THE JOURNAL OF TRANSPORT AND LAND USE http://jtlu.org
VOL. 3 NO. 2 [Summer 2010] pp. 5–37 doi: 10.5198/jtlu.v3i2.113

Disaggregate models with aggregate data

Two UrbanSim applications

Zachary Patterson
Concordia University, Montréal, Canada [a]

Marko Kryvobokov
Laboratory of Transport Economics (LET), CNRS, Lyon, France

Fabrice Marchal
Laboratory of Transport Economics (LET), CNRS, Lyon, France

Michel Bierlaire
Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Abstract: UrbanSim has significant data requirements. In particular, it requires disaggregate data (traditionally at the 150 meter by 150 meter gridcell level) for employment, households, and buildings. While such data are not always easily available, most regions have readily available data in a more aggregate form, often at the level of traffic analysis zone (TAZ) or other municipal divisions. This paper describes two UrbanSim applications for the cities of Brussels, Belgium and Lyon, France that adopted different approaches of using aggregate data. In Brussels, aggregate zonal data were disaggregated to the gridcell level. In the Lyon application, the zone was used as the unit of analysis and as such, each zone corresponds to one gridcell. The objectives of this paper are: 1) establish whether an UrbanSim model can be developed using aggregate data; 2) describe two different approaches to using aggregate data with UrbanSim and evaluate; and 3) evaluate the advantages and disadvantages of using aggregate data, as well as the two different approaches described. In doing so, it advances knowledge in the field of transportation and land use modeling by helping modelers evaluate the use of an increasingly popular integrated transportation land use modeling option. Several conclusions flow from this work. First, aggregate data can be used to develop UrbanSim models. Second, only a limited amount of disaggregate information can be drawn from aggregate data. In the context of UrbanSim, this is manifested in models with relatively few variables and dubious simulation results—in other words, while it is possible to develop an UrbanSim application with aggregate data, it should not be used for applied analysis. Finally, the development of such models can be a relatively low-cost exercise to gain familiarity with UrbanSim's functioning and data requirements. As a result, it can also be seen as an important first step to developing or evaluating UrbanSim for application in a new region.

Keywords: Data requirements, Data disaggregation, UrbanSim, Integrated transportation/land use modeling

[a] Principal contact; Zachary.Patterson@concordia.ca

1 Introduction

UrbanSim (Waddell 2002; Waddell *et al.* 2007a) is an increasingly popular integrated transportation/land use model that has been under development since the late 1990s by Paul Waddell at the University of Washington. Two features make UrbanSim particularly interesting for planners and researchers: 1) it is open source, meaning that anyone can freely use the software and, if desired, access, modify and redistribute its code, and 2) it is disaggregate. The model is implemented at the level of individual households, jobs, and real estate developments and operates at fine geographical detail, traditionally at 150 m by 150 m gridcell.

Operating at such a fine level of detail means that it requires a great deal of disaggregate data. While allowing for a rich analysis, this can present significant challenges to model implementation. This paper is written by two research teams, one from the École Polytechnique Fédérale de Lausanne (Switzerland) that developed a model for the Brussels region in Belgium, and the other from the Laboratoire d'Économie des Transports in Lyon (France) that developed a model for Lyon in France. Both research teams were interested in better understanding and developing UrbanSim models; however, given the complexity and data requirements of UrbanSim, each was also hesitant to launch directly into full-scale model development. As a result, both teams independently decided to develop models with aggregate data they had available.

The objectives of this research were as follows: 1) establish whether an UrbanSim model can be developed using aggregate data; 2) describe two different approaches to using aggregate data with UrbanSim and evaluate; and 3) evaluate the advantages and disadvantages of using aggregate data, as well as the two different approaches described. In doing so, it advances knowledge in the field of transportation and land use modeling by helping modelers evaluate the use of an increasingly popular integrated transportation land use modeling option. The two case studies describe two different approaches to using aggregate data in UrbanSim.

