The Journal of Transport and Land Use http://jtlu.org
Vol. 11 No. 1 [2018] pp. 1101–1123

Estimating bid-auction models of residential location using census data with imputed household income

Benjamin Heldt (Corresponding Author)
DLR Institute of Transport Research
benjamin.heldt@dlr.de

Pedro Donoso
Universidad de Chile–Santiago
pedrodonosos@gmail.com

Francisco Bahamonde-Birke
DLR Institute of Transport Research
and Technical University of Berlin
francisco.bahamondebirke@dlr.de
bahamondebirke@gmail.com

Dirk Heinrichs
DLR Institute of Transport Research
and Technical University of Berlin
dirk.heinrichs@dlr.de

Abstract: Modeling residential location as a key component of the land-use system is essential to understanding the relationship between land use and transport. The increasing availability of censuses such as the German *Zensus 2011* has enabled residential location to be modeled with a large number of observations, presenting both opportunities and challenges. Censuses are statistically highly representative; however, they often lack variables such as income or mobility-related attributes as in the case of *Zensus 2011*. This is particularly problematic if missing variables define utility or willingness-to-pay functions that characterize choice options in a location model. One example of this is household income, which is an indispensable variable in land-use models because it influences household location preferences and defines affordable location options. For estimating bid-auction location models for different income groups, we impute household income in census data applying an ordered regression model. We find that location models considering this imputation perform sufficiently well as they reveal reasonable and expected aspects of the location patterns. In general, imputing choice variables should thus be considered in the estimation of residential location models but is also promising for other decision problems. Comparing results for two imputation methods, we also show that while applying the deterministic first preference imputation could yield misleading results, the probabilistic Monte Carlo simulation is the correct imputation approach.

Article history:
Received: October 31, 2016
Received in revised form: August 14, 2017
Accepted: November 24, 2017
Available online: November 20, 2018

http://dx.doi.org/10.5198/jtlu.2018.1040
ISSN: 1938-7849

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies. This paper is also published with additional sponsorship from WSTLUR.

1 Introduction

Being essential for understanding land use–transport interactions and support planning decisions, residential location models are rooted in a long-standing history (Acheampong & Silva, 2015; Cordera, Ibeas, dell'Olio, & Alonso, 2017; de la Barra, 1989; Hunt, Kriger, & Miller, 2005; Iacono, Levinson, & El-Geneidy, 2008; Timmermans, 2003; Wegener, 2011, 2014). In urban planning and science, they are valuable tools that complement transport models by incorporating accessibility measures as factors influencing location choice (Martínez, 1995; Moeckel, 2018; Ortúzar & Willumsen, 2011; Wegener, 2014).

Such models follow basically three different theories and types: spatial interaction models, econometric models, and microsimulation models (de la Barra, 1989). One of the recently most-applied econometric models is the bid-auction discrete choice model (Martínez, 1992, 1996; Martínez & Donoso, 2010; Martínez & Henriquez, 2007), which is based upon bid-rent theory (Alonso, 1964) and an alternative approach to maximizing utility (Ellickson, 1981; McFadden, 1978). The relevant difference to utility-based discrete choice models is the inverse relationship between options and choosers. While in classical location models households or firms are choosers, in the bid-auction model they are options; simply speaking owners of locations choose residents (households or firms). Maximizing-utility and bid-auction approaches are equivalent under a demand-supply equilibrium condition (Martínez, 1992), i.e., households and firms are located in those real estate options that provide them maximum utility and are also highest bidders if all agents are located somewhere within the city.

Following the bid-auction approach, the household residing in a dwelling is the highest bidder among all households interested in being located there and the paid rent is the bid of this highest bidder. Then, rent is an endogenous outcome of the location process. In this light, residential bids reflect the preferences of households in line with their budgets and determine location and rents simultaneously. It is usually assumed that bids are Gumbel-distributed functions, which yields a closed-form of location probabilities: a multinomial logit model. Its parameters are commonly estimated on the basis of revealed location preferences. The result of this estimation is a set of bid functions which describe preferences of different agents such as households segmented by number of persons or income.

Location models are very data-demanding and their reliability depends on the data used for parameter estimation (Heldt, Gade, & Heinrichs, 2014). While several governmental and private institutions increasingly obtain large data sources at high spatial resolution, rising concerns regarding data protection complicate the use of detailed geocoded information. Since location models are to a great extent rooted in transportation research, they traditionally build upon travel-survey data. This includes information on mobility behavior and resources but often lacks crucial information regarding location, such as attributes of the dwelling or the neighborhood, which in turn depends on the spatial detail of the data. Census data that includes real estate information helps to overcome this by its large number of observations, enabling detailed geocoding without data protection issues. Such data is nonetheless associated with a lack of variables, including those that allow defining household groups with different location patterns. This limits the formation of alternatives in the bid-auction residential model and of choosers in the classical utility-based models.

One variable that is often missing in such data sources is household income. At the same time it is a crucial variable to differentiate households in location models as it defines available resources and thus affordable locations. Many studies of residential location choice include household income as a variable, either to segment choosers or define location characteristics or both (see the summary by Schirmer, Van Eggermond, and Axhausen (2014); examples are: Hurtubia and Bierlaire (2013), Ben-Akiva and Bowman (1998), Bhat and Guo (2007), Guo and Bhat (2004), Martínez (1996), Hunt et al. (2005), Zondag, de Bok, Geurs, and Molenwijk (2015), Zondag and Pieters (2005)). Furthermore, simulation studies require income as a variable in order to assess the effects of different compositions of income in the population and on urban structure and mobility, which is of increasing concern due to the rising divide in the income distribution in todays' societies. Household income is particularly relevant because it is related to household size and to certain life styles and therefore location preferences (Bhat, 2015). Finally, income defines to a large extent which mobility resources a household disposes of and is therefore indispensable when considering questions on the association between residential land use/mobility and transport systems. On this latter topic there is a considerable body of research (DeSalvo & Huq, 1996; LeRoy & Sonstelie, 1983; Paleti, Bhat, & Pendyala, 2013).

In the following paper, we show our approach to estimating a bid-auction location model based on imputed household income for Berlin, Germany. The first part of the paper introduces the main data sources, *Zensus 2011* and *Mikrozensus 2010*, and identifies lacking key variables. Subsequently, we describe our approach, which sequentially combines the estimation of an ordered logit model for income imputation, and a multinomial logit model for residential location choice applying either the deterministic first preference approach (henceforth FP-approach or FP) or the probabilistic Monte Carlo simulation (henceforth MC-approach or MC). In the next section we introduce the Berlin-specific context by providing general descriptive statistics of key variables of *Mikrozensus* and discuss the income imputation model and the resulting spatial distribution. This leads us to the formulation of hypotheses on the association between location preferences — in particular accessibility — and household income. Finally, we discuss the results of the location models for the two imputation approaches.

2 Data sources

The residential location bid function included in the auction-based location probability model depends on attributes of the households (which are the options in this discrete choice model) and attributes of the real estates and their locations, such as dwelling characteristics, zonal characteristics and accessibility measures. A comprehensive database is therefore needed that comprises individual households with their locations and all mentioned attributes (Heldt et al., 2014). Such data could be either gathered by an own survey or combined from existing public data sources. Existing surveys for our study area (Berlin) only covered very specific locations due to each survey's specific purposes — thus we decided to use public data. In the following sections, we introduce the main data sources available in Germany and describe their suitability for estimating location models. Subsequently, the process to link the different sources is described.

2.1 Mikrozensus 2010

The micro census is a Germany-wide survey of 1% of all households carried out annually in order to provide the administration with the main statistical numbers. Every four years, additional information is garnered on different topics, including housing. The corresponding dataset for Berlin (RDC of the Federal Statistical Office and Statistical Offices of the Länder, 2015a) includes about 15,000 households and dwellings. The data are geocoded at the coarse level of twelve districts, which is due to small sample size and thus cannot be used to directly estimate location bid functions with detailed spatial information. However, *Mikrozensus* [1] includes net household income information and other household attributes, and hence is very useful as auxiliary data.

[1] Henceforth, we refer to the micro census dataset of the year 2010 and for Berlin as *Mikrozensus*.

