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

Benjamin Heldt [a, *]
Francisco Bahamonde-Birke [a]
Pedro Donoso [b]
Dirk Heinrichs [a]

* corresponding author, email: benjamin.heldt@dlr.de

a DLR Institute of Transport Research
b Universidad de Chile - Santiago, Departamento de Ingenieria Civil

Abstract

Residential location is a key component of the land-use system and modelling residential location is essential to understand 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 the choice options in a location model. For estimating bid-auction location models for household income groups and comparing them to a similar model for household size, we impute household income in census data by ordered regression. We find that the income model performs sufficiently well as it reveals reasonable aspects of the location patterns that the size model does not cover. Imputing choice variables should thus be considered in the estimation of residential location models.

Submitted for the WSTLUR 2017 Conference, Brisbane.

1. Introduction

Being essential for understanding land-use–transport interactions and support planning decisions, residential location models are rooted in a long-standing history (for overviews see e.g., Acheampong and Silva, 2015, Wegener, 2014, Wegener, 2011, Iacono et al., 2008, Hunt et al., 2005, Wegener, 2004, Timmermans, 2003, De la Barra, 1989) and complement transport models by incorporating accessibility measures as factors influencing location choice (Martínez, 1995, Wegener, 2014, Ortúzar and Willumsen, 2011).

Most scholars identify three different strands of research: spatial interaction models, econometric models, and microsimulation models (De la Barra, 1989)[1]. One of the recently most-applied econometric models is the bid-choice model (Martínez, 1992, Martínez, 1996, Martínez and Henriquez, 2007, Martínez and Donoso, 2010). By applying Ellickson's (1981) bid-auction approach, the framework combines discrete-choice (McFadden, 1978) and bid-rent theory (Alonso, 1964), which were shown to be equivalent and complementary under equilibrium conditions (Martínez, 1992). The relevant difference to utility-based discrete choice models is the inverse relationship between options and choosers. While in classical location models households are choosers, in the bid-auction model they are options; simply speaking owners of locations choose households.

The model assumes that an equilibrium state can be attained between the demand for and the supply of residential locations. In the framework, all agents behave rationally and are located somewhere within the city – an equilibrium state is reached. Households or firms are located in those real estate options that provide them maximum utility and for which they are the highest bidders. On the other side, suppliers offer location options to maximize their profits, which are rents deducted by real estate costs. Following the bid-auction approach rents are endogenous values, calculated as maximum bids for each location option. Demand for

[1] For a detailed explanation of these types of models please refer to the studies mentioned above.

residential locations in the bid-auction form is thus described by a set of bid functions, each of which reflects the preferences of the corresponding group of households, which is defined by characteristics such as income or number of persons. Revealed location preferences of each household type are determined by estimating the parameters of a multinomial logit model. Location models are very data-demanding and their reliability depends on the data used for parameter estimation (Heldt, Gade, & Heinrichs, 2014). Although 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 spatial detail. 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 define household groups. Depending on the type of residential location model, this implies that information for either classifying the choosers (in classical utility-based models) or the alternatives (in bid-auction models) is missing.

In the following paper, we show our approach to estimating and comparing two location models for Berlin, Germany: one based on imputed income data and the other based on (observed) household size. 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 imputation, and a multinomial logit model for discrete location choice. In the final sections we discuss the results of the imputation model and the comparison of both location models.

2. Data sources

2.1. Mikrozensus 2010

Mikrozensus is a Germany-wide survey carried out annually in order to provide the administration with main statistical numbers. Every four years, additional information is garnered on different topics, including housing. *Mikrozensus* information for Berlin (FDZ, 2010) is geocoded at the coarse level of twelve districts, which is due to a small sample size and thus cannot be used to estimate location bid functions. However, *Mikrozensus* includes net household income and rent information, and hence is very useful as auxiliary data.

2.2. Zensus 2011

In 2011, the German Federal Statistical Office conducted a nationwide "register-based" census of population and buildings and dwellings (FDZ, 2011). Instead of directly surveying persons and households, information was fused from several administrative registers, such as the population register, registers of employment agencies, etc. The total census for Berlin includes 3.3 million observations of persons and 1.8 million households, and can therefore be assigned to a spatial reference system with relatively high resolution, such as traffic analysis zones (SenStadtUm, 2012).[2] For defining a neighborhood, this is sufficiently small and allows the inclusion of detailed spatial characteristics and accessibility measures. *Zensus* 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 models.

