CHALLENGES OF DATA REQUIREMENTS FOR MODELLING RESIDENTIAL LOCATION CHOICE: THE CASE OF BERLIN, GERMANY

Benjamin Heldt
Kay Gade
Dr.-Ing. Dirk Heinrichs
German Aerospace Center, Institute of Transport Research

1 INTRODUCTION

The locations of households and firms play a significant role in daily mobility patterns. Location decisions as long-term mobility decisions considerably affect short-term mobility decisions such as destination choice or mode choice (Wegener & Fürst, 1999). For choosing a location, households or persons consider a variety of aspects regarding or affecting the real estate (e.g. land price, neighbourhood) as well as transport-related factors such as accessibility of locations for daily activities. Land-use models in general, and residential location models in particular, are rooted in a long-standing history (Alonso, 1964; Lowry, 1964; McFadden, 1978; Anas, 1982; cp., e.g., Wegener, 1994; Pagliara & Wilson, 2010). Applying them helps to better understand the corresponding relations of land use and residential mobility to transport, especially in the urban context. The identification of influencing factors and their intensity is helpful in understanding the urban system. The actual application of such models also enables the prediction of land use and transport interactions subject to policy interventions (Lowry, 1965; Batty, 2009).

While land-use models in general, and specifically location models, are very useful tools to analyse and plan urban development, they imply considerable data requirements (Wegener, 2011). Although data availability is improving, selecting suitable data is still a challenging process and often requires compromises. This paper addresses a set of related questions: First, what criteria are best suited to evaluate and select available data for location-choice modelling? Second, what are the 'typical' limitations of such data and what are the available methods to overcome them? The paper explores these questions empirically by setting up a residential location model for Berlin, Germany. Berlin, a city with a dense and well-functioning transport system, has recently featured a strongly growing population and thus provides an interesting case for analysing the interactions between long-term and short-term mobility.

The paper first discusses criteria for data evaluation and relates them to data requirements of urban land-use models, with particular reference to the model Cube Land, which is used for the Berlin case (Section 2). Next, the paper introduces the relevant data sets available for Germany and Berlin, analyses their quality against the requirements of estimating a residential location model and identifies the main limitations (Section 3). We show, firstly, that data frequently lack attributes and/or their interrelations and, secondly, access to data is often limited as their use is subject to conditions. Following this, we demonstrate and discuss approaches to overcome these restrictions (Section 4). The paper concludes with a short summary of the main findings and their implications for modelling location choice. We find that our approach helps to consciously select data according to their advantages and disadvantages.

2 DATA REQUIREMENTS OF URBAN LAND-USE MODELS AND CRITERIA TO ASSESS DATA QUALITY

As models are based on data, their outcome interrelates with the quality and quantity of input data. To model the complex urban system, urban land use models require a high quantity of different types of data. In recent years, data availability has improved as the digital revolution facilitates the collection and processing of big data. However, models also have become more complex and thus data-demanding (Wegener, 2011). *High-quality data* must match model requirements and be suited to answering its research questions, but land use models differ by purpose and thus type. Since the community has developed at least 15 distinct operational urban models (Hunt et al., 2005), applying criteria for defining their data requirements seems appropriate. In the following, model characteristics described in several reviews of land use and transport interaction models serve to derive evaluation criteria. To illustrate the application of these criteria with our case, we determine the data requirements of the land use model 'Cube Land'.

2.1 Data quality

For evaluating data for urban land use models, we apply the general dimensions of data quality for deriving data-relevant model requirements. Quality of data comprises the following aspects: relevance, accuracy, level of detail comparability, coherence, completeness, clarity and accessibility, etc. (for a detailed description of the definitions of these aspects, refer to Ehling & Körner, 2007; Herzog et al., 2007; Ortúzar & Willumsen, 2011; Veregin, 1999). Data demanded by urban land use models must primarily suit the models' purpose. Accordingly, we address *suitability* of the data as the most important aspect of data quality in more detail below.

Trustworthiness and *accessibility* are other important aspects to consider. High-quality data should of course produce trustworthy outcomes, and hence feature accuracy, coherence, completeness, clarity, and comparability. Accuracy refers to 'the degree to which a measurement or model result matches true or accepted (valid) values' (Ortúzar & Willumsen, 2011, p. 10). Accuracy should be complemented by clarity, i.e. meta information that enables understanding of the data and its production. Trustworthiness also involves coherence and comparability, i.e. the ability to combine different data sources without producing inconsistent outcomes, internal to one data source (coherence) and among multiple datasets referring to the same observations (comparability) (Ehling & Körner, 2007). Finally, data must be complete, i.e. exhibiting no missing values (Herzog et al., 2007). In addition to suitability and trustworthiness, accessibility describes the conditions under which data can be acquired and used (Ehling & Körner, 2007).

With no doubt trustworthiness and accessibility of data are very important aspects to consider for selecting data sources. However, since the evaluation of trustworthiness and accessibility is more general than model-specific, we will focus here on data suitability, which differs slightly regarding estimation and simulation of a land use model. Nevertheless will accessibility play a role for data limitation as discussed in Section 4.