2.2 Zensus 2011

In 2011, the German Federal Statistical Office conducted a nationwide "register-based" census of population and buildings and dwellings, called *Zensus 2011* (RDC of the Federal Statistical Office and Statistical Offices of the Länder, 2015b). Instead of directly surveying persons and households, information was merged from several administrative registers, such as the population register, registers of employment agencies, etc. Census data for Berlin includes 3.3 million observations of persons and 1.8 million households. Since the original data is geocoded at block level, households could be assigned to a spatial reference system with relatively high resolution, such as traffic analysis zones [2] (TAZ)(Senatsverwaltung für Stadtentwicklung und Umwelt Berlin, 2012). This allows, in contrast to *Mikrozensus* the inclusion of detailed spatial characteristics and accessibility measures. *Zensus* [3] unfortunately lacks information regarding household resources, i.e., household income and related attributes. Other auxiliary data sources are therefore required in order to include household resources, which are important in defining options in bid-auction location models. Another disadvantage of *Zensus* is that it is updated only every ten years, but this does not pose a problem to estimating the bid-auction discrete choice model, because this is a static, not a dynamic model. However, this characteristic of census data should be kept in mind when applying it to dynamic models.

2.3 Accessibility and spatial indicators

Accessibility and spatial attributes which are defined in detail in Section 4.2 were calculated from a number of different data sources, including OpenStreetMap network data (OpenStreetMap contributors, 2016), a survey of retail establishments conducted by the Senate Department for Urban Development and the Environment of Berlin, land-use data from the Senate's Environmental Atlas (Senatsverwaltung für Stadtentwicklung und Wohnen Berlin, 2016b), and activity locations from commercial data providers.

2.4 Data processing

Information from data sources are processed and combined using an R-based computer program. *Zensus* and *Mikrozensus* person characteristics such as age are aggregated to household level using the concept of the household representative. We define the household head as the person who is assumed to decide where to locate, who in our model is either the oldest employed person if one or more members are working, or the oldest person if no household member is working. Other household characteristics are simple aggregations of dummy variables, such as the number of children or the number of persons by age. In order to calculate the number of dwellings, cases are aggregated at the building level and the resulting figure assigned back to the dwellings. Zonal attributes and accessibility indicators are added to dwellings from external sources at TAZ-level (cp. Section 4.2). Since *Zensus* defines a household as all persons who live together in a dwelling, dwellings and households are basically equivalent and dwelling attributes link directly to household attributes.

3 Methodology

As the aforementioned data source that is used to estimate residential location (*Zensus*) lacks income information, it is necessary to rely on an imputation process. This way, income categories will be imputed for this dataset by employing a model estimated with *Mikrozensus* data.

[2] To give an impression of the size of these zones: the City of Berlin (3.3 million inhabitants as of 2011) has 1,223 zones with an average size of 0.72 km^2.

[3] In the rest of the text, we use the term *Zensus* for the German census in 2011 for the City of Berlin.

Several data fusion approaches, including record linkage, multiple regression imputation, etc., have been developed during the last years to deal with missing data (Herzog, Scheuren, & Winkler, 2007; Rubin, 1987; Rubin & Little, 2002). These approaches are applicable in different situations depending on the type of the missing variable. Especially promising appear methods allowing for simultaneous imputation and estimation (Bahamonde-Birke & Hanappi, 2016; Sanko, Hess, Dumont, & Daly, 2014), which jointly consider the imputation model and the objective function (which in this case would be the location model and includes the missing variable).

In the case of the bid-auction probability models considered in this study, income characterizes all options in the choice set and therefore influences which one is chosen. As the lack of such a variable is associated with the dependent variable rather than the explanatory ones, it is not possible to rely on the aforementioned approaches. We apply a sequential imputation of missing data to deal with this problem. Hence, we first estimate an income imputation model, and then apply two approaches, the deterministic FP-approach and the probabilistic MC-approach to impute the missing variable, which is then used for estimating the bid functions in the context of the location bid-auction model. In the following, we describe both models in more detail.

3.1 Income imputation

In this study, each *Zensus* household needs to be classified into an income category to specify location bid functions differentiated by this agent attribute. For achieving that, we use an ordered regression model (McCullagh, 1980; Wooldridge, 2010) to estimate the income level for each household of the *Mikrozensus* dataset. Since all explanatory variables of this model are present in both databases (*Mikrozensus* and *Zensus*) we can apply this imputation model to finally impute income categories in the *Zensus* dataset.

The probability of belonging to a given income category is defined on the basis of a latent variable (y) taking the following linear expression:

$$y=x^{T} \gamma+\varsigma \qquad (1)$$

where x is a vector of explanatory variables, γ the vector of parameters to be estimated, and ς an error term, whose distribution depends on the assumptions for income. The income-class probabilities can then be expressed in the following manner:

$$\begin{aligned} &P(I=n|x;\gamma,\varsigma) \qquad (2)\\ &=P(\psi_{n-1}< x^{T} \gamma+\varsigma \leq \psi_{n})\\ &=P(\psi_{n-1}- x^{T} \gamma < \varsigma \leq \psi_{n}- x^{T} \gamma)\\ &=F_{\varsigma} (\psi_{n}- x^{T} \gamma) -F_{\varsigma} (\psi_{n-1}- x^{T} \gamma). \end{aligned}$$

Here, $P(n)$ indicates the probability of a household belonging to income class n and ψ are thresholds to be estimated. F_{ς} is the cumulative distribution function of ς. Assuming m different income levels, $\psi_{0}= -\infty$ and $\psi_{m}=\infty$, the intermediate thresholds increase monotonically. Depending on the specification of the error term ς, which is usually assumed to be either normally or logistically distributed, with mean zero and diagonal covariance matrix $\sum_{I}$, equation (2) will lead to an ordinal probit or ordinal logit framework, respectively. For the purpose of this work we will assume logistically distributed error terms.

The estimation of the bid model with *Zensus* data requires the income category for each household in the dataset. The parameters of each location model specification are estimated testing two alternative income imputation methods. FP predicts the income category as a point estimate reflecting income by the class with the highest probability, i.e., category = n if and only if $P(n) > P(m)\ \forall\ m \neq n$. Then, the first location model (referred to as FP-model) is estimated with a single vector of income category values resulting from the FP-approach. The second method, the MC approach, uses Monte Carlo simulation to assign for each household 100 times the income category based on the probabilities calculated from the ordered regression model. The second location model considers these 100 income values for each household to generate multiple models. The resulting model (MC-model) includes the average coefficient estimates of these 100 simulations.

3.2 Bid-choice model

The location model in our application is of the bid-auction type. The model is based on Ellickson (1981) hedonic formulation of households being assigned to houses. Martínez (1992) uses an extended approach of the aggregate logit version of Ellickson's model that considers dwelling types in zones as locations. The probability $P_{h|dz}$ that household category h is assigned to location d, z (dwelling type d in zone z) is defined as:

$$P_{h/dz} = \frac{H_h \exp(\mu B_{hdz})}{\sum_g H_g \exp(\mu B_{gdz})}$$

with g representing all household categories including h, H_h the number of households of category h in the population, scale parameter μ (set to 1 without loss of generality), and B_{hdz} the bid of household category h for location d,z. The bid is defined as a function of attributes that are assumed to explain residential location choice. Commonly applied attributes are household (X_k) and dwelling (D_l) characteristics as well as accessibility (A_m) and zonal (Z_n) indicators (Hurtubia, 2012; Hurtubia & Bierlaire, 2013; Hurtubia, Gallay, & Bierlaire, 2010; Schirmer et al., 2014). The linear-in-parameters bid function is thus defined as:

$$B_{hdz} = \beta_{0h} + \sum_k \beta_{hk} * X_{hk} + \sum_l \beta_{hl} * D_{dl} + \sum_m \beta_{hm} * A_{zm} + \sum_n \beta_{hn} * Z_{zn}$$

with attribute indices k, l, m and n. Betas differ by household category and attribute while attributes may also differ by dimension investigated (household type h, dwelling type d, or zone z). The bid function reflects a household category's willingness to pay for this type of location.

4 Results

The following sections describe the results of the model's application to empirical data, i.e., *Zensus* and *Mikrozensus*, first for the income imputation model, and then for the location models. After outlining the assumptions about the relations between dependent and independent variables, we show descriptive statistics. Then we turn to the models which at the end will also be compared against direct probabilities at the zonal level.