The paper begins with some background, a literature review, and a brief introduction to how UrbanSim works and what this implies for the data required to run it. It then describes the Eugene template distributed with UrbanSim and a description of the two case study regions and the data that were available for the analysis, before explaining how aggregate data were used in the two applications and providing reports on the results of the UrbanSim model estimation and simulations. The penultimate section describes the effort required to develop the UrbanSim application and evaluates the two approaches, and the final section presents the main conclusions about what can be learned from the development of UrbanSim models with aggregate data.

2 Background and Literature Review

In the words of (Wegener 1995), who recognized the tendency of transition from zonal to spatially disaggregate data structures in transportation/land use modeling, "We are after (or still in the process of?) a quantum leap in terms of disaggregation of variables and spatial and temporal resolution." According to Wegener, disaggregate models are easier to implement and calibrate, more practical in their data needs (because they can work with sample data or synthetic micro data), more flexible with respect to testing new hypotheses or policies, and easier to communicate to non-experts and decision makers. In the overviews of Wegener (2004) and Hunt *et al.* (2005), UrbanSim is analyzed among the selected contemporary frameworks representing the state-of-the-art in transportation/land use modeling. The latter review notes that UrbanSim is the most disaggregate of the frameworks reviewed.

Its disaggregation allows detailed analyses, but also implies significant requirements for data—in particular, UrbanSim requires disaggregate data (traditionally at the 150 meter by 150 meter gridcell level) for employment, households, and buildings. The data needed to run UrbanSim and their availability in the United States context is discussed in Duthie *et al.* (2007), who note that disaggregate data required for UrbanSim may take months or even a few years to refine to an acceptable level of reliability. While such data are not always easily available, most regions normally have readily available data in a more aggregate form, often at the level of traffic analysis zone (TAZ) or other municipal divisions.

In addition to the above sources, many articles and reports have been written about UrbanSim in the formal and grey literatures. For a recent description of this literature, refer to Patterson and Bierlaire (2010). To summarize, the work on UrbanSim has tended to concentrate on descriptions of UrbanSim itself and its applications (e.g. Waddell 2001; Waddell *et al.* 2007a), computing science aspects of UrbanSim (e.g. Noth *et al.* 2003), and methodological developments that have used data relating to, or resulting from, UrbanSim (e.g. de Palma *et al.* 2007). However, within this literature, none looks directly at the suitability of using aggregate data in this disaggregate context (although the issue is broached by Duthie *et al.* 2007). This paper directly addresses this question.

Spatial disaggregation issues are most commonly discussed by geographers and natural scientists, whose models are based on cellular automata approach (see the review of Irwin and Geoghegan 2001). Although there is no explicit propagation from a gridcell to its neighbors, UrbanSim nevertheless has some features in common with cellular automata, e.g. its "within walking distance" concept. It is worth noting that cellular models themselves are viewed by some authors as capable of bridging the gap between aggregate and disaggregate description (Couclelis 1985) or between absolute and relative space through geo-algebra (Couclelis 1997; Takeyama and Couclelis 1997). The practical implementation of the latter ideas can be seen today in fully operational GIS tools, such as ESRI's Spatial Analyst. In the context of the current study, an interesting example of the application of cellular automata is Liu and Andersson (2004), where, as highlighted by Benenson and Torrens (2004), the influence of nearest neighbors is considered at the highest resolution, while an increase in the distance of neighbors' influence is considered in a more aggregate manner.

Briassoulis (2001) has considered the question of data disaggregation in the integrated analysis of land use change. The data available and the methods used to disaggregate the data in the two case studies are described using her framework of analysis. In particular, we discuss:

- the availability of adequate and proper data to disaggregate,
- the georeferencing of disaggregated data, and
- the ease and cost of disaggregation.

We also apply her criteria of data compatibility, consistency, and reliability to the data used in the two case studies and consider her general question, "How to insure that the disaggregate data measure the concept of interest?"