2.3. Accessibility and spatial indicators

[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.

Accessibility and spatial attributes were calculated from a number of different data sources, including OpenStreetMap network data (OpenStreetMap contributors, 2016) for the computation of accessibility measures, a survey of retail establishments conducted by the Senate Department for Urban Development and the Environment, Berlin, land-use data from the Senate's Environmental Atlas (SenStadtUm, 2016), and activity locations from commercial data providers. [3]

3. Methodology

As the aforementioned data source that will be used to estimate residential location (*Zensus 2011*) lacks income information, it is necessary to rely on an imputation process. This way, income categories will be imputed for this dataset on the basis of the information collected in *Mikrozensus 2010*.

Several approaches, including record linkage, multiple regression imputation, etc., have been developed during the last years to deal with missing data (Rubin, 1987, Rubin and Little, 2002, Herzog et al., 2007). 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 (Sanko et al., 2014, (Bahamonde-Birke & Hanappi, 2016)), which jointly consider the imputation model and the objective function (which includes the missing variable).

However, as in the case of bid-auction models, income defines the dependent variable; it is thus not possible to rely on the aforementioned approaches. For that reason we apply a sequential imputation of missing data. Hence, we first estimate an income imputation model,

[3] The computation of the most important of these attributes is given in the following sections; computation of the remaining attributes can be retrieved from the authors.

imputing the missing variable as a point estimate in the census data, which is then used for estimating the bid functions in the context of the location choice model.

3.1. Income imputation

In this study, we impute income categories on the basis of an ordered regression model (cp. (McCullagh, 1980; Wooldridge, 2010). Using categorized data is required as bid-auction models rely on discrete variables. The estimation of the ordered model is based on information contained in *Mikrozensus 2010* and includes variables present in both *Mikrozensus* and *Zensus*. The resulting imputation model is applied to census data in order to predict income probabilities. The class with the highest probability is assumed to be the predicted income category.

The probability of belonging to a given income category is then defined on the basis of an auxiliary function (*Y*) taking the following shape:

$$Y = X \cdot \gamma + \varsigma \quad , \tag{1}$$

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

$$\begin{aligned} &P(I = n \mid X;\gamma,\varsigma) = P(\psi_{n-1} < \gamma \cdot X \le \psi_n \mid X;\gamma,\varsigma) \\ &= P(\psi_n < \gamma \cdot X \mid X;\gamma,\varsigma) - P(\psi_{n-1} < \gamma \cdot X \mid X;\gamma,\varsigma) \end{aligned} \tag{2}$$

Here, *P(I=n)* indicates the probability of an individual belonging to income class *n* and Ψ are thresholds to be estimated. Assuming *m* different income levels, $\Psi_0 = -\infty$ and Ψ_m

$= \infty$, with the intermediate thresholds increasing 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 Σ_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.

3.2. Bid-choice model

The location model in our example is of the bid-auction type. The model is based on Ellickson's (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 type h is assigned to location (d, z – dwelling type d in zone z) is defined as:

$$P_{(h|dz)} = \frac{H_h \exp(\mu B_{hdz})}{\sum_g H_g \exp(\mu B_{dz})} \quad (3)$$

with g representing all household types including h, H_h the number of households of type h in the population, scale parameter μ (set to 1 without loss of generality), and B_{hdz} the bid of household type 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 dwelling (D_l) and household (X_k) characteristics as well as accessibility (A_m) and zonal (Z_n) indicators (Hurtubia and Bierlaire, 2013, Hurtubia, 2012, Hurtubia et al., 2010, Schirmer et al., 2014). The linear-in-parameters bid function is thus defined as:

$$B_{hvi} = \beta_{0h} + \sum_k \beta_{h,k} * X_{h,k} + \sum_l \beta_{h,l} * D_{z,l} + \sum_m \beta_{h,m} * A_{z,m} + \sum_n \beta_{h,n} * Z_{z,n} \quad (4)$$

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

4. Results

4.1. Income imputation model

Assuming that households choose among residential locations depending on their disposable income (after deducting taxes), we explain *Mikrozensus* net household income categories by means of an ordinal logit model which we use later to impute income in the *Zensus*.