4.1 Income imputation model

Assuming that households determine their location bids differentiated by disposable income (after deducting taxes), we have to impute income for each of the *Zensus* households since it is not included. The imputation is based on an ordinal logit model that explains *Mikrozensus* categories for household net income by sociodemographic household variables included in both datasets. Net household income is codified in 24 categories, which is impractical for the estimation of a multinomial logit model since it would define an unmanageable number of alternatives and thus bid functions and parameters. For illustrative purposes, we aggregate these 24 categories into four, which represent the distribution of households across income in Berlin in 2011 quite well and correspond to typical national classifications (cp. Table 1).

Table 1: Income groups (2010 values)

Group	Range	Description	Relative frequency
1	below 900 €	Low income	17 %
2	900 to below 1,500 €	Lower medium income	28 %
3	1,500 to below 2,600 €	Upper medium income	32 %
4	2,600 € and above	High Income	23 %

(Source: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Mikrozensus, survey year 2010, own calculations; frequencies are unweighted)

Education, professional background, and professional experience are some of the most important variables explaining personal income (Baldemir, Ozkoc, Bakan, & Yesildag, 2012; Mincer, 1974). Unfortunately, corresponding variables are not included in *Zensus* and therefore cannot be used directly for imputation. Proxies for this variable are age and occupation of household members. What is more, household income obviously correlates with household size. Thus, we expect this variable to also have a considerable influence. Households with only one person should be associated with rather low household income and thus also help to classify the income level of these households. Taking into account experience proxies, the influence of the number of household members should vary by specific age groups related to life phase, such as children, students, or adults established in the workforce and pensioners which all have different degrees of education and experience and thus imply different income levels. A more intuitive variable would be the number of workers, since unemployed persons usually earn much less than employed ones.

However, applying variables related to the working situation in *Zensus* requires caution because it is a register-based data source that only identifies employment status for individuals registered with employment agencies and therefore ignores unregistered employment types such as self-employment. In practical terms this means that the "non-working" group actually consists of both unemployed and self-employed persons. This complicates the prediction of the lowest and highest income groups.

Table 2: Industry groups according to German WZ 2008 classification (Federal Statistical Office Germany, 2008) with Sections in brackets

Group	Sections included
0	"non-working"
1	agriculture (A), administrate activities in private sector (N), accommodation (I) and household-related services (T)
2	construction (F), wholesale and retail (G), transportation (H), health and social activities (Q) and other services (S)
3	manufacturing (C), water supply and waste (E), education (P) and arts (R)
4	mining (B), information and communication (J), science and professional services (M), public services (O) and extraterritorial activities (U)
5	electricity supply (D), finance and insurance (K)

Other variables that we expect to have an influence on income can be attributed to the household representative (cp. Chapter 2). Due to the different role of working associations, productivity and contracts, wages vary according to industry where a person is employed—and so should household income. However, since industry of employment comprises 21 categories, we aggregated them into five branches according to household income similarity, added by one category for (according to the data) "non-working" persons (cp. Table 2).

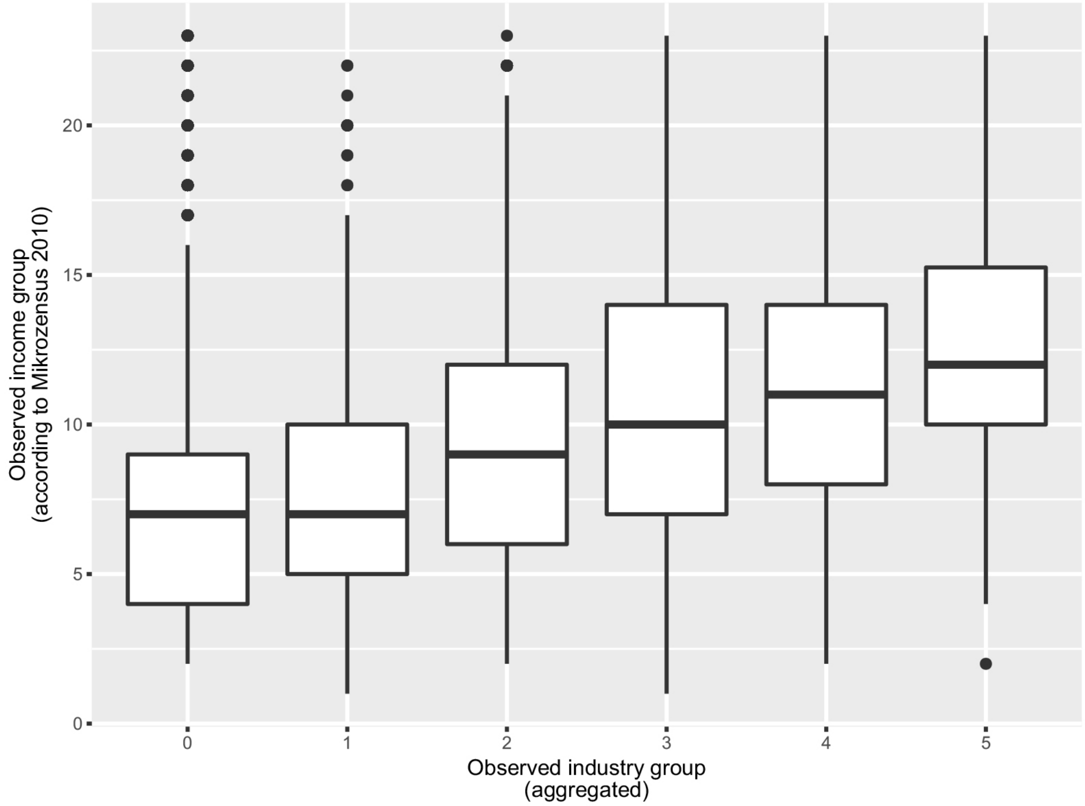


Figure 1: Distribution of 24 *Mikrozensus* income groups for six industry groups (including non-working). (Source: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Mikrozensus, survey year 2010, own calculations; frequencies are unweighted)

The association between the original 24 groups of *Mikrozensus* household net income and aggregated industry group of the household representative is shown in Figure 1. The boxplot confirms significant differences between branches. Including "non-working" household heads in Industry group 0, i.e., all household representatives for whom we cannot identify whether they are working, shows that this group has the lowest median income which is, however, not different from the income of Industry group 1. The large range between 0.25 and 0.75 quantiles and the high number of outliers indicate that in the "non-working" group we might also find households with high income, such as self-employed persons. Accordingly, we expect a coefficient close to zero for Industry group 1 and positive and increasing coefficients for the remaining industry groups. Given that there are many households with more than one employed member it is likely that the employment industry of the second household representative (second-oldest employed person) may also turn out as a significant variable to explain the income group of the household.

The cultural differentiation of households may also have an effect on income. We expect that migration background of the householder has a negative impact on income level, which could be due to lacking integration in society and facing disadvantages in compensation (Brenke, 2008). However, an analysis of migration background by country of origin may reveal differences but was beyond the scope of this paper.

Table 3: Descriptive statistics for continuous variables in the imputation model

Variable	Proportion or mean	$Q_{0.05}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.95}$	Standard deviation
Number of household members:							
below the age of 18	0.25	0	0	0	0	2	0.65
at the age of 18 until the age of 30	0.29	0	0	0	0	1	0.55
at the age of 31 until the age of 50	0.52	0	0	0	1	2	0.70
at the age of 51 until the age of 64	0.32	0	0	0	0	2	0.59
at the age of 65 and above	0.37	0	0	0	1	2	0.65

For a better understanding of the data and model results, in the following, we list descriptive statistics of variables to be included in the income imputation model. Mean household income in Berlin in 2010 was 1,525 € (Amt für Statistik Berlin-Brandenburg, 2011). Since we use the sample of *Mikrozensus* households, it may well be that low-income households are underrepresented there in comparison to the population. As for the employment of the household head, the distribution is the following: 49% of the household representatives are not employed according to census data. Of the remaining 51% of households 14% work in Industry group 1. 39% are employed in group 2, and 22% of the employed heads make their living in groups 3 and 4 respectively. Apparently, household representatives employed in Industry groups 1 and 5 are rare — the last group represents only 4% of all working household heads. 14% of all households have a second representative who is employed. The distribution of these persons across industries is almost the same when comparing it to the household head. Berlin is a multicultural city, as a considerable proportion of about 15% actually immigrated from elsewhere. Looking at the distribution of the number of persons in a household confirms that Berlin is the capital of singles in Germany, in 52% of the *Mikrozensus* households lives only one person.