Estimation and simulation

Later in this paper we will demonstrate the application of data quality criteria for evaluating data to estimate a residential location model (Section 3). This serves as one part of the 'Berlin Land use Model' (BLUME, see Section 3.1), applying the framework

Cube Land (cp. Section 2.2). For understanding the specific requirements of estimation base data, a short explanation of the concept of estimation is given here. Model development generally comprises two parts, which have different data quality requirements: estimation and simulation. *Estimation* aims at statistically explaining observed patterns by calculating the influence of parameters (cp. e.g., Cambridge Systematics, 2010, p. 1-4). *Simulation* uses the resulting coefficients in order to reproduce the observed choice behaviour under changing conditions. Consequently, to produce the best attainable outcomes during simulation, the error should be as low as possible during estimation. In general, estimation requires more comprehensive data regarding the variety of attributes than simulation. In simulation, only information about those parameters found significant during estimation is needed, while estimation needs to consider all possibly influencing parameters. Due to the intention to explain observations, estimation generally analyses observed data, e.g. a survey.

Criteria for the evaluation of data suitability

The characteristics of a model can be used to derive suitability criteria. *Suitability* here refers to relevant and temporally fitting data. Statistical organisations consider data as *relevant* to the extent that they meet the needs of their users (Ehling & Körner, 2007). Translated to modelling, relevant data must fit the research questions addressed by the model. We discuss this concept in more detail below. Of further importance is the temporal fit, i.e. the lag between the time at base of model and the time the data is referring to.

For evaluating data-relevance, a subset of criteria was derived from a number of reviews of urban land use and transport-interaction models (Hunt et al., 2005; Iacono et al., 2008; Pagliara & Wilson, 2010; Timmermans, 2003; Wegener, 1994; Wegener & Fürst, 1999; Wegener, 2011). Comparing the criteria these authors used to evaluate models against each other, we determined those aspects that relate to data. These are *spatial level of detail and spatial system*, *model components*, *factual level of detail*, *interrelations of model components and other models*, and *parameters*.

Spatial level of detail and spatial system

Spatial level of detail and spatial system determine the degree of detail of location-influencing land use (here also referred to as neighbourhood or zonal) variables (Hunt et al., 2005, Ortúzar & Willumsen, 2011). However, according to Wegener (2011) more micro is not necessarily better, as microscopic modelling frequently adds stochastic variation due to random sampling. How much spatial detail is appropriate rather depends on the purpose of the model. Not only is the spatial level of disaggregation important, it is also relevant to know the spatial system, and hence be aware of the extent to which systems used geometrically refer to each other. This ensures the ability to combine them without having intersections. Thus, the more spatially disaggregate data are, the better, since this simplifies their aggregation to the spatial system of the model.

Model components

The components of a model include the physical system modelled and decision-makers, among others (Hunt et al., 2005). Which components a model comprises, relates to the underlying theory. Of considerable importance are the observations or observed patterns that the simulation intends to reproduce; in other words, the dependent variables the estimation should explain. As an example, in residential location choice

models, observed units can be households. Hence, data that describe persons is not well-suited to such a model.

Factual level of detail
Transport models and land use models alike, regarding their factual level of detail, are often referred to as microscopic or macroscopic; i.e. do they depict the behaviour of individuals one by one or aggregated as the sum? (Ortúzar & Willumsen, 2011; Lowry, 1965; Wegener, 2011) Accordingly, factual level of detail describes the level of disaggregation at which observed units or model components are measured. Similar to spatial level of detail, model quality generally improves with increasing factual level of detail, but data requirements and error rise as well. An additional consideration regarding data is, the more detailed a model is, the more disaggregated the data must be. Depending on the type of model, it thus requires (disaggregate) microdata or aggregate data, representing model components and their interrelations (see below)

Interrelation of model components and other models
The theory underlying each model also suggests the interaction between the model in question and other models, as well as interrelations between its own components. Hence, the more integrated a model is, the more data sources have to fit each other regarding their level of detail and attributes. As an example, land use models that treat the real estate market as households located in dwellings require a data source that includes the combination of observed households and dwellings. What is more, land use models frequently relate to a transport model, which can be either integrated (land use transport interaction model), or connected. Connected transport models require less, but still consistently, structured data in the land use model.

Parameters
Parameters refer to the independent variables explaining the observed pattern of the model components and are reflected by attributes. Such attributes exhibit average or distinct values depending on the factual level of detail of the corresponding model component.

In addition to these criteria, particularly when using surveys for estimation, information should be representative of the simulated system. Thus, sample data needs to include a weight for extrapolating it to the population. In Table 1 we summarise the criteria and relating questions being applied for identifying suitable data. How to find information answering these questions is addressed in Section 3.2.

Table 1: Criteria and questions for evaluating data suitability for land use modelling

Relevance	• Which are the model components, i.e. observed units? • What is the spatial level of detail? • What is the factual level of detail? • How do model components interrelate? • Which attributes describe the observed units?
Temporal fit	• What is the base year?
Represen-tativeness	• How can sample data be extrapolated to the whole population?

2.2 The model Cube Land and its data requirements

By developing BLUME, the model and software *Cube Land* is applied to the city of Berlin. The aim of Cube Land is to predict the location of households and firms in a study area. In conjunction with a transport model, Cube Land is able to model the interaction between location choices and the transportation system.