Among the most important variables to explain personal income are education and professional background or experience (Baldemir et al., 2012, Mincer, 1974). Unfortunately, corresponding variables are not included in the *Zensus* and therefore cannot be used directly for imputation. Since household income obviously correlates with household size, we expect this variable to have a considerable influence, which should be different when considering specific age groups related to life phase, such as children, students, or adults established in the workforce and pensioners, who all have different degrees of experience and thus different income levels. A more intuitive variable would be the number of workers, since unemployed persons usually only receive social welfare and therefore earn much less than employed ones[4].

Other variables can be attributed to the person representing the household, who is also assumed to decide where to move. In our model, the household's head is either the oldest employed person if one or more members are working, or the oldest person if no one is working. We expect several characteristics of the household representative to have an

[4] Applying this variable in the *Zensus* requires caution because the *Zensus* is a register-based data source that only includes individuals registered with employment agencies and therefore neglects self-employed persons.

influence on income level. Wages differ according to industry and so should household income. Indeed, cross-tabulating income with industry where the household representative is employed yields significant differences. The sex of the household representative may also play an important role, as men still receive higher wages than women in general (Busch and Holst, 2013). Additionally, migration background is expected to have a negative impact on income level, which could be due to lacking integration in society and facing disadvantages in compensation (Brenke, 2008).

We tested several specifications including interactions between gender and household size which did not prove significant. Testing different number of classes ranging from four to seven, we aggregate adjacent groups resulting in four categories. Another aggregation is necessary for 21 employment industry dummies, which we consolidate to five branches according to income similarity.

The final model is specified with the following variables:

Table 1: Income and industry groups

Income groups	
1	below 900 €
2	900 to below 1,500 €
3	1,500 to below 2,600 €
4	2,600 € and above

Industry groups	
Industry categories according to German WZ 2008 classification (in brackets) aggregated by coefficients in disaggregated models:	
0	unemployed
1	agriculture (A), administrate activities in private sector (N), accommodation (I), household-related services (T)
2	construction (F), wholesale and retail (G), transportation (H), health and social activities (Q), other services (S)
3	manufacturing (C), water supply and waste (E), education (P), arts (R)
4	mining (B), information and communication (J), science and professional services (M), public services (O), extraterritorial activities (U)
5	electricity supply (D), finance and insurance (K)

Coefficients are estimated using the *vgam* package in R (Yee, 2010) applying the proportional odds assumption. Table 2 shows the results of the ordinal logit model. Thresholds are as expected and all coefficients are significant and their signs are negative for one-person households and heads with migration background only, as expected. There is one remarkable exception. Gender does not seem to have a significant influence on household income level, disproving our expectations. Regarding household size and age groups, we find expected relative differences for number of children, and students. However, the number of working household members differentiated by age does not seem to have a different effect, which questions the role of experience. Also noticeable is that the highest coefficient is related to the

number of seniors in a household. The household representative's employment industry has a very strong influence. 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. In summary, the model has a sufficient number of significant coefficients with different levels and signs and can thus be used for imputing income.

Table 2: Coefficients of the imputation model for net household income

Variable	β	t
Threshold 1	-0.631	*(-5.430)*
Threshold 2	-2.704	*(-23.225)*
Threshold 3	-5.089	*(-41.448)*
Household representative is employed in industry group 1	0.651	*(9.675)*
Household representative is employed in industry group 2	1.190	*(24.362)*
Household representative is employed in industry group 3	1.960	*(32.201)*
Household representative is employed in industry group 4	2.366	*(38.174)*
Household representative is employed in industry group 5	2.998	*(21.128)*
Household is a single-person household	-0.723	*(-10.741)*
Household representative is male	0.048	*(1.486)*
Household representative has a migration background	-0.837	*(-17.606)*
Number of household members below the age of 18	0.438	*(13.152)*
Number of household members at the age of 18 until the age 30	0.867	*(15.632)*
Number of household members at the age of 31 until the age 50	1.820	*(31.046)*
Number of household members at the age of 51 until the age 64	1.835	*(30.623)*
Number of household members at the age of 65 and above	2.363	*(31.173)*
Number of observations	15,046	
Log-likelihood at Convergence	-15,473	
Log-likelihood at Zero	-20,465	

4.2 Location models

In our case study, we compare two location models, one where the choice is a household of an imputed income level (*income model*) and one where observed household size defines the choice (*size model*). Before coming to the results, in the following we describe our variables and derive our expectations regarding specification based on several location choice studies including household size or income.