3 Functioning and data requirements

In order to understand data requirements and how these requirements were overcome in these case studies, a basic understanding of UrbanSim is needed. UrbanSim is composed of a number of models that together predict the location of households, jobs, and new real estate developments. The study region is structured geographically by at least two levels: at a more aggregate level, the region is structured at the TAZs of the transportation model used with UrbanSim, whereas at the disaggregate level, the region is structured by gridcells[1]. Gridcells are geographical units that have traditionally been 150 mby 150 m. Households, jobs, and buildings are located in gridcells. Thus, using rather traditional modifiable geographical units, UrbanSim also includes the spatially non-modifiable elements, though their resolution is not completely "atomic" as in geosimulation (Benenson and Torrens 2004).

For simulation of future years, UrbanSim requires exogenous data—control totals for population and employment. The probabilities that households or jobs will move from their current location are user-defined. Every simulation year, lists of the new and relocating households and jobs, as well as new development projects, are created. The households, jobs, and real estate developments are then placed with multinomial logit models (MNL)[2]—the location choice models. The land price model (LPM) is used to update the value of land, while the residential land share model (RLSM) calculates the share of residential land across the region. The LPM and the RLSM are ordinary least squares (OLS) regression models. Estimation of these models for the region studied is how UrbanSim is "tailored" to each application.

The backbone of UrbanSim is the base year database. In this relational database, the main UrbanSim data are found in six tables: gridcells, households, jobs, buildings, development event history, and development constraints tables. The gridcells table is central and links all the other tables. It identifies and characterizes each gridcell in the urban system (see following section).

Each record of the households table represents one household. Households are characterized by socio-economic attributes and the gridcell in which they are found. Each record of the jobs table represents one job. Jobs are characterized by industrial sector and location. Each record of the buildings table identifies the building's location and its characteristics.

The development event history table contains information on historical developments, characterizing them and noting where they took place. The development project transition model samples from this table to create developments in simulation years.

The development constraints table identifies what constraints are placed on different types of gridcells. These can be zoning constraints, physical constraints (e.g. no building in stream buffers), or idiosyncratic individual gridcell constraints.

UrbanSim data requirements include at a minimum for a region, a record for:

- each job, its characteristics, and in which gridcell it is located,
- each household, its characteristics, and in which gridcell it is located,
- each building, its characteristics, and in which gridcell it is located, and
- each gridcell with its many characteristics.

[1] In the latest development release of UrbanSim, it is possible to use land parcels as a geographical unit, but stable versions of UrbanSim use gridcells. The use of zonal models remains experimental and unstable.

[2] See Ben-Akiva and Lerman (1985) or Ben-Akiva and Bierlaire (2003) for more on discrete choice models.

4 The Eugene package as a template

The first UrbanSim application was developed for the Eugene/Springfield, Oregon region in the United States (Waddell 2000). An example base year dataset of Eugene/Springfield is distributed with UrbanSim. Both applications described here used the Eugene dataset as a starting point. The approach adopted was to gain experience and understand the data requirements of UrbanSim through the examination of the Eugene dataset; it was then used as the structure on which to build the new applications. First, the main data (see Section 3 above) of the Eugene base year database was replaced with data for the new region, then the various models were replaced after having been estimated with the new data.

The disaggregate data in the Eugene package is comprehensive. The gridcells table includes not only data on environmental and planning attributes, but also such detailed data as area of premises occupied by different real estate types and property value with separation between land and improvement values. The buildings table contains such detailed data as area, building type, year built, number of residential units, and improvement value for more than 9000 buildings. The households table in the Eugene dataset includes about 200 000 people. Moreover, there is comprehensive development history data, development constraints data, and various UrbanSim constants, which refer to geographical, demographic, and planning data, as well as to core models parameters. The list of working UrbanSim models in the Eugene package consists of the household location choice model (HLCM), the employment location choice model (ELCM), the development project location choice model (DPLCM), the RLSM, and the LPM.