The underlying model used by Cube Land is based on discrete choice theory and simulates a market where demand and supply meet in a bid-auction process (Martinez, 1992; Martinez & Henriquez, 2007; Martinez & Donoso, 2010). It presupposes rationally behaving consumers (households and firms) and real estate suppliers, both aiming at maximising their benefit. For consumers, the benefit comprises the obtained utility of a location, reflected in monetary surplus, whereas for suppliers the benefit is represented by monetary profits. Both are interdependent as the suppliers' profit depends on the consumers' willingness to pay, expressed as bid in the bid-auction process. Within this process all consumers provide a bid for every available location, the consumer with the highest bid will be assigned to the location. Market equilibrium is achieved if there are no incentives for consumers to change locations or suppliers to produce new real estate.

In Cube Land, the complex mechanism is divided into the three sub-models demand, supply and rent. In the following, the first two sub-models will be explained in more detail, with a focus on the derived data requirements. The latter one contains the equilibrium solution algorithm, which is based on the utility function provided by the demand model and the supply function provided by the supply model.

The demand model determines the consumers' bids for each available location. This bid is the result of the utility function, which takes into account consumer characteristics and preferences. Consumers are represented by clusters based on similar socioeconomic characteristics (households) or economic activity and business size (firms). For example, households can be clustered by income and amount of persons living together. The parameters of such a utility function represent a variety of attributes with a possible and plausible impact on location choice. Table 2 gives an overview of attribute groups with examples.

Table 2: Attributes used for modelling residential location with Cube Land

Model components and attribute groups	Examples for attributes
Consumer	Household income, number of persons in the household, phase of life
Property	Building type, age of building
Zone – exogenous	Accessibility, attractiveness
Zone – endogenous	Average income in the zone, built area

As will be explained in Section 3.1, data is required for the estimation of these parameters, which combines consumer attributes with property attributes, preferably from a single source. Zonal attributes can be added from different data sources. In the case of the residential location model – the focus of this paper – data needs to encompass household attributes and property attributes as shown in Table 2. Moreover, high data quality is crucial, as the resulting parameter for each attribute influences the simulation result directly.

In contrast to the estimation, the simulation only necessitates the number of consumers as an exogenous variable, real estate supply can be determined as the outcome of the bidding process.

As the counterpart to the demand model, the supply model determines the different real estate types and their amount per zone, under the condition of maximizing the suppliers' profit. The supply is therefore described by property type and zone, each differentiated unit offering a discrete option for the consumers. The decision of which real estate type to build depends on the achievable profit for the supplier as well as possible restrictions by the state, such as housing density. The profit is the result of achievable rent reduced by costs. Therefore the supply model needs data regarding the costs of buildings such as construction and maintenance, differentiated by property type and zone.

In the next section, we illustrate the characteristics and requirements of Cube Land with applying its demand model for residential location choice to Berlin.

3 EVALUATING DATA SOURCES FOR MODELLING RESIDENTIAL LOCATION CHOICE IN BERLIN, GERMANY

3.1 Berlin residential location model and its data requirements

Residential location models are a crucial part of urban land use models attempting to simulate a distribution of residential locations based on several input data. As demonstrated in Section 2.2, Cube Land's residential location model simulates the distribution of households across real estate types and zones. Table 3 (next page) summarises its characteristics, giving answers to the questions shown in Table 1. Accordingly, the model for Berlin comprises interrelated grouped households and dwellings as model components. These are located in a zone in any zonal system representing the spatial level of detail. Model components are described by parameters that explain residential location, such as: household and dwelling characteristics, access measures and zonal attributes. As Cube Land's demand and rent models can be estimated simultaneously including rent data in estimation base data is advantageous. The specific attributes in Table 3 were identified in part by analysing literature on empirical research about residential mobility; in part data availability determines them.

Following a description of our findings from model development regarding main data providers and the retrieving of metadata, a comparison of Berlin-specific data sources against the data requirements just described determines the adequacy of data sources.

Table 3: Model characteristics and data requirements of Berlin residential location model

Which are the model components, i.e., observed units?	Consumers (households) Properties (dwellings)
What is the spatial level of detail?	Any zonal system
What is the factual level of detail?	Disaggregate or aggregate (disaggregate preferred)
How do model components interrelate?	Households with dwellings Households with zones Dwellings with zones
Which attributes describe the observations?	• Household attributes Income Mobility Household size Householder age • Dwelling attributes Building type (form) Building size (apartments) Building age Dwelling size (rooms) Dwelling size (area) Rent • Zonal attributes • Accessibility
What is the base year?	2010
How can sample data be extrapolated to the whole population?	Model needs extrapolation of data for estimation/simulation

3.2 Retrieving data and metadata

In this Section we describe our model-building experiences on where to find important data sources and how to determine information for data evaluation.

Data providers

Frequently, institutions developing models are not disposed of sufficient resources to carry out their own surveys providing data for estimation. Surveys additionally often lack information necessary for model-building. Consequently, acquiring data is normally inevitable. As a result of our research, we identified three major data providers:
public official sources, i.e. the government and associated institutions, the general public (open data), and private data brokers.

The majority of data can be retrieved from public official sources. In the case of the residential location model, these are, for example, censuses of the population serving as estimation base data, satellite images and cadastres providing land use information, and tax information describing the real estate market. However, the access to these data depends on the purpose of using it and thus on the type of organisations both providing and asking data. Particularly micro data, i.e. data describing single persons or households, is often restricted to scientific non-profit purposes. Even the scientific use of these data is limited, due to the protection of privacy (cp. Section 4.2). This implies considerable problems for building residential location models, as we will demonstrate

when evaluating the data for Berlin. For scientific surveys, the access is subject to the decision of the institute that has funded data collection.