We define several dwelling, zonal and accessibility attributes to include in location models based on reviews of the studies cited in Section 1. One group of variables frequently cited in studies on the association between built environment and transport are the *D's* (density, diversity, design, etc.,see Ewing and Cervero (2010), Cervero and Kockelman (1997)). We assume that some of these variables also have an influence on residential location choice. In particular, we include household type densities and a land-use entropy index (Cervero and Kockelman, 1997) of the Shannon form that is calculated as follows:

$$(-\sum_{k} p_k * \ln(p_k)) * \frac{1}{\ln(K)}, \quad (5)$$

where p_k are proportions of square meters of land use categories *k* (residential, commercial, public use, parks and recreation areas, water, and industrial use); and K is the number of land use categories. The resulting normalized land-use entropy index is in the range between 0 and 1 where 0 means homogenous (only one land use) and 1 equally mixed. Dwelling variables included in our analysis consist of two indicators of dwelling size, one for the number of rooms, and one for the actual size in square meters. This value was calculated from a variable that categorizes floor space of dwellings in 10-square-meter steps by computing the midpoint of these categories. Zonal attributes include school propensity, leisure facility density and proportion of waterfront area. Several accessibility measures were included in the analysis, in

particular walking accessibility to groceries which is the grocery store floor space that can be reached from a location within 10 minutes walking (Heldt, Gade, & Heinrichs, 2016).[5]

Income is a crucial variable to differentiate households as it defines available resources and thus affordable locations. Many studies of residential location choice include income as a variable, either to segment choosers or define location characteristics or both (see the summary by Schirmer et al. (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 et al., 2015). 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. Income is assumed to not only be related to household size but also 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 mobility and transport.

According to urban economic theory, preferences for positively perceived location characteristics increase with income, while those for negative ones decrease (Ellickson, 1981). While the association with dwelling variables is straightforward – households with higher income can afford larger homes – it is not so clear for other variables. According to Ellickson, positive variables school density, land use entropy and accessibility of grocery store floor space should show positive coefficients, although this may differ by household type. For concentration, De Palma et al. (2005) show that households with low income tend to cluster.

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

Another important variable for the explanation of household location is household size, which was addressed by Rossi (1955), who found associations between location and household life cycle operationalized by number of household members. According to his study, households of different sizes have different location preferences, particularly in terms of dwelling size, which was affirmed by other empirical studies (Lee and Waddell, 2010). Furthermore, we assume that concentration plays an important role for household size as well, i.e., one-person households cluster, as do three- and four-person households. In total, we assume that the coefficients for household size as compared to income do not show a similarly clear pattern regarding magnitude, direction and signs of positively perceived as opposed to negatively perceived attributes.

Coefficients are estimated using *Biogeme* (Bierlaire, 2003). Tables 3 and 4 show the results of the parameter estimations for the *size model* and the *income model*, respectively. For identification of the specification that fits the data best, we include variables stepwise in the *size model* according to the significance of likelihood-ratio tests. Within this process, coefficients for age of household head proved insignificant, as did those for proportion of water frontline, density of leisure activities, and several accessibility measures to name but a few.

We now turn to each of the models separately before comparing them qualitatively. Coefficients are all significant at 1 % in the *size model*; the estimate for the proportion of single-person households is lower than the others, but still highly significant. Regarding accessibility to groceries, each extra hectare of grocery store floor space is associated with a higher willingness-to-pay, rising with household size. Estimators for dwelling-size attributes increase as expected with number of household members. Having a look at zonal attributes, including the zonal proportion of the household type as index of spatial concentration, .