5 The case study regions

The motivation for research in both cases was to create a working application of UrbanSim, while balancing the challenges of time-consuming data preparation and model development with limited time and human resources. In other words, it was necessary to start modeling as quickly as possible, even if initial results were not completely realistic. Two teams working on two cities faced the same challenge and found solutions in the use of aggregate data, but chose different approaches.

The two case study regions were chosen because of the availability of significant land use (employment, population, real estate, land price, etc.) and transportation data. Brussels was chosen as a case study because of the availability of land use and transportation data that had been used for the development of a TRANUS-integrated model of the Brussels region. Lyon was chosen because significant aggregate land use data existed, and also transport data could be calculated, exploiting the available Lyon transportation model. While both teams had access to aggregate data, they approached the use of these data in differently. In the Brussels case, aggregate data were disaggregated from the zonal (TAZs) to the gridcell level. In Lyon, TAZs were considered as the unit of analysis, i.e. as gridcells.

Summary statistical information for Brussels and Lyon is presented in Table 1.

Brussels is the capital of Belgium and home to many international organizations. It is perhaps best known for its importance relative to the European Union—among other EU institutions, the European Commission and the Council of the European Union are located in Brussels. The study region covers 4361 km^2 around the city of Brussels and includes 139 town-

Table 1: Summary information on case study regions.

Parameter	Brussels	Lyon
Population (million)	2.9	1.6
Area (km^2)	4361	3325
Households (thousand)	1293	662
Jobs (thousand)	1353	697

ships in parts of Wallonia (French-speaking area to the south of the region) as well as Flanders (Flemish-speaking area to the north). In addition to Brussels, the region also incorporates a number of other important cities, with Mechelen, Aalst, and Leuven being the largest. As such, the study region represents roughly 15 percent of the entire country of Belgium.

Figure 1 shows household and employment density across the region. The highest concentration of households and jobs are found in and around the center of the region. There is also significant employment and household density near and around the larger population centers of the region. Circles (generally with higher densities) are found inside the larger townships and represent the central area of these townships. With the inclusion of these township centers, there are 152 total TAZs.