Another important public data source is the general public, i.e. so-called user-generated content. Open Street Map and Wikipedia are the best-known examples. Such data are also valuable for calculating land use attributes; but usually they do not serve as estimation base data as they do not comprehend microdata. The sheer mass of user-generated data, such as information collected by smartphones, however, leads to them becoming more and more important in the near future, not only for residential location modelling (Lyons, 2014).

Finally, private data brokers are either specialised in selling their own data, or they market the data of others, including public data. Combining public data and information coming from lotteries, raffles, or mail order, they additionally often carry out their own consumer surveys for market research. Such companies may also sell household-based information, but usually this is excluded by law and thus not useful for scientific purposes.

Information for evaluation

The evaluation of data against model requirements requires the availability of metadata. While model components and their factual level of detail become obvious through the name or description of a data source, code plans describe the attributes that provide the parameters. Regarding the interrelation of model components, one data source can contain multiple observed units or only their attributes. Alternatively, the data source may provide key fields that enable to link it to other data sources representing other components. Finally, surveys usually contain spatial level of detail as attributes describing location. The availability of metadata depends on the providing institution. This sometimes demands persistent enquiring, in particular, regarding private data brokers.

3.3 Description and evaluation of data sources for estimating Berlin residential location model

For the estimation of Berlin residential location model, four main data sources have been identified. Three of them are public or scientific surveys, with each having different purposes. The fourth is a data collection by a private data broker. Before comparing them, a short description of each data source will be given. Additionally, an introduction to the spatial levels at which data for Berlin is available facilitates understanding the data we present. In Table 4 we summarise the comparison of our data sources.

Spatial reference systems in Berlin

Varying by data provider and purpose, different spatial reference systems are applied in the city of Berlin (cp. Bömermann 2012). Administratively, the city is divided into 12 districts, which originated from 23 sub-districts that were consolidated during the administrative reform in 2001. For spatially fine-grained analyses, the local government developed a so-called regional reference system, consisting of addresses, streets, partial blocks and blocks, as well as two statistical systems. Until 2007 the city was separated into 195 statistical areas, which could be further subdivided into 338 traffic analysis zones and 1193 partial traffic analysis zones. In 2007, the senate decided to introduce a new spatial system called ‘LOR’, which caters more to the socio-spatial

reality (Bömermann 2012) and comprises of three related levels: 60 prognosis areas, 138 intermediate areas, and 447 planning areas. However, the two statistical systems can only be linked by aggregating partial traffic analysis zones to prognosis areas.

Description of Berlin-specific data sources

Zensus 2011
In 2011, a census of population and buildings was carried out in Germany, which complied with the requirements of the European Union. It includes a complete survey of buildings and dwellings as well as a 10 %-household survey, also containing person-related information. Access to microdata is envisaged for 2015, but limited to scientific use only.

Mikrozensus 2010
The Mikrozensus is a survey of 1 % of households located in Germany. It is carried out each year and additionally contains every four years questions on living conditions with the last such survey dating back to 2010. Admission is limited to scientific research, either in a controlled-access area, or by a 'Scientific Use File', which contains a 70 % sub-sample without appropriate geo-location.

Survey on mobility in cities 2008 ('SrV')
Travel surveys often provide data on mobility behaviour in connection with living conditions. In Germany, the most comprehensive travel survey regarding urban areas is the SrV, which is conducted every five years by Technical University Dresden in cooperation with the concerned federal states. The latest SrV available from 2008 consists of three parts: a household survey, a questionnaire for the person, and a travel questionnaire. Availability of SrV is subject to the decision of the institutions in charge.

infas geodaten 2010
The only data source coming from a private data broker contains products from 'infas geodaten' (from 2014 on 'nexiga'). These consist of three different data sets about Berlin: 359,200 addresses of buildings with building age, number of households, building type; a description of 3,900 neighbourhoods with attributes, such as number of households by income group, householder age, etc., and 168,500 businesses, with names and addresses, their economic activity, and number of employees.

Evaluation of these data sources
To evaluate these data sources, we compared them against the data requirements of the residential location model of Cube Land, as highlighted in Section 3.1.

Metadata is openly available for Mikrozensus, Zensus, and SrV. Infas geodaten delivered a description of attributes, but did not provide sufficient information on data background. In particular, understanding the response categories of public data and their corresponding implications may require consulting the institution that holds the data source. Table 4 gives an overview of the results outlined below. The comparison leaves out some aspects for the data from infas geodaten. Since this is no microdata but the distribution of households within a zone by a single attribute, it cannot be used for estimating location choice, but for simulation.

Observations
Data for estimating residential location choice in Cube Land includes observations of households and dwellings in zones (cp. Section 3.1). Mikrozensus and Zensus both consist of such information; nonetheless, Mikrozensus only includes a sample of households described by dwelling attributes, while Zensus features a household sample and a survey of dwellings and buildings which can be connected. In contrast to both data sources, SrV describes households located in zones, but without dwellings or any related information.

Factual level of detail
Both factual and spatial level of detail depend on the conditions of using the data. In general, spatial resolution is exchanged for factual level of disaggregation. Thus, although our main databases contain geocoded microdata, these must be aggregated in order to protect privacy (cp. Section 4.2). One exception is scientific microdata from SrV, which retains geocodes and spatial level of detail even when working in-house.

