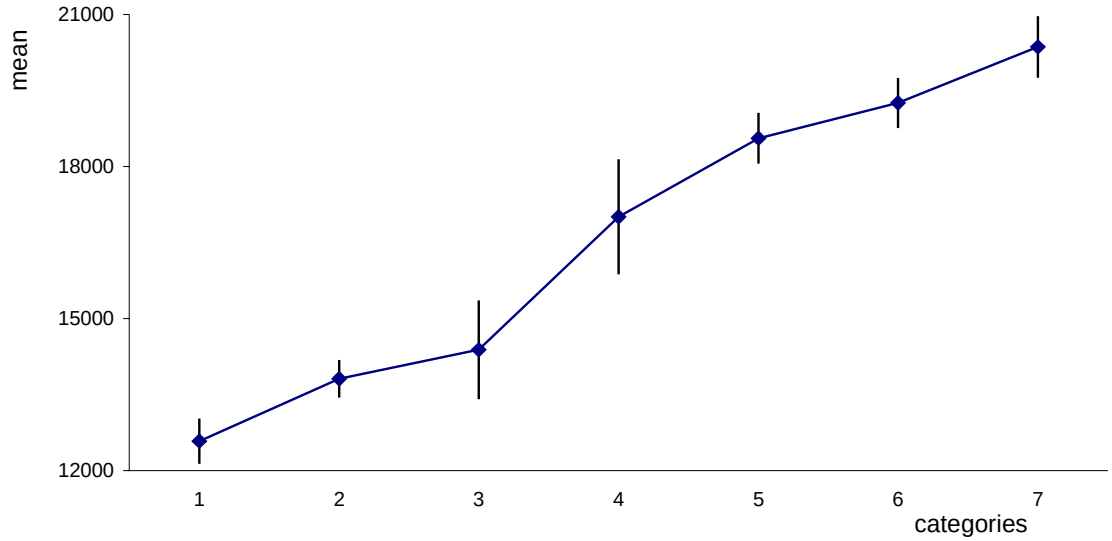
Figure A1 conforms to our expectations. In order to make the figure clearer, the lengths of the branches are uniform rather than proportional to the resulting deviance at each split. The final nodes show the estimated value of the response. Labels at the nodes point out which predictor was used to split the data. To get an idea of the robustness of our model we computed the average distance driven in each category of the new predictor for our test set. We then compared it with the one predicted by the model (checking individual residuals for such a smooth model does not seem to make much sense as the model we arrived at is a very smooth one and fits the individual data well). This comparison is reported in table A2. Our observed and predicted values are very similar.

Table A2: Results of applying the estimated regression tree on the test set.

Categories	Predicted	Observed	Low. C. I.	Up. C. I.
1	12660	12434	11712	13156
2	13790	13859	13279	14440
3	14130	14814	13254	16373
4	16120	18664	16823	20504
5	18570	18541	17747	19336
6	19220	19326	18543	20110
7	20350	20392	19406	21379

In Figure A2 we plot the mean distance driven for the various classes of the new variable. We did this for the whole data set. The new predictor seems to discriminate pretty well between the different groups.

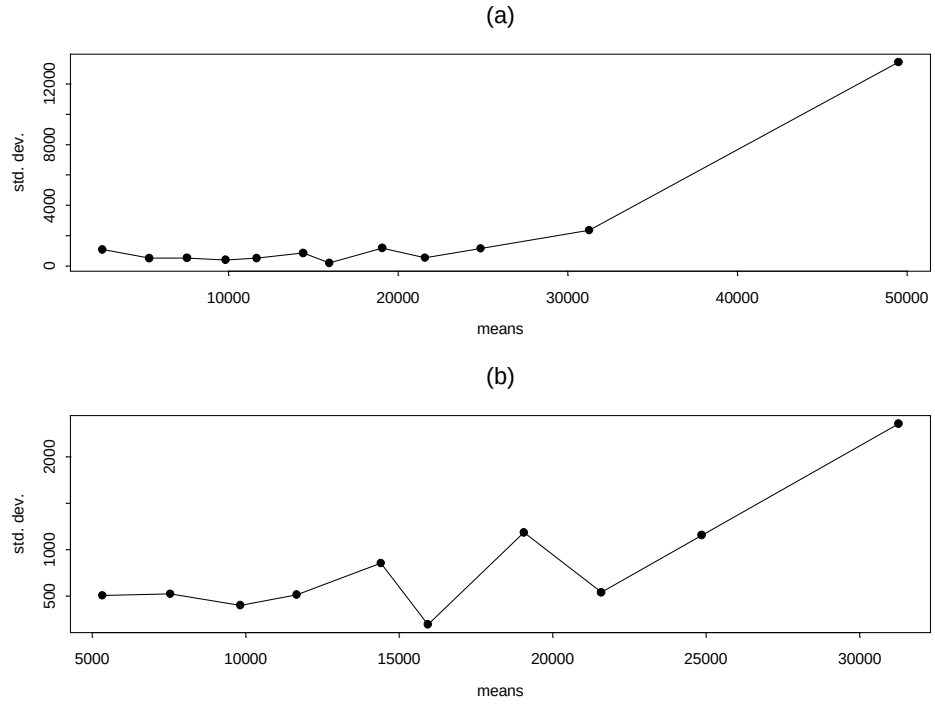
Figure A2: Average distance driven in the class of the new variable with their confidence intervals.