Table 3 shows the descriptive statistics for continuous variables applied in the imputation model. There are few multi-person households involving at least one member between 18 and 30. The number of households by age shows that Berlin is a city with rather few young and many medium-aged and older households.

Table 4: Coefficients of the imputation model for net household income

Variable	**y**	**t-value**
Thresholds:		
1	0.543	*(4.74)*
2	2.61	*(22.64)*
3	5.06	*(41.47)*
Household representative is employed in:		
Industry group 1	0.579	*(8.52)*
Industry group 2	1.08	*(21.59)*
Industry group 3	1.84	*(29.11)*
Industry group 4	2.25	*(35.20)*
Industry group 5	2.93	*(20.08)*
Second household representative is employed in:		
Industry group 1	0.268	*(1.99)*
Industry group 2	0.956	*(11.47)*
Industry group 3	1.37	*(10.68)*
Industry group 4	1.93	*(13.35)*
Industry group 5	1.71	*(4.61)*
Household is a single-person household	-0.628	*(-9.45)*
Household representative has a migration background	-0.793	*(-16.12)*
Number of household members:		
below the age of 18	0.477	*(14.05)*
at the age of 18 until the age of 30	0.692	*(12.39)*
at the age of 31 until the age of 50	1.64	*(27.93)*
at the age of 51 until the age of 64	1.71	*(28.69)*
at the age of 65 and above	2.30	*(36.87)*
Number of observations 14,922		
Log-likelihood at convergence -15,124		

Coefficients are estimated using the *vgam* package in R (Yee, 2010)[4] applying the proportional odds assumption. Table 4 shows the results of the best [5] ordinal logit model. Thresholds are significantly different from zero and from each other indicating that the model splits groups sufficiently well. All coefficients show significance and their signs are negative for one-person households and heads with migration background only, as expected. Regarding household size and age groups, we find expected relative differences for number of children, and students. However, the coefficient for the number of household members in the age groups 31-50 and 51-64 do not seem to have a different effect, which questions the role of experience for explaining household income. Also noticeable is that the highest coefficient is related to the number of seniors in a household. The household representative's employment industry has

[4] For the actual application in the location model, we later re-estimate the model using *PythonBiogeme* yielding the same results (see Section 4.2).

[5] The best model has been identified by sequentially including additional parameters and applying each time the likelihood ratio test.

a very strong influence as has the industry group of the second employed person. Working in companies registered in the financial sector or electricity increases the odds for a higher income category much more than agriculture or arts, e.g., coefficients are generally higher for industries with higher wages, as we expected. The order of coefficients for employment industry of the first and second household representative differ remarkably for Industry group 5 suggesting that second household representatives earn less in this industry than household heads. In summary, the model has a sufficient number of significant coefficients with different levels and signs and can generally be used for imputing income.[6] Before including the imputation model in the location model, we also check the plausibility of the results when employing the model.

Applying the imputation model to census data enables assessing the suitability of the model in general and in terms of the two approaches, FP and MC, for estimating location choice. This check consists of the following steps: 1. Apply the imputation model to census data and predict the probability of each household belonging to an income group. 2. Aggregate probabilities by zone and income group calculating the share of each group in each zone. 3. Apply the FP-approach to deterministically and the MC-approach to probabilistically determine the income group of each household. 4. Compare resulting proportions at the zonal level.

[6] Other variables tested such as gender, dummies or the presence of a person in the different age groups were tested but did not turn out significant.

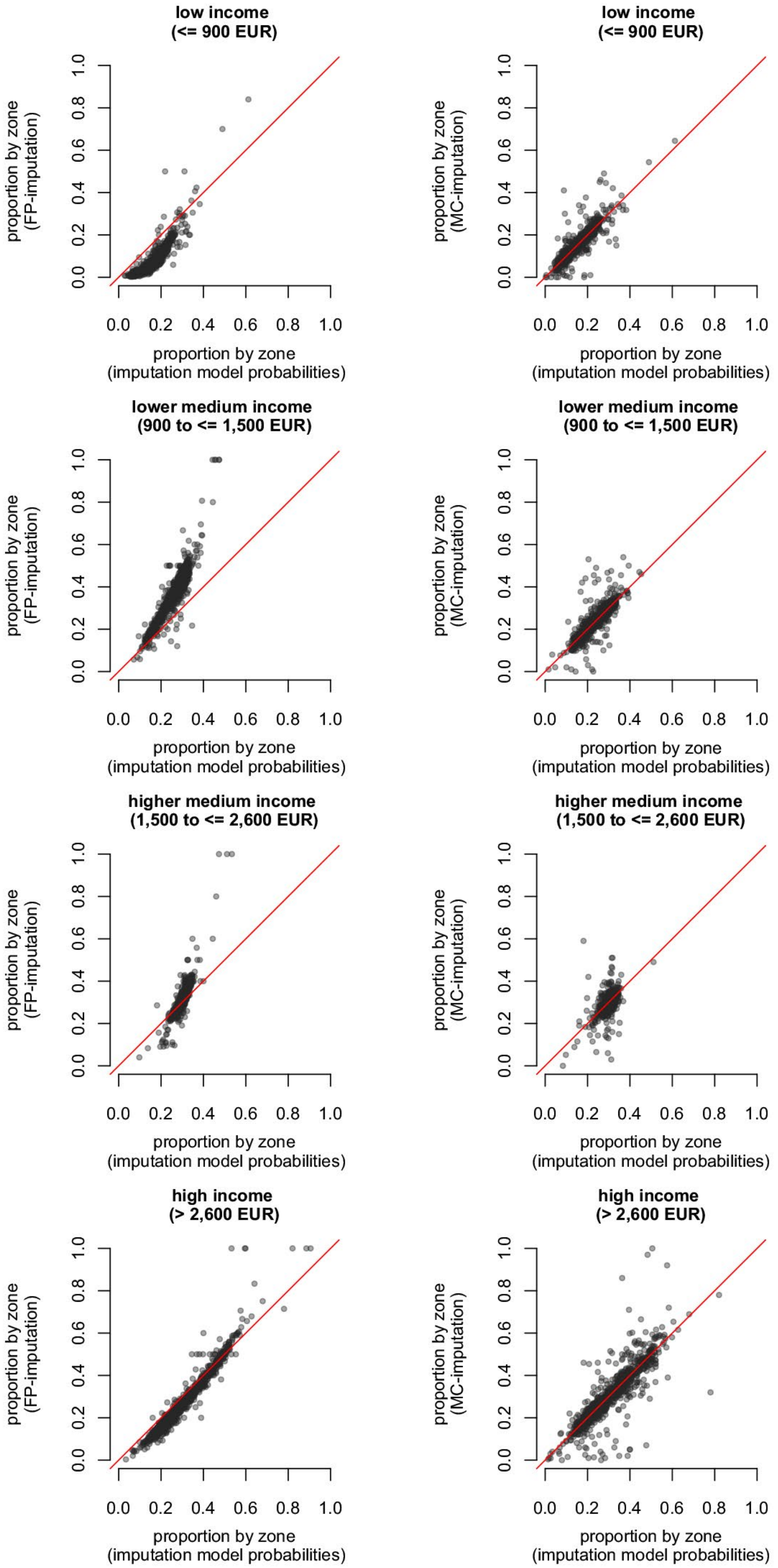


Figure 2: Validation plots of imputed income probabilities for each zone
(Sources: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Mikrozensus, survey year 2010, and RDC of the Federal Statistical Office and Statistical Offices of the Länder, Zensusgesamtdatenbestand Berlin, survey year 2011, own calculations)

The results of this analysis are shown in Figure 2. Predicted direct probabilities at the zonal level should not differ from those generated from applying the FP-approach or the MC-approach. Hence, in Figure 2, for a good fit, all dots should be distributed along the red line. Comparing both figures indicates that MP (right column) in aggregate terms better predicts income groups 2 and 3 in which the majority of households fall. FP is more suitable to predicting the groups at the margins and strongly overpredicts groups 2 and 3 since the approach favors categories with high probabilities. This suggests MC to be slightly superior when imputing income for location choice modeling.