Spatial level of detail
Cube Land does not require a specific spatial system or level of disaggregation. However, considering zones as entities that influence location decisions, data with high spatial level of detail is likely to explain location choice better. The most spatially detailed data comes from infas geodaten but is not useful for estimation. Observed data, provided by SrV, is geocoded at statistical areas, but representative only for sub-districts, while Mikrozensus is geocoded at sub-districts, but weight factor extrapolates the data only to districts. Since the Zensus contains the largest household sample, it is expected that microdata is geocoded at least at districts, rather more detailed at LOR-level.

Interrelations
Table 4 shows the relations between the observed units representing model components. Mikrozensus and Zensus both relate household and dwelling attributes, while SrV does not contain such information. Linking between the different data sources, including infas geodaten, is possible by location only. The transport model provides access measures and both models must therefore be consistent regarding model design and data.

Attributes
In contrast to all the other data sources, SrV describes the mobility of households, such as number of licenses, car ownership level, etc. Only Zensus lacks the very important variable household income; in return it includes a better description of buildings and dwellings. As part of living conditions, Mikrozensus is the only data source that asks for rent. All data sources lack attributes describing zones and only SrV includes some accessibility measures; however, both these attributes can be joined by location (see Section 4.1).

Concerning temporal fit, Mikrozensus matches best, followed by Zensus and SrV.

Table 4: Evaluation of data suitability for estimating Berlin residential location model

Descriptor	Data required by Cube Land	Data: Zensus 2011[a]	Data: Mikrozensus 2010	Data: SrV 2008	Data: infas geodaten[e]
Sample size households (records)		n/a[b]	14.700 / 15.600[d]	19.350	-
Model components	Consumers (households) Properties (dwellings)	Households Dwellings	Persons Households Dwellings	Persons Households Trips	n/a[e]
Spatial level of detail (zone)	Any zonal system	n/a[b]	23 districts	195 statistical areas	3886 neighbourhoods
Factual level of detail	Disaggregate or aggregate (disaggregate preferred)	Disaggregate or aggregate[c]	Disaggregate or aggregate[c]	Disaggregate	Aggregate
Relations	Households in dwellings	**X**	**X**	-	-
	Households in zones	**X**	**X**	**X**	-
	Dwellings in zones	**X**	**X**	-	-
Attributes of observed units	• Household attributes				
	Income	-	**X**	**X**	**X**
	Mobility	-	-	**X**	-
	Household size	**X**	**X**	**X**	**X**
	Householder age	**X**	**X**	**X**	**X**
	• Dwelling attributes				
	Building type (form)	**X**	-	-	**X**
	Building size (apartments)	**X**	**X**	-	**X**[f]
	Building age	**X**	**X**	-	**X**
	Dwelling size (rooms)	**X**	-	-	-
	Dwelling size (area)	**X**	**X**	-	-
	Rent	-	**X**	-	-
	• Zonal attributes	-	-	-	-
	• Accessibility	-	-	**X**	-
Base year	2010	2011	2010	2008	2010, 2011
Representativeness	model needs extrapolation of data for estimation/simulation	n/a[b]	only for 12 districts	only for 23 subdistricts	already extrapolated

(X) partially contained or can be approximated from the other attributes

[a] for Zensus only marginal sums are available, characteristics of final microdata set for Berlin not yet clear (publication envisaged for 2015)
[b] Zensus samples 10 % of the German households, spatial level of detail for Berlin is thus likely better than districts, at LOR
[c] depending on the confidentiality level of the analysis, microdata need to be aggregated
[d] depending on version (refer to description of Mikrozensus 2010 above)
[e] no microdata and not based on a survey – no observations
[f] approximated as number of households per address

To conclude, the most suitable data source will, once available, likely be Zensus due to its spatial resolution, large sample size, and comprehensive description of dwellings. Mikrozensus ranks second, as its spatial level of detail is unsufficient. SrV is the least preferred data source because it misses real estate attributes. As Zensus microdata wil not be available before 2015, we focus on Mikrozensus. Hence, BLUME's spatial level of detail at the time being is districts.

As has been demonstrated, all these data sources have their advantages and disadvantages. Limitations are mainly caused by survey design or confidentiality requirements, resulting in the need for aggregating or adding data, which is discussed subsequently.

4 APPROACHES TO OVERCOME LIMITATIONS

Evaluating the data sources, we find two main problems limiting data for residential location modelling. On the one hand, data limitations arise from missing attributes depending on the survey design. On the other hand, the conditions of the use of microdata require aggregating data and information loss. For both problems, the challenge is to find methods helping to control the loss of information and constructing a database still adequate for residential location modelling.