Table 3: Estimation results of the observed-choice - location model household size

Variable	Households with a household size of							
	1 person		2 persons		3 persons		4 and more persons	
	β	*t*	*β*	*t*	*β*	*t*	*β*	*t*
Constant	*Reference (0)*		-3.920	*(-53.340)*	-4.780	*(-68.750)*	-6.170	*(-92.660)*
Floor space of grocery stores accessible in 10 min walking (ha)	*Reference*		0.113	*(4.400)*	0.276	*(7.750)*	0.582	*(15.320)*
Number of rooms in the dwelling	*Reference*		0.396	*(48.920)*	0.589	*(53.290)*	0.831	*(67.490)*
Dwelling size (m²)	*Reference*		0.021	*(58.880)*	0.026	*(58.040)*	0.029	*(59.850)*
Proportion of household type 1 in the zone	0.207	*(2.710)*	-		-		-	
Proportion of household type 2 in the zone	-		3.680	*(29.400)*	-		-	
Proportion of household type 3 in the zone	-		-		6.630	*(26.620)*	-	
Proportion of household type 4 in the zone	-		-		-		6.710	*(36.410)*
Number of schools per 1,000 persons	*Reference*		-0.101	*(-8.960)*	-0.228	*(-12.850)*	-0.265	*(-13.400)*
Land use entropy	*Reference*		0.205	*(8.040)*	0.342	*(9.380)*	0.474	*(11.590)*

Number of observations	179,500
Final log-likelihood	-182,117
Null log-likelihood	-248,830

Table 4: Estimation results of the imputed-choice – location model for household net income

Variable	Households with an income of							
	less than 900 €		900 to 1,499 €		1,500 to 2,600 €		2,600 € or more	
	β	*t*	*β*	*t*	*β*	*t*	*β*	*t*
Constant	*Reference (0)*		-0.771	*(-13.800)*	-3.280	*(-45.160)*	-4.790	*(-85.870)*
Floor space of grocery stores accessible in 10 min walking (ha)	*Reference*		-0.107	*(-3.110)*	-0.015	*(-0.400)*	0.007	*(0.170)*
Number of rooms in the dwelling	*Reference*		0.198	*(15.180)*	0.502	*(37.200)*	0.711	*(48.940)*
Dwelling size (m²)	*Reference*		0.012	*(17.790)*	0.029	*(42.760)*	0.038	*(54.110)*
Proportion of household type 1 in the zone	7.100	*(45.220)*	-		-		-	
Proportion of household type 2 in the zone	-		0.940	*(10.080)*	-		-	
Proportion of household type 3 in the zone	-		-		1.860	*(11.750)*	-	
Proportion of household type 4 in the zone	-		-		-		2.080	*(26.110)*
Number of schools per 1,000 persons	*Reference*		-0.029	*(-2.380)*	-0.070	*(-5.250)*	-0.137	*(-8.390)*
Land use entropy	*Reference*		0.042	*(1.130)*	0.121	*(3.100)*	0.267	*(6.110)*

Number of observations	179,500
Final log-likelihood	-204,056
Null log-likelihood	-248,830

increases the model's explanatory power. Except for those with only one person, households appear to have a higher willingness to pay for living where similar households live. The direction of the indicator for school propensity differs from what we expected. It appears that larger households have a lower willingness to pay for zones with a high school propensity than smaller households. A possible explanation for that is that larger households are formed mostly by people that do not go to schools or the quality of schools is of higher importance for family households than their quantity. Coefficients for the final indicator, land-use entropy, also increase with household size, implying that an increase from 0 to 1, i.e. from homogenous to mixed zones, is associated with higher willingness to pay if there are more members in the household.

Turning to the income model, again most coefficients are statistically significant, except for some accessibility and zonal attributes. Importance of accessible grocery store floor space, for example, does not decrease with income except between the two lowest income categories. This could due to low-income households having a higher willingness to pay for being near to groceries as they do not dispose of the same means of transport as households with higher income.

The relevance of all other attributes in the bid function increases with income except school-propensity coefficients, which exhibit the opposite pattern. This seems reasonable when considering that high-income households tend to live in low-density neighborhoods. For land-use entropy, we show that land-use diversity is highly and positively correlated with income. Coefficients for dwelling-size attributes are as expected. Moreover, the high coefficient for proportion of low-income households confirms De Palma's (2005) findings.

Comparing both models reveals that final log-likelihood is better for the *size model* which also has more statistically significant coefficients.[6] Looking at the coefficients for each variable, we notice several differences. Accessible grocery store floor space is more important when considering household size. While dwelling size measured in rooms similarly increases willingness to pay with increasing size and income in a household, it seems to be still more important for larger households (as opposed to smaller) than for richer ones (as opposed to poorer). The same is true for available square meters, although this value increases more with income, which shows the relevance of dwelling size for households with high income. A considerable difference exists between concentration indicators. Low-income households tend to have a much higher willingness to pay for locations when increasing concentration of low-income households than other income groups for zones with high proportions of their type; one-person households show the opposite pattern. School-propensity and land-use entropy coefficients are problematic regarding interpretation in both models and are probably proxies for other variables.