4.2 Location model

In our case study, we apply the income imputation model to *Zensus* for Berlin, Germany, testing two imputation methods in location models with the same variables. This produces two location models with the same parameters but different coefficients. The location model predicts the probability of households belonging to a specific group to be located in a specific dwelling type and TAZ. Both approaches are tested using the same sample of 1% (i.e., 18,000) randomly selected *Zensus* households.

Before coming to the results, in the following we derive our expectations regarding the parameters of the location model based on location choice studies that analyze the influence of household income. Additionally, we consider descriptive statistics of variables included in the location sample.

Household income is associated in different ways to residential location preferences. According to Alonso's bid-rent theory (Alonso, 1964) and its extension by Muth (1969), land is generally cheaper farer away from the city center which is why households that have the choice are willing to trade off accessibility (or commuting cost) for land (cp. the discussion in Diamond, 1980). Regarding other location characteristics, urban economic theory suggests that preferences for positively perceived attributes increase with income, while those for negative ones decrease (Ellickson, 1981). A number of studies analyze the association between location choice and the socioeconomic structure of the neighborhood considering either the proportion of households of certain socioeconomic categories or the average income level. In general, authors find that households are located close to similar ones. De Palma, Motamedi, Picard, and Waddell (2005) show that households with low income tend to co-locate, while Zondag and Pieters (2005) find that zones with households of higher income attract all households in general, but in particular those with high income.

We define several dwelling, zonal and accessibility attributes to include in location models based on reviews of the studies cited in Chapter 1. Two variables characterize dwellings in our analysis: one dummy for single-family housing (one dwelling per building) and one quasi-metric variable for the floor space of dwellings in multifamily housing.[7] Several attributes describe the neighborhood. They have been calculated for the 1,223 transport analysis zones and include number of creative industry businesses per 1,000 inhabitants in the zone [8] and four variables corresponding to proportions of households in each income category calculated as the proportion of the sum of probabilities resulting from the ordered logit model for household income. Additionally, several accessibility measures are considered in the analysis including travel times to activity locations (train stations, grocery stores, city center, jobs) and cumulative opportunities measures as well as other local and regional accessibility indicators (Handy, 1992). All accessibility indicators are calculated based on the actual travel time between address-fine locations and averaged across all households within a zone (Krajzewicz & Heinrichs, 2016).

[7] This variable transforms the original *Zensus* floor space categories (given in 10-square-meter steps) to their midpoints. Since it is interacted with the single-family housing dummy, both need to be included at once in the bid function. We therefore assume that floor space in single-family houses has a rather uniform distribution and is mostly larger than in multifamily houses

[8] Creative industry businesses are defined as establishments of firms that belong to industry groups with divisions 91 through 93 according to WZ 2008 (Federal Statistical Office Germany, 2008), for instance theaters, museums, libraries, gambling halls, locations for sports and entertainment.

We test the following regional accessibility indicators:

- Travel time to city centers: Average car travel time to city centers East and West in units of 10 minutes.
- Travel time to next rail station: Travel time by public transport[9] (including ingress by walking[10]) to the next commuter rail train station in units of 10 minutes.
- Job accessibility (public transport): Number of jobs within 30 minutes travel time by public transport in units of 100,000 jobs.
- Job accessibility (public transport / foot): Number of jobs within 30 minutes travel time by public transport (including ingress by walking) in units of 100,000 jobs.
- Job accessibility (car): Number of jobs within 30 minutes car travel time in units of 100,000 jobs.

Other indicators relate to the access to local activities such as:

- Travel time to next grocery store: Walking time to closest grocery store of at least 200 m^2 sales area, in minutes.
- Bus lines: Number of bus lines at stops within 400 meters around address, averaged across the zone.
- Rail lines: Number of local rail and metro lines at stops within 2,000 meters and light rail lines within 400 meters around address, averaged across the zone.
- Public transport lines: Number of local rail and metro lines at stops within 2,000 meters and light rail and bus lines at stops within 400 meters around address, averaged across the zone.

While the association between household income and dwelling variables is straightforward—households with higher income can and do afford larger homes and coefficients for single-family housing and floor space are thus assumed to increase with household income—this association is not monotonous for other variables. According to Ellickson (1981), in general, positively perceived variables such as accessibility in terms of accessible jobs should show increasing coefficients, while the opposite applies to negatively perceived ones such as travel times to activity locations. Regarding accessibility and location choice there is a considerable body of literature, many of which suggests that accessibility in general has a rather low influence as compared to other location characteristics (Zondag & Pieters, 2005).

After deriving theory-based expectations about the location model, we now turn to the description of location patterns in the study area. Since 1920, Berlin has been a city with fixed borders (with the exception of the separation in East Berlin and West Berlin and some minor changes), which only grows internally.[11] In recent years, Berlin has taken up considerable pace in population growth and housing demand increased tremendously which is why development occurs internally, but also externally in suburban areas which, however, lie in the federal state of Brandenburg.

Because of the separation, the city has two main city centers, one in the eastern part of the city and one in the western part. In Berlin, like in many other cities, the more affluent neighborhoods which also feature larger (family) households can be found in the outer city. This implies a positive influence of income on travel times which has been analyzed and confirmed for several urban contexts including Germany (Gutiérrez-i-Puigarnau, Mulalic, & van Ommeren, 2016). For local accessibility, we do not expect to find significant coefficients since households by income are not concentrated in specific local centers.

[9] This includes bus, tram (light rail), S-Bahn (local rail), and U-Bahn (metro).

[10] We assume a walking speed of five kilometers per hour.

[11] One reason for that is that Berlin is at the same time a city, a municipality, and a federal state.

The following figures describe the structure of the sample regarding explanatory attributes (also see Table 5 for descriptives of metric variables). Most dwellings are located in multifamily housing; however, Berlin's housing stock also consists of a considerable proportion of single-family houses (8.5% of locations in the sample). An average dwelling in multifamily housing has about 63 m² floor space. This number is significantly higher in single-family housing (not shown here). The very majority of single-family houses is located at the city's outskirts (Senatsverwaltung für Stadtentwicklung und Umwelt Berlin, 2015; Senatsverwaltung für Stadtentwicklung und Wohnen Berlin, 2016a). Correlation analysis supports the assumed positive association between floor space and income category. Regarding zonal attributes, low-income households (considering imputed income) are the smallest group. The number of creative industry businesses varies strongly as their density is much higher in city centers. On average, Berliners travel 22 minutes by car to the city's main centers. In 30 minutes, they can reach 211,000 jobs using public transport, 122,000 jobs when considering walking ingress, and 583,000 jobs by car. People need to walk 13 minutes to get to the next supermarket of at least 200 m² size. Berlin has a very good transit network since, on average, sampled households have access to more than four bus lines in their neighborhood and 14 lines of public transport in total.

Table 5: Descriptive statistics for metric variables in the location model

Variable	Mean	$Q_{0.05}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.95}$	Standard deviation
Dwelling attributes:							
Size of dwelling in multifamily housing	63.27	0	4	65	7	115	31.72
Floor space category (10 m² steps)	5.79	2	45	5	75	12	3.05
Zonal indicators:							
Proportion of households (in population) within TAZ for:							
Income group 1	0.17	0.08	0.15	0.17	0.20	0.24	0.05
Income group 2	0.27	0.18	0.26	0.29	0.30	0.32	0.04
Income group 3	0.31	0.28	0.30	0.31	0.32	0.33	0.02
Income group 4	0.25	0.16	0.20	0.23	0.27	0.43	0.08
Number of creative industry businesses (by 1,000 persons)	1.54	0.30	0.62	1.07	1.82	3.70	2.80
Accessibility indicators:							
Travel time (car) to city centers in 10 minutes	2.22	1.11	1.53	2.03	2.88	3.71	0.84
Job accessibility (public transport) in 100,000 jobs	2.11	0.13	0.51	1.89	3.57	4.80	1.64
Job accessibility (public transport / foot) in 100,000 jobs	1.22	0.06	0.20	0.91	2.05	3.26	1.12
Job accessibility (car) in 100,000 jobs	5.83	1.70	4.34	6.76	7.46	7.86	2.03
Travel time to next grocery store in minutes	12.69	6.73	8.86	11.40	14.76	22.41	6.63
Travel time to next rail station in 10 minutes	2.60	1.53	2.02	2.40	2.99	4.27	0.95
Bus lines (number)	4.40	1	3	4	6	9	2.46
Rail lines (number)	9.53	1	4	9	14	20	6.33
Public transport lines (number)	13.93	4	9	13	19	25	6.33

(Source: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Zensusgesamtdatenbestand Berlin, survey year 2011, own calculations)

Coefficients for the imputation model and the FP and MC location models are estimated using *PythonBiogeme* (Bierlaire, 2003; Bierlaire & Fetiarison, 2009). Table 6 shows the results of the parameter estimations for the FP-model and the MC-model, respectively. For finding the specification that fits the data best, we include variables stepwise in the estimation according to the significance of likelihood-ratio tests. Within this process, coefficients for dwelling attributes, single-family housing dummy and floor space in multifamily housing yielded the best model.[12] Besides, including the proportion of high-income households within a neighborhood significantly increased the final log-likelihood. Subsequently testing several accessibility measures, we found that only one measure at a time improves the model significantly. For illustrative purposes we chose the indicator travel time to city centers by car.