4.1 Incomplete data

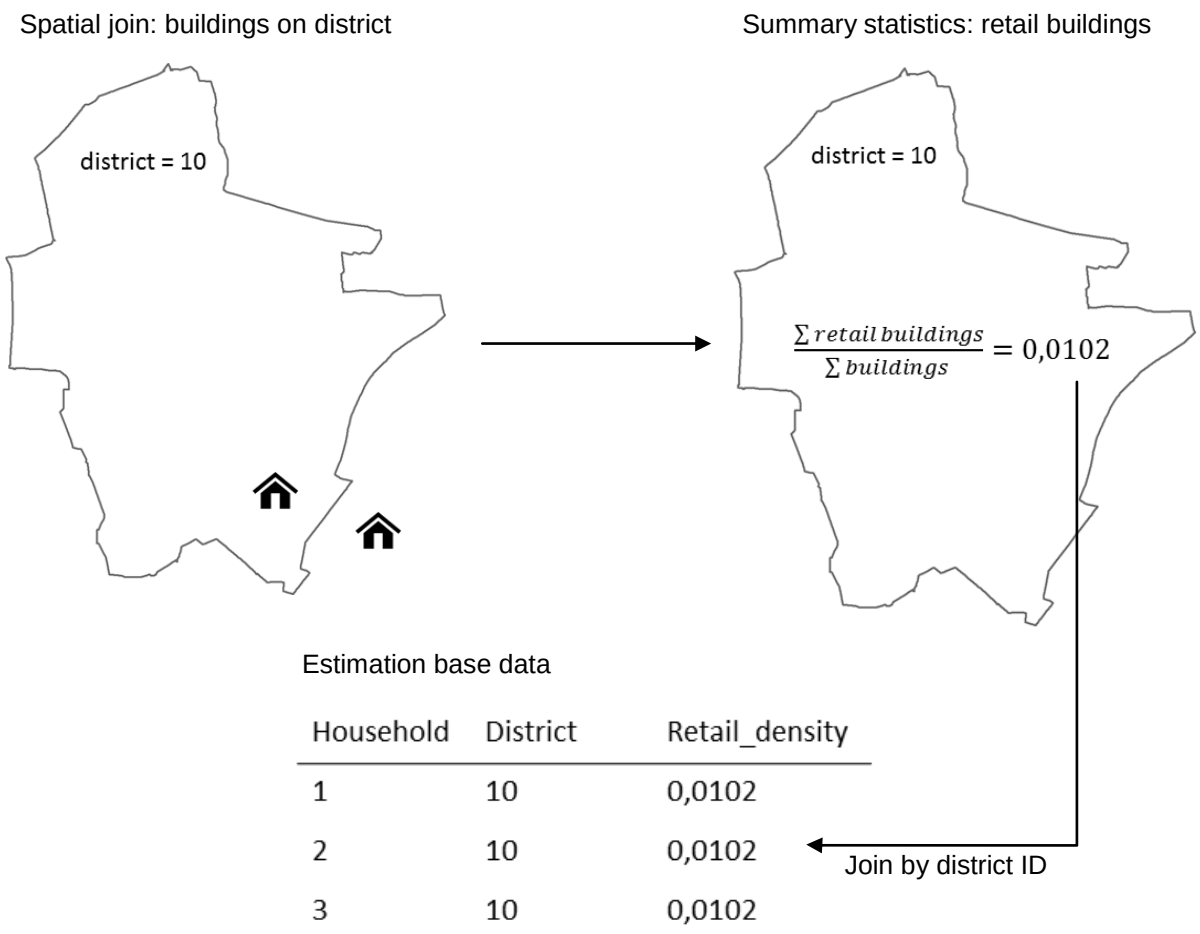
Thus, although household-based data is generally available, sources not intended for land use modelling likely lack the required information. Travel surveys such as SrV accordingly do not include dwelling attributes, while demographic surveys are lacking mobility attributes. In general, two groups of attributes may be missing. First, those that describe households or dwellings, i.e. they qualify the factual units of analysis (cp. Section 2.1). Second, attributes that represent characteristics of the neighbourhood, i.e. the spatial unit of analysis, or access measures.

Several approaches may contribute to completing data, namely record linkage, imputation, and (spatial) join (e.g., Herzog et al., 2007). Record linkage is a technique that identifies identical or similar records in two or more data sources that share information, such as several surveys of the same households. Identifiers or several matching attributes enable the joining of datasets corresponding to the same record. Imputation is another methodology that serves to add missing values or characteristics, e.g. of households or dwellings, based on data contained in a data source. For this purpose, a statistical model analysing a second data source calculates the probability of the observed unit having a characteristic, e.g. car ownership. However, this sub-model does not actually attach more information to the data and hence will not improve the model (cp., e.g., Herzog et al., 2007; Schafer, 1999). Finally, location or other attributes provide the link for adding information that is more aggregate than the factual unit of analysis. Relating matching spatial systems to each other by reference systems or by geoprocessing operations, such as spatial join and intersect allows such information to be added.

In the Berlin case, our database consists of sources whose corresponding terms of use prohibit linking them to one another. Primarily concerning microdata (Mikrozensus, Zensus), this is again due to data confidentiality. Thus, we cannot use record linkage.

Since imputation does not improve model estimation, we only used spatial joins for the data preparation for BLUME. While combining several microdata sources is not possible, it is generally less of a problem to add aggregate data, based on information already included in the data source. Hence, in adding land use data from other public official sources such as the environmental atlas and the cadastre, we complemented zonal attributes based on location. These data sources are spatially disaggregate at the block or building level. Thus, we aggregated them to the district level, using summary statistics in ESRI ArcGIS. We attached the calculated zonal attributes to the dwellings via the identifier for the zone, in our case the district. Figure 1 below shows this procedure for retail density. Aggregating the amount of retail buildings in a district and dividing it by the corresponding total number of buildings, we added this number to the dataset at base of the estimation.