Appendix B

First of all we try to test -at least in an empirical way - the assumptions under the linear model (Venables and Ripley, 1999).To check homoscedasticity of the response variable (distance in kilometres driven during a one year period) we split up the set of its values into several groups. We used the first thirteen percentiles as cut points. For each subset of values we computed the mean and the standard deviation of the response. The results are plotted (together with an interpolating line) in Figure B1.a. It is quite clear that the response variance increases with the mean. Even if we drop the first and last subset (Figure B1.b), i. e. the smallest and the biggest (about) 8% of the response values, where the outliers are located, the assumption of homoscedasticy does not seem very plausible.

Figure B1. Standard deviations versus means of twelve intervals of the response variable values.



Thus we transform the data according to a logarithmic transformation. The same plot (as Figure B1) for such transformed data is reported in Figure B2.
We actually ran two linear regressions with the original and the transformed response, including in the predictor all the variables listed in Table 2 of section 4. The analysis of residuals for those models (not reported here) suggested again that log transformation makes the linear model more suitable for the considered data.
The estimated distribution of the response variable is depicted in Figure B3.a with the continuos line being the kernel estimator (Silverman 1985). The distribution of the logarithmic transformation of the response is shown in Figure B3.b.

Figure B2. Standard deviation versus mean of twelve intervals of the response variable values transformed according to a logarithmic transformation

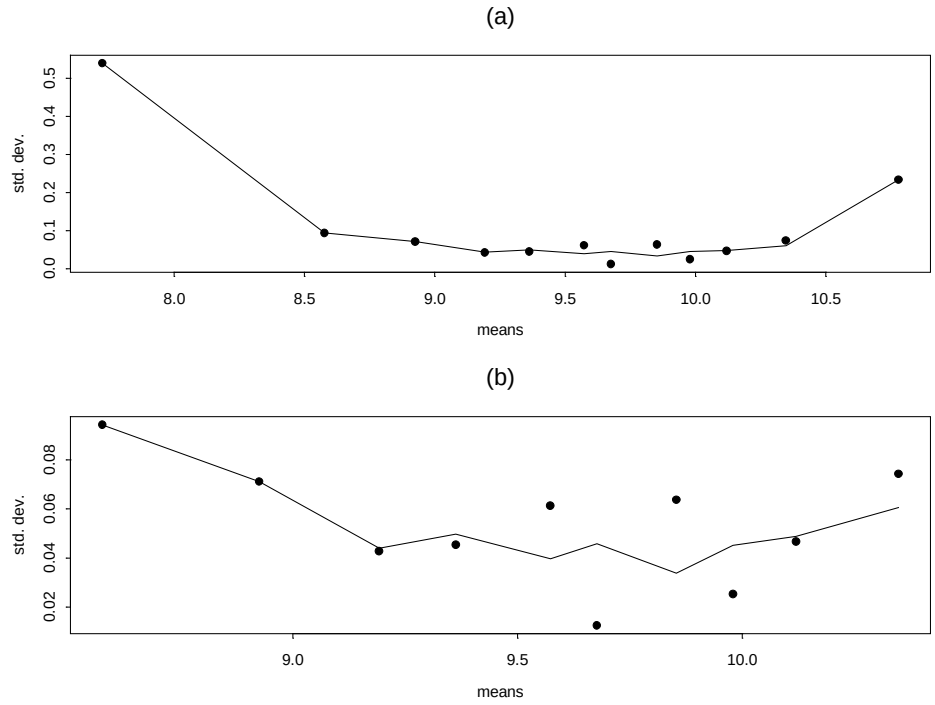
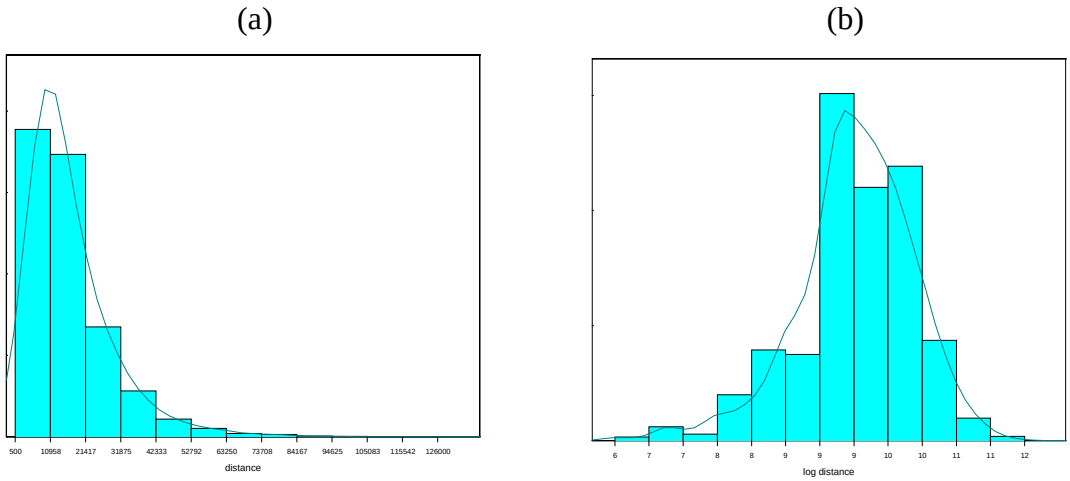


Figure B3. Estimated density distribution for distance driven (a), and its logarithmic transformation.



The estimated model performs quite well according to a standard model diagnostic. Figure B4.a and B4.b give and overall picture of this. We present the plot of predicted versus observed values and the quantile-quantile plot for residual normality, respectively.

Figure B4: (a) predicted versus observed and (b) qq plot