Table 6: Bid-function coefficients for four household income groups comparing FP-model and MC-model

Variable	**FP-model**		**MC-model**	
	Coefficient / (*t-value**)		Coefficient / (*t-value**)	
Constant$_1$	0	*(fixed)*	0	*(fixed)*
Constant$_2$	-1.15	*(-6.62)*	-0.00149	*(-0.05)*
Constant$_3$	-3.67	*(-19.93)*	-0.876	*(-19.73)*
Constant$_4$	-6.06	*(-31.2)*	-2.830	*(-46.54)*
Dwelling attributes				
Floor space if multifamily housing$_1$	0	*(fixed)*	0	*(fixed)*
Floor space if multifamily housing$_2$	0.0263	*(11.43)*	0.00689	*(21.70)*
Floor space if multifamily housing$_3$	0.0548	*(22.63)*	0.0172	*(32.69)*
Floor space if multifamily housing$_4$	0.0691	*(27.55)*	0.0290	*(42.01)*
Single-family housing$_1$	0	*(fixed)*	0	*(fixed)*
Single-family housing$_2$	2.73	*(5.17)*	0.689	*(26.35)*
Single-family housing$_3$	5.24	*(9.92)*	1.70	*(30.2)*
Single-family housing$_4$	6.79	*(12.79)*	3.03	*(34.22)*
Zonal attributes				
Proportion of high-income households$_1$	0	*(fixed)*	0	*(fixed)*
Proportion of high-income households$_2$	3.25	*(4.64)*	0.855	*(7.58)*
Proportion of high-income households$_3$	3.72	*(5.18)*	1.78	*(9.72)*
Proportion of high-income households$_4$	6.05	*(8.17)*	3.20	*(12.64)*
Accessibility attributes				
Travel time to city centers$_1$	0	*(fixed)*	0	*(fixed)*
Travel time to city centers$_2$	0.216	*(4.85)*	0.0625	*(7.44)*
Travel time to city centers$_3$	0.321	*(6.91)*	0.124	*(8.63)*
Travel time to city centers$_4$	0.383	*(7.61)*	0.165	*(7.71)*
Number of observations	17,949		17,949	
Log-likelihood at zero	-24,883		-24,883	
Log-likelihood at convergence	-20,491		-22,937	

* This value corresponds to the robus t-statistic as indicated by *PythonBiogeme* (Bierlaire, 2016).

(Sources: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Mikrozensus, survey year 2010 and RDC of the Federal Statistical Office and Statistical Offices of the Länder, Zensusgesamtdatenbestand Berlin, survey year 2011, own calculations)

[12] Prices which are often included in hedonic models cannot be included here as the bid function itself represents the willingness to pay. In MUSSA / Cube Land, the simulations' results are prices and thus endogenous.

We now turn to the interpretation of the estimation results of the FP-model. Coefficients are all significant at 1% level and show a clear pattern across income. Ceteris paribus, adding one square meter of floor space in multifamily housing increases the willingness to pay of households as their incomes increase[13] which reflects their higher desire and ability to afford such dwellings. Living in single-family housing also differentiates household types by income with the same tendency. According to the FP-model, the higher its income compared to the lowest income group the more a household is willing to concentrate around households with the highest income level.

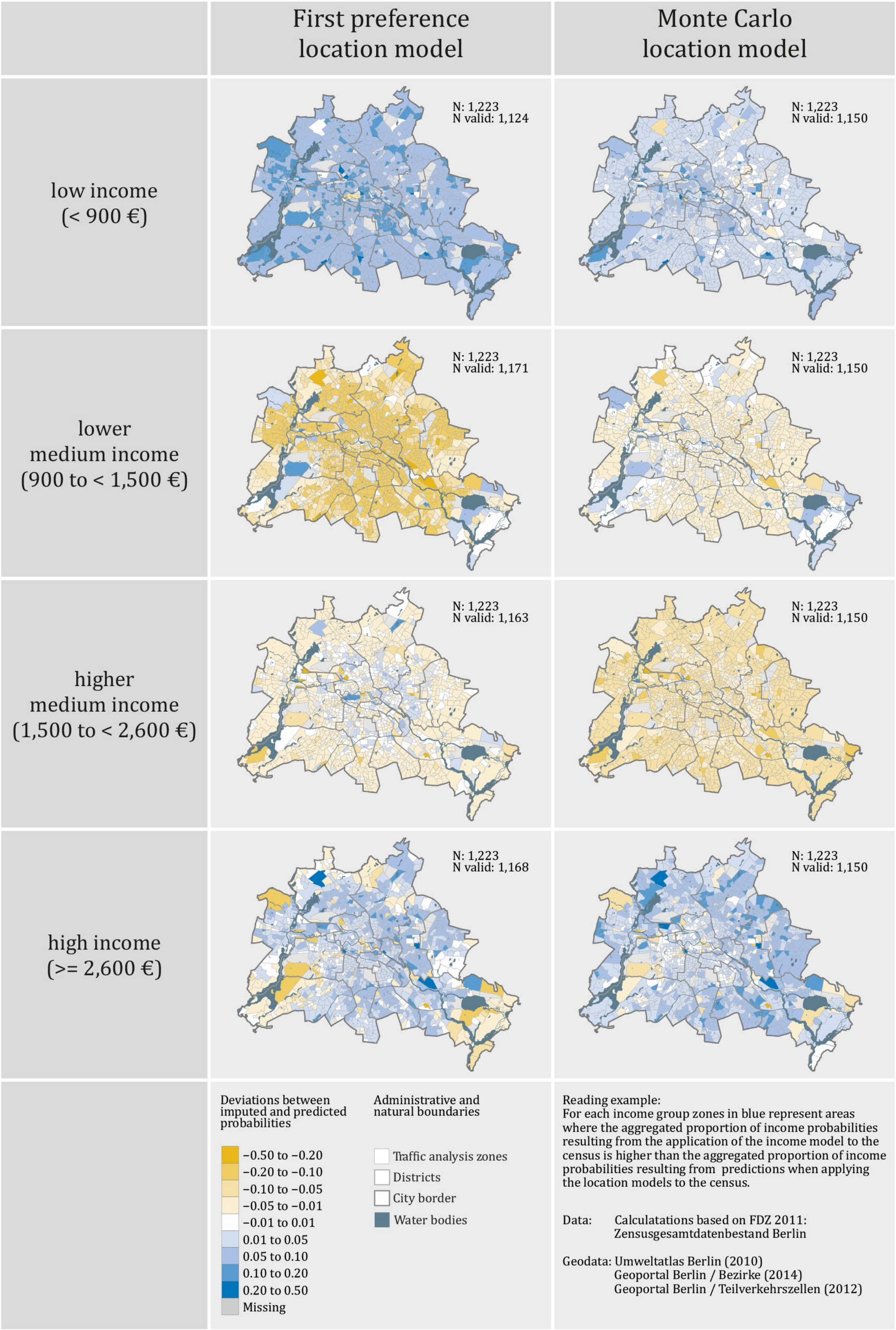
So far, all our assumptions are confirmed by the location model, e.g., the high positive coefficient for proportion of high-income households supports Zondag and Pieters (2005) findings that households cluster by income and Ellickson's theory that positively perceived variables are valued even more by households with high income.