Figure 1: Complementing data by spatial join



4.2 Data aggregation and information loss

In Germany, data availability and accessibility are frequently limited due to the protection of data privacy. To use disaggregate data like Mikrozensus, it is required to edit data in a way that there is no identification of individuals or households possible. Therefore data needs to be either selectively deleted or aggregated.

Model design and derived data requirements are other reasons for an aggregation of data. As mentioned in Section 2.2, Cube Land models the location choice of household types instead of single households. Therefore survey data for individual households has to be aggregated to household types, clustering households with similar behaviour regarding their location choice decision.

As a consequence of such aggregation, a loss of information will occur, either by defining a classification too broadly, deleting less important parameters or deleting data which still does not fulfil privacy concerns. Therefore, the aggregation process has to consider that the identified clusters are homogeneous in themselves and heterogeneous as compared to each other; but types still need to represent a broad variety, and must be able to answer the underlying research questions. In case of models like Cube Land, these clusters further need to be predictable for simulations as well.

Inductive approach for finding appropriate clusters

Different approaches to find such groups are summarised under the term ‘cluster analysis’. In general, cluster analysis consists of a definition of similarity measures and a clustering algorithm, which defines the rules for clustering the objects based on the inter-object similarities (Backhaus, 2006; Dillon, 1984).

As described before, Cube Land models the decision of household types. Moreover, dwellings need to be grouped in order to reduce the number of alternatives. Therefore the challenge is to cluster the Mikrozensus data in a way that the amount of clusters is maximized (to represent a high variety) and to fulfil the following constraints:

- Identical range of values for the attribute household income between estimation and simulation, as it is expected to be an influencing variable for location choice decisions.
- As the land use model will interact with the transport model, both will use the same population as input data. Only a further aggregation of the pre-defined range of values for household income is possible, due to the model requirements and source data of the transport model.
- Due to data privacy concerns, every cluster must contain at least three objects.

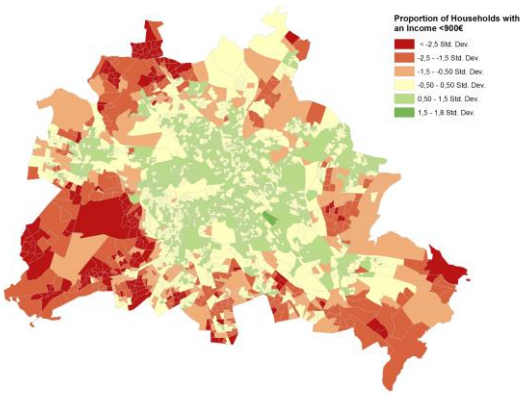
We chose an inductive approach consisting of a step-by-step reduction of involved variables, a method similar to the hierarchical cluster analysis, as well as a GIS-based visual similarity analysis.

In a first step, the Mikrozensus data were aggregated according to relevant location factors and their classifications, data availability, and the model specification of the connected transport model. This resulted in a first cross table of observed household types in dwelling types and zones. As the resulting clusters did not fulfil the constraint of data privacy, a further aggregation was necessary. Household attributes which could be explained by other attributes were excluded, as well as attributes with an expected low influence on location choice. Remaining clusters of household types still not fulfilling data privacy were grouped against dwelling types by using hierarchical cluster analysis with Euclidean distance as the similarity measure. Further aggregation was achieved by GIS-based visual analysis of the spatial distribution of household income, as follows.

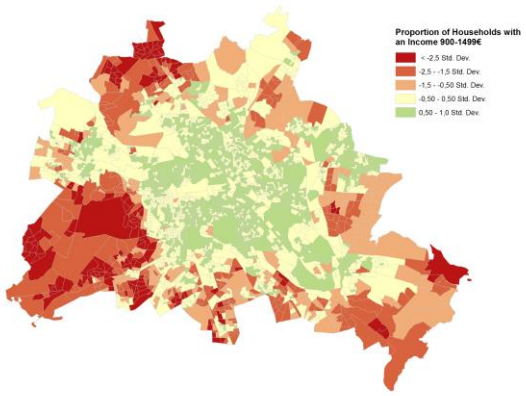
GIS-based visual analysis

Geographic information systems provide a tool for analysing big spatial data. This way, it is also possible to determine the similarity of spatial patterns that different attributes or characteristics of observed units exhibit. For aggregating households in Mikrozensus in order to comply with confidentiality requirements (cp. previous section), we visually compared the spatial distribution of several income categories. At the base of this analysis is aggregate data from a private data broker, describing neighbourhoods by the number of households exhibiting a specific class of income (cp. Section 3.2). Standard deviation serves to qualify the distribution of one income category across the city and can be used to also spatially compare different income categories. Figure 2 shows the results of this analysis. Yellow polygons depict neighbourhoods with a number of households of a specific income category that is very close to the mean, while red symbolises negative deviations, and green positive ones. Aggregating income categories showing similar colour patterns, households with an income of less than €900 and households with an income of €900 to €1499 can be aggregated (Figure 2). This methodology has its pitfalls. On the one hand, the accuracy of the analysis is of course subject to the quality of data. On the other, it only compares the single characteristics of a single variable, not taking into account other location attributes, such as dwelling information, e.g. being addressed by cluster analysis.