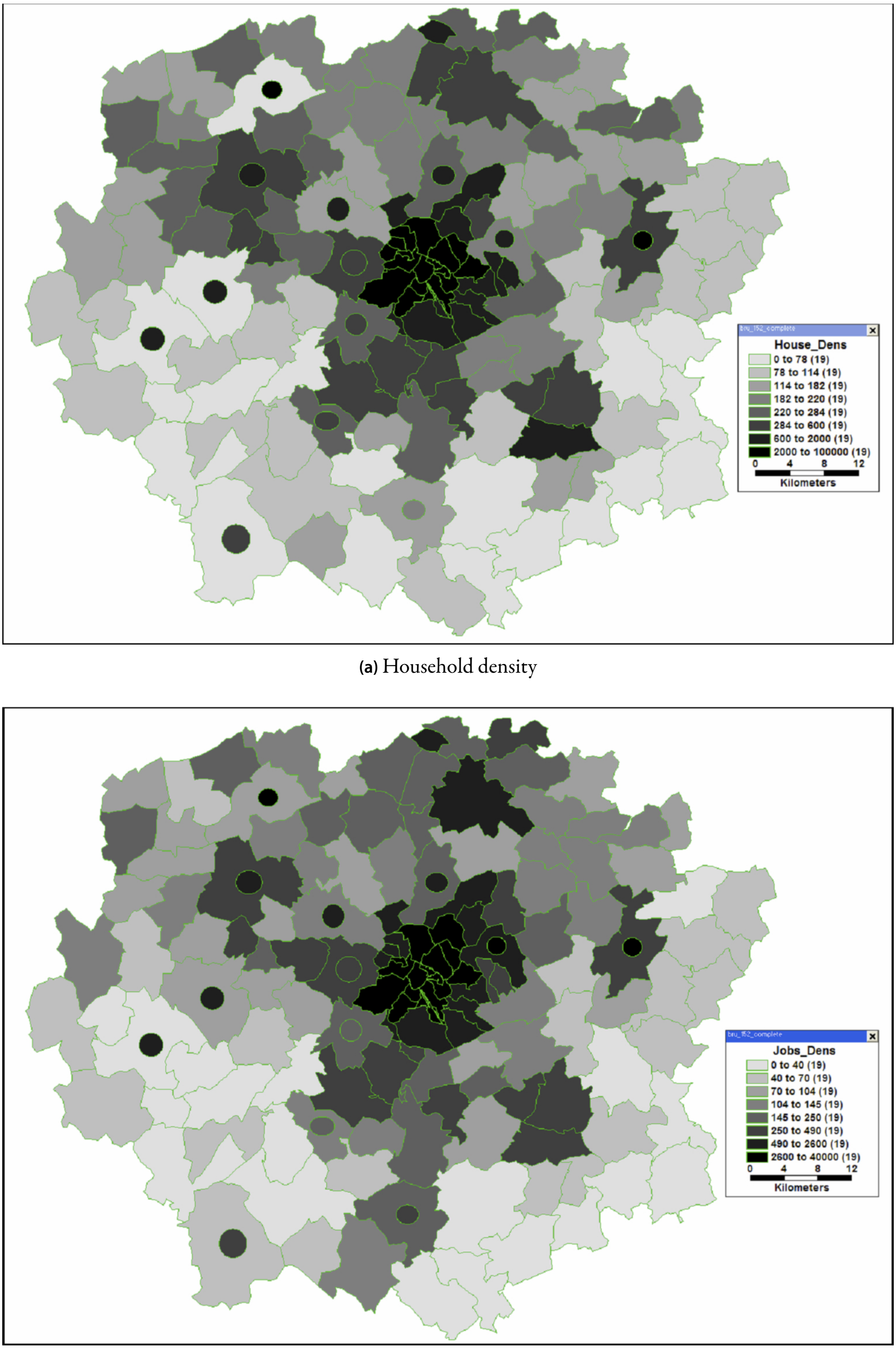
With 1.6 million inhabitants, the Lyon urban area is the second largest agglomeration by population in France. Administratively, the territory occupies parts of the départements of Rhône, Ain, Isère, and Loire. Figure 2 shows population and employment density across the region. The central part of the agglomeration, with a population of 613 000 people, consists of the cities of Lyon and Villeurbanne. These two cities, having a common planning structure and transportation network, make up the core of the region and have the highest concentration of population and employment. The rest of the region is much less urbanized, primarily agricultural, and has lower population densities. Apart from few cases mainly in the south, there are no significant centers in the periphery. There are 777 TAZs in the region.

6 Available data

6.1 Available data in Brussels

Thanks to a research partnership with the Brussels transportation engineering firm Stratec, the research team had access to data that had been used for another integrated model, TRANUS. TRANUS works with larger geographical units (152 zones for the Brussels region). Almost all of the data used for UrbanSim came from the TRANUS data.