To correctly interpret the results it is necessary to take the following into consideration. Since we specified linear-in-parameter willingness to pay functions, if all household coefficients of a real estate or zonal attribute are shifted by the same amount, the location probability does not change. Then, the differences between these parameters and not their absolute values (and signs) can be identified. Regarding travel time to city centers by car, the higher the household income the lower is the willingness to pay for a travel time reduction as expected from the spatial distribution of households by imputed income. If we add a constant of -0.383 to all accessibility coefficients, we observe that households with lower incomes income are willing to pay more for a travel time reduction than households with higher incomes. Hence, they seem to prefer to live in the city center. This location pattern may result from several causes. One is the budget constraint: poorer households cannot afford locations with higher travel costs or large or high-quality real estate. Another explanation refers to the structure of the city. Travel time to city centers reflects the spatial distribution of households and activities in the city since the street network of Berlin is similarly good everywhere in the city. We know from the *Zensus* that many of the larger households (which usually also dispose of higher household incomes) are rather found apart from the city center, at the outskirts, while less affluent households are mostly found in the inner city, except for some areas (vom Berge et al., 2014). This may indirectly reflect a combination of attributes that we have not tested, such as proximity to water, quietness, etc. Including such variables in a future study might yield a better picture of the location factors at play. Furthermore, knowing the actually available transport mode for each household might help to address this in more detail. A final explanation for this result is that households with high income are mostly family households who locate in single-family housing. As we could see, such a housing type is almost only available at the outskirts of the city. Thus, high income households may be willing to accept a higher distance to the city center in exchange for living in a single-family house in a family-friendly environment. This consideration explains the greater coefficient of households with higher income (and more household members) and confirms Alonso's and Muth's theories, and the empirical study by Gutiérrez-i-Puigarnau et al. (2016).

Now coming back to the question, which methodology achieves which result for the location model, we can see that the MC-model yields very similar relative effects but with significantly lower absolute coefficients. The reason for this is that the FP-model assumes that income is an observed and not an uncertain variable - which it actually is. Thus, goodness of fit of the MC-model which includes the uncertainty regarding household income is worse due to lower final log-likelihood. Actually, the MC-model reflects the true patterns better and is the correct model whereas the results of the FP-model are more subject to measurement noise and thus misleading.

As for the assessment of the imputation method (Section 4.1), location models can be evaluated using aggregated probabilities at the zonal level for each income group. Applying both models to census data, we determine location probabilities which we aggregate to zones. Figure 3 shows the difference between these values for each zone with positive deviations (underpredictions) in blue and negative ones (overpredictions) in yellow and the saturation reflecting the intensity.

[13] Note that the value of the coefficient in the bid function cannot be interpreted directly.



Reading example:
For each income group zones in blue represent areas where the aggregated proportion of income probabilities resulting from the application of the income model to the census is higher than the aggregated proportion of income probabilities resulting from predictions when applying the location models to the census.

Data: Calculatations based on FDZ 2011: Zensusgesamtdatenbestand Berlin

Geodata: Umweltatlas Berlin (2010)
Geoportal Berlin / Bezirke (2014)
Geoportal Berlin / Teilverkehrszellen (2012)

Figure 3: Deviations between aggregated zonal probabilities from the imputation model and aggregated zonal probabilities from location predictions

(Source: RDC of the Federal Statistical Office and Statistical Offices of the Länder, Zensusgesamtdatenbestand Berlin, survey year 2011, own calculations)

Maps for differences between aggregated probabilities and FP-model results reveal high deviations for income categories 1 and 2 in the magnitude of more than ± 10% for many zones. Applying the MC-model results in much lower deviations for most zones in income groups 1 and 2. For income group 3, however, deviations are higher. Finally, differences for the highest income group are similar. From these considerations we conclude that the MC-model yields more accurate results than the FP-model confirming what we already found out for the imputation itself (cp. Section 4.1). To conclude, our results show that imputation can be used to find significant estimates in bid functions for households differentiated by unobserved characteristics and if this approach is used Monte Carlo simulation should be included to determine the location model.

Conclusion

With increasing data availability, imputation of variables becomes more important to add attributes that are lacking e.g., due to confidentiality. We estimate bid-auction based location choice models for households with different income levels using German *Zensus 2011*. As *Zensus* does not contain household income, we impute this variable based on an ordered regression model calculated from observed income in *Mikrozensus 2010*. The two location models employ different imputation approaches: one with imputed income derived deterministically from the highest probability and one which assigns income categories probabilistically using Monte Carlo simulation.

We show that with each of our approaches it is possible to generate significant and plausible findings and thus imputation is an option to deal with the lack of an important choice variable in large data sources such as censuses. As our results demonstrate, models of imputed choice can perform well if other data sources exist that help to develop a high-performing imputation model. This methodology can be applied to many other situations where separate data sources include parts of the choice situation.

Comparing both approaches reveals that the MC-model has a lower goodness of fit than the FP-model. However, the main reason for that is that the latter "pretends" that the income group is observed, which actually is not true since this method underpredicts categories with low probabilities and overpredicts such with high probabilities. Furthermore, comparing aggregated probabilities resulting from the imputation model with the outputs of the imputation by FP and MC reveals higher accuracy of the latter method. This is confirmed by the same analysis considering predictions from respective location models.

Our methodology could be improved with respect to increasing the performance of both models by including further explanatory variables. Attributes such as the country of the origin of the household representative, or other specifications of the distributions of the error term, e.g., in an ordinal probit model, could improve the former. In the location models, measures that better reflect the different urban structures of inner and outer city such as other accessibility measures that consider different cost factors could achieve a better fit as well as including the resources of household types, and additional dwelling attributes, such as the quality or age of the house.

Another limitation is the definition of household groups. It is possible that other household types better reflect location patterns than income, such as household structure. What is more, the inclusion of household attributes or interactions between them and other attributes could improve the location models and should be considered in the future. Finally, upcoming studies that deal with data lacking important variables, could apply new models that simultaneously impute choice-relevant variables and estimate location choice rather than sequentially. Using direct probabilities instead of Monte Carlo simulation might eventually further reduce computational effort and increase accuracy of the results.

In summary, our study shows that imputation of choice variables in large data sources is a low-cost option that should be considered for analyzing decision situations with missing variables, including location choice modeling, particularly if a large data source such as a census is available.

Acknowledgements

We very much thank the team of the research data center at the Amt für Statistik Berlin-Brandenburg for their patient cooperation when providing data, executing our programs, and checking the results. This study has been conducted in the framework of the institutionally funded project "Transport and the Environment (VEU)" of the German Aerospace Center (DLR) which was funded by the Helmholtz Association of German Research Centres. We also thank two anonymous reviewers who heavily contributed to further improving the scientific quality of the article.

Data

Due to its very high degree of detail, microdata such as individual records from *Zensus 2011* and *Mikrozensus 2010* underlie strict confidentiality provisions according to the *Bundesstatistikgesetz*. The authors are thus not able to provide this data to the public.

References

Acheampong, R. A., & Silva, E. (2015). Land use–transport interaction modeling: A review of the literature and future research directions. *Journal of Transport and Land Use*, *8*(3), 11–38.

Alonso, W. (1964). *Location and land use. Toward a general theory of land rent.* Cambridge, UK: Harvard University Press.

Amt für Statistik Berlin-Brandenburg. (2011). E*rgebnisse des Mikrozensus im Land Berlin 2010. Haushalte, Familien und Lebensformen*. Retrieved from https://www.destatis.de/GPStatistik/servlets/MCRFileNodeServlet/BBHeft_derivate_00000294/SB_A1-11_j01-10_BE.pdf

Bahamonde-Birke, F. J., & Hanappi, T. (2016). The potential of electromobility in Austria: Evidence from hybrid choice models under the presence of unreported information. *Transportation Research Part A: Policy and Practice, 83*, 30–41.

Baldemir, E., Ozkoc, H., Bakan, H., & Yesildag, B. (2012). An application of ordered logit model and artificial neural networks in an income model. *Current Research Journal of Economic Theory, 4*(3), 77-82.

Ben-Akiva, M., & Bowman, J. (1998). Integration of an activity-based model system and a residential location model. *Urban Studies*, 35(7), 1131–1153. doi:10.1080/0042098984529

Bhat, C. R. (2015). A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. *Transportation Research Part B: Methodological, 79,* 161–188.