Figure 2: Distribution of the number of households by income group across Berlin



Distribution of income group *€1 - 899*



Distribution of income group *€900 - 1499*

5 CONCLUSION

This article is aimed at deriving criteria for evaluating and selecting data suitable to estimating residential location models. A further objective was to identify limitations of data and find methods for dealing with them. The model characteristics (spatial level of detail, model components, factual level of detail, interrelations, and parameters) serve to determine whether data suits a model's purposes. This way we introduced a framework for the description of a model's data requirements. In Cube Land, spatial level of detail can be handled flexibly as any zonal division of the study area. Interrelated households and dwellings are the model components and related data should preferably be at disaggregate level. Finally, parameters such as household and dwelling characteristics, access measures and zonal variables explain location choice. The application of this framework to the available different data sources for the Berlin case identified

Mikrozensus as the most adequate data source that is currently accessible. The framework reveals, however, that all compared data sources have their limitations. Either they do not include all the necessary attributes for residential location modelling, or data confidentiality requires their aggregation, leading to loss of information. Three methods help to add information to the data source. For the Berlin case, protection of privacy excludes record linkage, while imputation does not actually add information. The third method, spatial join, served to add zonal attributes. For aggregating data, we used statistical analyses and thereby balanced the loss of data and the information needs of the model.

In general, this paper demonstrates that surveys suited to residential location modelling are rare, in particular with reference to the linkages between longer term residential mobility and daily mobility decisions. This shows the need for the inclusion of corresponding questions in surveying and closer integration of land use and transport domains. Selecting and applying data not corresponding to a survey tailored to land use modelling is always a compromise.

The application of the criteria developed in this article helps to critically evaluate data sources and assess their advantages and disadvantages regarding spatial and factual depth of information. In this way, this method may be of value to increase awareness of a model's quality depending on the data at base.

BIBLIOGRAPHY

Alonso, W. (1964) *Location and Land use*, Harvard University Press, Cambridge (United States).

Anas, A. (1982) *Residential Location Markets and Urban Transportation,* Academic Press, New York.

Batty, M. (2009) Urban Modelling, *International Encyclopedia of Human Geography,* Elsevier*,* Amsterdam.

Backhaus, K., Erichson, B., & Plinke, W. (2006). Multivariate Analysemethoden. Eine anwendungsorientierte Einfuhrung. 11., überarb. Auflage

Bömermann, H. (2012) Stadtgebiet und Gliederungen, *Zeitschrift für amtliche Statistik Berlin Brandenburg*, 2012 (1+2) 76-87.

Cambridge Systematics, Inc. (2010) *Travel Model Validation and Reasonableness Checking,* US Federal Highway Administration, Washington, D.C.

Ehling, M., & Körner, T. (2007), *Handbook on Data Quality Assessment Methods and Tools*, European Commission - Eurostat*,* Wiesbaden.

Dillon, W. R. (1984), *Multivariate Analysis*, John Wiley & Sons, New York.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007), *Data Quality and Record Linkage Techniques*, Springer, New York.

Hunt, J. D., Kriger, D. S., and Miller, E. J. (2005), Current Operational Urban Land-use–Transport Modelling Frameworks: A Review, *Transport Reviews*, 25 (3) 329-376.

Iacono, M., Levinson, D., and El-Geneidy, A. (2008) Models of Transportation and Land use Change: A Guide to the Territory, *Journal of Planning Literature*, 22 (4) 323-340.

Lowry, I. S. (1964) *A Model of Metropolis,* Rand Corporation, Santa Monica.

Lowry, I. S. (1965) A Short Course in Model Design, *Journal of the American Institute of Planners*, 31 (2) 158-166.

Lyons, G. (2014) Transport's Digital Age Transition, *Journal of Transport and Land use*, forthcoming.

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

Martínez, F., Henríquez, R. (2007) The RB&SM: a random bidding and supply land use equilibrium model, *Transport Research B,* 41 631-651.

Martínez, F., Donoso, P. (2010) The MUSSA II land use auction equilibrium model, in F. Pagliara, J. Preston & D. Simmonds, eds, *'Residential Location Choice', Advances in Spatial Science*, Springer Berlin Heidelberg, pp. 99–113.

McFadden, D. (1978) Modelling the Choice of Residential Location, *Spatial Interaction Theory and Planning Models*, 75-96.

Ortúzar, J. de D., Willumsen, L. G. (2011) *Modelling Transport*, John Wiley & Sons, Chichester (UK).

Pagliara, F., & Wilson, A. (2010) The State-of-the-Art in Building Residential Location Models, *Residential Location Choice,* 1-20.

Schafer, J. L. (1999) Multiple Imputation: A Primer, *Statistical methods in medical research*, 8 (1) 3-15.

Timmermans, H. (2003) The Saga of Integrated Land use-Transport Modeling: How Many More Dreams Before We Wake Up?, *Keynote paper, Moving through nets: The Physical and Social Dimension of Travel, 10th International Conference on Travel Behaviour Research.*

Wegener, M. (1994) Operational Urban Models State of the Art *Journal of the American Planning Association*, 60 (1) 17-29.

Wegener, M. (2011) From Macro to Micro – How Much Micro Is too Much?, *Transport Reviews*, 31 (2) 161-177.

Wegener, M., Fürst, F. (1999) Land use Transport Interaction: State of the Art, *Berichte aus dem Insititut für Raumplanung*, 46.

Veregin, H. (1999) Data Quality Parameters. *Geographical Information Systems*, 1, 177-189.