This comparative analysis shows that in spite of the uncertainty associated with imputation, the *income model* raises reasonable aspects of the location pattern that are not fully captured by the *size model*, confirming aditionally, the importance of including income to model residential location decision making

5. Conclusion

We estimate two bid-auction based location choice models using German *Zensus 2011*: one for household income and one for household size. As *Zensus* does not contain household income, we impute this variable based on *Mikrozensus 2010*. The comparison of the two

[6] Note that this is only an indication since due to different choice variables both models cannot be compared by performance indicators such as AIC or pseudo rho-squared directly.

location models must be treated with caution, as they rely on different theories. However, for a location model that explains imputed choice, the number of significant variables is surprisingly high. Although household size and income are generally correlated they show different but reasonable patterns, which means that the *income model* can discover plausible effects in the data. Therefore, we conclude that regression imputation is an option to deal with the lack of choice variables in large data sources such as censuses. As our results have shown, models of imputed choice can perform well if another data source exists that helps to develop a high-performing imputation regression model.

Further explanatory variables could further improve the imputation as well as the location models. Variables 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, population density, better accessibility measures that consider congestion or other cost factors, and additional dwelling attributes, such as the quality or age of the house, might achieve a better fit.

In future studies better location model performance could be attained by either developing new models that simultaneously impute choice-relevant variables and estimate location choice rather than sequentially. In summary, our analysis shows that imputation of choice variables in large data sources is a low-cost option that should be considered for modelling location choice with missing variables.

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**,** 11-38. Online: https://www.jtlu.org/index.php/jtlu/article/view/806, last visited on 30th of October 2016.

ALONSO, W. 1964. *Location and land use. Toward a general theory of land rent,* Cambridge, United Kingdom, Harvard University Press.

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**,** 77-82.

BEN-AKIVA, M. & BOWMAN, J. 1998. Integration of an activity-based model system and a residential location model. *Urban Studies,* 35**,** 1131 - 1153.

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**,** 506-526.

BIERLAIRE, M. 2003. BIOGEME: a free package for the estimation of discrete choice models. *Proceedings of the 3rd Swiss Transportation Research Conference*. Ascona, Switzerland. Online: https://infoscience.epfl.ch/record/117133/files/bierlaire.pdf, last visited on 30th of October 2016.

BRENKE, K. 2008. Migranten in Berlin: Schlechte Jobchancen, geringe Einkommen, hohe Transferabhängigkeit. *Wochenbericht des DIW Berlin,* 35**,** 496-507.

BUSCH, A. & HOLST, E. 2013. Geschlechtsspezifische Verdienstunterschiede bei Führungskräften und sonstigen Angestellten in Deutschland: Welche Relevanz hat der Frauenanteil im Beruf? / The Gender Pay Gap in Leadership and Other White-Collar Positions in Germany: Putting the Relevance of Women's Share in Occupations into Context. *Zeitschrift für Soziologie,* 42**,** 315-336.

CERVERO, R. & KOCKELMAN, K. 1997. Travel demand and the 3Ds: density, diversity, and design. *Transportation Research Part D,* 2**,** 199-219.

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 | Transporti Europei,* 31**,** 67-82.

ELLICKSON, B. 1981. An alternative test of the hedonic theory of housing markets. *Journal of Urban Economics,* 9**,** 56-79.

EWING, R. & CERVERO, R. 2010. Travel and the built environment - A meta-analysis. *Journal of the American Planning Association,* 76**,** 265-294.

FDZ 2010. Mikrozensus 2010. Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, dataset received on 30th of April 2016.

FDZ 2011. Zensusgesamtdatenbestand Berlin 2011. Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, dataset received on 29th of July 2016.