Bhat, C. R., & Guo, J. Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological, 41*(5), 506–526. doi:http://dx.doi.org/10.1016/j.trb.2005.12.005

Bierlaire, M. (2003). BIOGEME: *A free package for the estimation of discrete choice models.* Paper presented at the 3rd Swiss Transportation Research Conference, Monte Verità, Ascona, Switzerland.

Bierlaire, M., & Fetiarison, M. (2009). *Estimation of discrete choice models: Extending BIOGEME.* Paper presented at the Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland.

Brenke, K. (2008). Migranten in Berlin: Schlechte Jobchancen, geringe Einkommen, hohe Transferabhängigkeit. *Wochenbericht des DIW Berlin, 35,* 496–507.

Cordera, R., Ibeas, Á., dell'Olio, L., & Alonso, B. (2017). *Land use–transport interaction models.* Boca Raton, FL: CRC Press.

de la Barra, T. (1989). *Integrated land use and transport modeling.* Cambridge, UK: Cambridge University Press.

De Palma, A., Motamedi, K., Picard, N., & Waddell, P. (2005). A model of residential location choice with endogenous housing prices and traffic for the Paris region. *European Transport, 31,* 67–82.

DeSalvo, J. S., & Huq, M. (1996). Income, residential location, and mode choice. *Journal of Urban Economics*, *40*(1), 84–99.

Diamond Jr, D. B. (1980). Income and residential location: Muth revisited. *Urban Studies, 17*(1), 1–12.

Ellickson, B. (1981). An alternative test of the hedonic theory of housing markets. *Journal of Urban Economics*, *9,* 56–79.

Federal Statistical Office Germany. (2008). *German classification of economic activities.* Wiesbaden, Germany: Federal Statistics Office Germany.

Guo, J., & Bhat, C. (2004). Modifiable areal units: Problem or perception in modeling of residential location choice? *Transportation Research Record, 1898*(1), 138–147.

Gutiérrez-i-Puigarnau, E., Mulalic, I., & van Ommeren, J. N. (2016). Do rich households live farther away from their workplaces? *Journal of Economic Geography, 16*(1), 177–201.

Handy, S. L. (1992). Regional versus local accessibility: Neo-traditional development and its implications for non-work travel. *Built Environment, 18*(4), 253–267.

Heldt, B., Gade, K., & Heinrichs, D. (2014). *Challenges of data requirements for modelling residential location choice: The case of Berlin, Germany.* Paper presented at the European Transport Conference, Frankfurt am Main, Germany.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques.* Wiesbaden: Springer Science and Business Media.

Hunt, J. D., Kriger, D. S., & Miller, E. J. (2005). Current operational urban land-use-transport modelling frameworks: A review. *Transport Reviews, 25*(3), 329–376.

Hurtubia, R. (2012). *Discrete choice and microsimulation methods for agent-based land use modeling.* (PhD), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Hurtubia, R., & Bierlaire, M. (2013). Estimation of bid functions for location choice and price modeling with a latent variable approach. *Networks and Spatial Economics, 14*(1), 47–65. doi:doi.org/10.1007/s11067-013-9200-z

Hurtubia, R., Gallay, O., & Bierlaire, M. (2010). Attributes of household, locations and real-estate markets for land-use modeling, *SustainCity Deliverable 2.7*. Lausanne, Switzerland: EPFL Lausanne.

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature, 22*(4), 323–340.

Krajzewicz, D., & Heinrichs, D. (2016). UrMo accessibility computer – a tool for computing contour accessibility measures. *Proceedings of the SIMUL 2016 conference*, Rome, Aug. 21-25.

LeRoy, S. F., & Sonstelie, J. (1983). Paradise lost and regained: Transportation innovation, income, and residential location. *Journal of Urban Economics, 13*(1), 67–89.

Martínez, F. (1992). The bid-choice land-use model: An integrated economic framework. *Environment and Planning A, 24,* 871–885.

Martínez, F. (1995). Access: The transport-land use economic link. *Transportation Research Part B, 29*(6), 457–470.

Martínez, F. (1996). MUSSA: Land-use model for Santiago City. *Transportation Research Record: Journal of the Transportation Research Board, 1552*(1), 126–134.

Martínez, F., & Donoso, P. (2010). The MUSSA II land use auction equilibrium model. In F. Pagliara, J. Preston, & D. Simmonds (Eds.), *Residential location choice* (pp. 99–113). New York: Springer.

Martínez, F., & Henriquez, R. (2007). A random bidding and supply land use equilibrium model. *Transportation Research Part B: Methodological, 41*(6), 632–651.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological), 42*(2), 109–142.

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars, & J. Weibull (Eds.), *Spatial interaction theory and planning models* (pp. 75–96). Amsterdam: North Holland Publishing.

Mincer, J. (1974). Schooling, experience, and earnings. *Human behavior and social institutions, No. 2.* Cambridge, MA: National Bureau of Economic Research.

Moeckel, R. (2018). Integrated transportation and land-use models. A synthesis of highway practice. (Synthesis 520). Retrieved from http://www.trb.org/Main/Blurbs/177870.aspx

Muth, R. (1969). C*ities and housing: The spatial pattern of urban residential land use.* Chicago, IL: University of Chicago Press.

OpenStreetMap contributors. (2016). OpenStreetMap. Retrieved from http://www.openstreetmap.org/

Ortúzar, J. D., & Willumsen, L. G. (2011). *Modelling transport.* Chichester, West Sussex, UK: John Wiley & Sons.

Paleti, R., Bhat, C., & Pendyala, R. (2013). Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record: Journal of the Transportation Research Board, 2382,* 162–172.

RDC of the Federal Statistical Office and Statistical Offices of the Länder. (2015a). Mikrozensus, survey year 2010.

RDC of the Federal Statistical Office and Statistical Offices of the Länder (2015b). Zensusgesamtdatenbestand Berlin, survey year 2011.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley and Sons.

Rubin, D. B., & Little, R. J. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley & Sons.

Sanko, N., Hess, S., Dumont, J., & Daly, A. (2014). Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of Choice Modelling, 12,* 47–57.

Schirmer, P. M., Van Eggermond, M. A. B., & Axhausen, K. W. (2014). The role of location in residential location choice models: A review of literature. J*ournal of Transport and Land Use, 7*(2), 3–21. doi:10.5198/jtlu.v7i2.740

Senatsverwaltung für Stadtentwicklung und Umwelt Berlin. (2012). *Teilverkehrszellen Berlin.* Retrieved from: http://www.stadtentwicklung.berlin.de/verkehr/datengrundlagen/verkehrszellen/

Senatsverwaltung für Stadtentwicklung und Umwelt Berlin. (2015). *Flächennutzung und Stadtstruktur. Dokumentation der Kartiereinheiten und Aktualsierung des Datenbestandes 2015.* Retrieved from https://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/download/Nutzungen_Stadtstruktur_2015.pdf

Senatsverwaltung für Stadtentwicklung und Wohnen Berlin. (2016a). *Berlin environmental atlas.* Urban structure - area types differentiated. Retrieved from https://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/ede607_05.htm

Senatsverwaltung für Stadtentwicklung und Wohnen Berlin. (2016b). *Berlin environmental atlas.* Retrieved from: http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/edua_index.shtml

Timmermans, H. (2003). *The saga of integrated land use-transport modeling: How many more dreams before we wake up?* Paper presented at the 10th International Conference on Travel Behavior Research, Lucerne, Switzerland.

vom Berge, P., Schanne, N., Schild, C.-J., Trübswetter, P., Petrovic, A., & Wurdack, A. (2014). *Eine räumliche Analyse für Deutschland: Wie sich Menschen mit niedrigen Löhnen in Großstädten verteilen.*

Wegener, M. (2011). From macro to micro—how much micro is too much? *Transport Reviews*, *31*(2), 161–177.

Wegener, M. (2014). Land-use transport interaction models. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of regional science, volume 2* (pp. 741–758). Berlin: Springer-Verlag.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Zondag, B., de Bok, M., Geurs, K. T., & Molenwijk, E. (2015). Accessibility modeling and evaluation: The TIGRIS XL land-use and transport interaction model for the Netherlands. *Computers, Environment and Urban Systems*, 49, 115–125.

Zondag, B., & Pieters, M. (2005). *Influence of accessibility on residential location choice.* Paper presented at the 84th Annual Meeting of the Transportation Research Board, Washington, DC.