GUO, J. & BHAT, C. 2004. Modifiable Areal Units: Problem or Perception in Modeling of Residential Location Choice? *Transportation Research Record,* 1898**,** 138-147.
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.
HELDT, B., GADE, K., & HEINRICHS, D. (2016). Determination of Attributes Reflecting Household Preferences in Location Choice Models. *Transportation Research Procedia, 19,* 119-134.
HERZOG, T. N., SCHEUREN, F. J. & WINKLER, W. E. 2007. *Data Quality and Record Linkage Techniques,* Wiesbaden, Springer Science + 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**,** 329-376.
HURTUBIA, R. 2012. *Discrete choice and microsimulation methods for agent-based land use modeling.* PhD, École Polytechnique Fédérale de Lausanne.
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**,** 47-65.
HURTUBIA, R., GALLAY, O. & BIERLAIRE, M. 2010. Attributes of household, locations and real-estate markets for land use modeling. *SustainCity Deliverable 2.7.* Lausanne: 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**,** 323-340.
LEE, B. H. Y. & WADDELL, P. 2010. Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation,* 37**,** 587-601.
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**,** 457-470.
MARTÍNEZ, F. 1996. MUSSA: Land use model for Santiago City. *Transportation Research Record: Journal of the Transportation Research Board,* 1552**,** 126-134.
MARTÍNEZ, F. & DONOSO, P. 2010. The MUSSA II land use auction equilibrium model. *In:* PAGLIARA, F., PRESTON, J. & SIMMONDS, D. (eds.) *Residential Location Choice.* Springer.
MARTÍNEZ, F. & HENRIQUEZ, R. 2007. A random bidding and supply land use equilibrium model. *Transportation Research Part B: Methodological,* 41**,** 632-651.
MCCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109-142.
MCFADDEN, D. 1978. Modelling the choice of residential location. *In:* KARLQVIST, A., LUNDQVIST, L., SNICKARS, F. & WEIBULL, J. (eds.) *Spatial interaction theory and planning models.* Amsterdam: North Holland.
MINCER, J. 1974. Schooling, Experience, and Earnings. Ney York: Columbia University Press.
OPENSTREETMAP CONTRIBUTORS 2016. OpenStreetMap project pages. Online: http://www.openstreetmap.org, last visited on 23rd of April 2016.
ORTÚZAR, J. de D. & WILLUMSEN, L. G. 2011. *Modelling Transport,* Chichester, West Sussex, UK, John Wiley & Sons, Ltd.
ROSSI, P. H. 1955. *Why Families Move: A Study in the Social Psychology of Urban Residential Mobility,* London, Free Press.
RUBIN, D. B. 1987. Multiple imputation for nonresponse in surveys.

RUBIN, D. B. & LITTLE, R. J. 2002. *Statistical analysis with missing data*, 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. *Journal of Transport and Land Use,* 7**,** 3-21 Online: https://www.jtlu.org/index.php/jtlu/article/view/740, last visited on 30th of October 2016.
SENSTADTUM 2012. Teilverkehrszellen Berlin. Senatsverwaltung für Stadtentwicklung und Umwelt Berlin, Online: http://www.stadtentwicklung.berlin.de/verkehr/datengrundlagen/verkehrszellen/, last visited on 30th of October 2016.
SENSTADTUM 2016. Environmental Atlas. Senatsverwaltung für Stadtentwicklung und Umwelt Berlin, Online: http://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/edua_index.shtml, last visited on 30th of October 2016.
TIMMERMANS, H. 2003. The saga of integrated land use-transport modeling: How many more dreams before we wake up? *10th International Conference on Travel Behaviour Research.* Lucerne.
WEGENER, M. 2004. Overview of land-use transport models. *In:* HENSHER, D. A. & BUTTON, K. (eds.) *Transport Geography and Spatial Systems.* Kindlington, UK: Pergamon / Elsevier Science.
WEGENER, M. 2011. From macro to micro - How much micro is too much? *Transport Reviews,* 31**,** 161-177.
WEGENER, M. 2014. Land-use transport interaction models. *In:* FISCHER, M. M. & NIJKAMP, P. (eds.) *Handbook of Regional Science.* Berlin: Springer-Verlag.
WOOLDRIDGE, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: The MIT Press.
YEE, T. W. 2010. The VGAM package for categorical data analysis. *Journal of Statistical Software,* 32**,** 1-34. Online: https://www.jstatsoft.org/index.php/jss/article/view/v032i10/v32i10.pdf, last visited on October 30th 2016.
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.