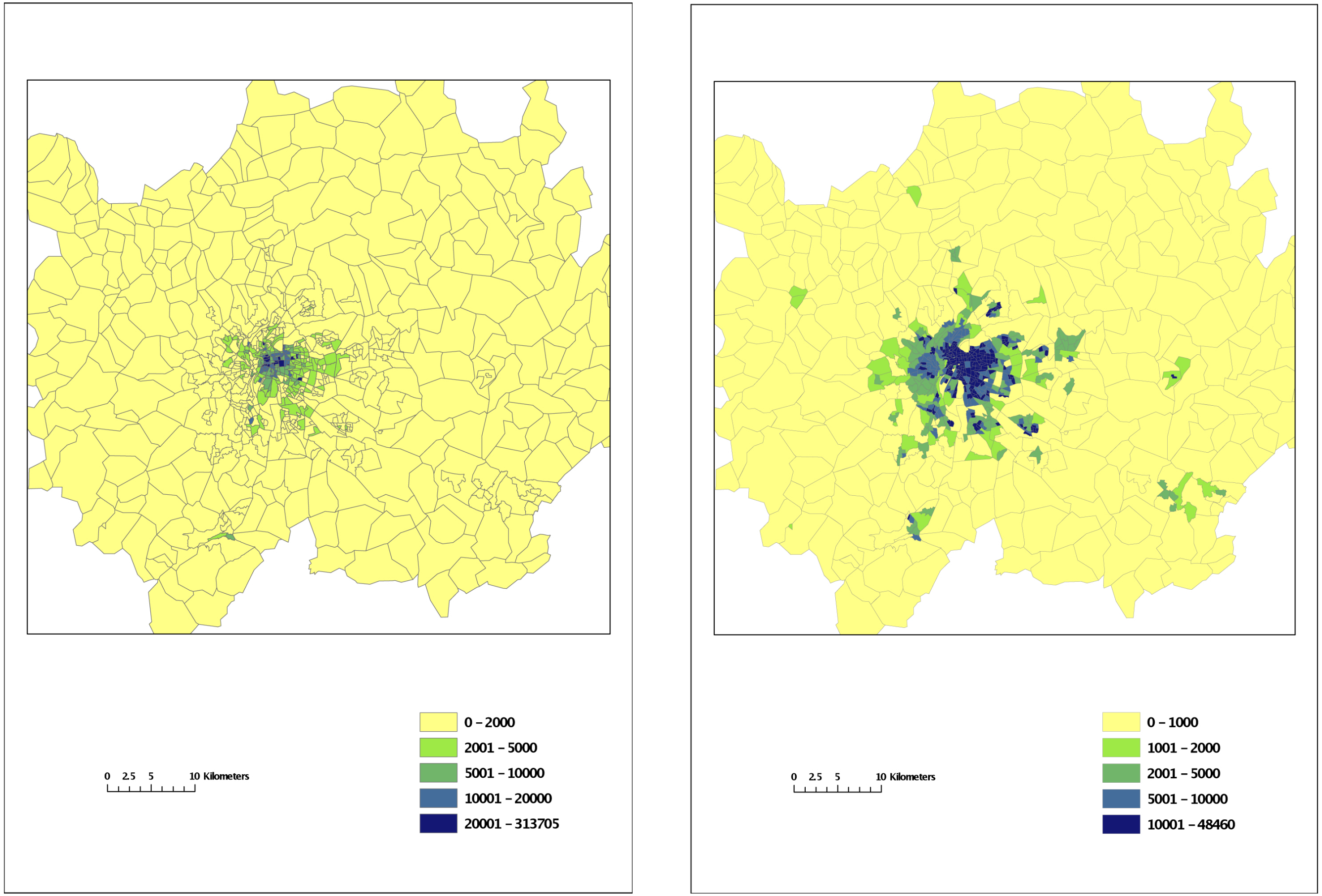
The TRANUS dataset included household and job totals per zone. Household data was relatively coarse, with households being divided into seven different types with categorical characteristics (e.g. families with children, families without children, etc.). Jobs were categorized by industrial sector, of which there were 13. There was also information on overall surface area by zone, but was not ultimately used. Other zonal data included land prices for residential and non-residential land, interzonal travel times, and generalized costs that were estimated with the TRANUS transportation model.



(a) Household density

(b) Employment density

Figure 1: Household and employment density in the Brussels region in 2001.



(a) Population density

(b) Employment density

Figure 2: Population and employment density in Lyon.

In addition to zonal data, there were GIS data on road infrastructure (highways and main arterials) as well as zoning data that covered the entire region. For some types of land use, the zoning data were quite good; however, there were limitations in zoning for commercial land uses. In particular, there were very few areas (and therefore gridcells) that were zoned as commercial, especially outside of the central part of the region.

Finally, there were some data that were essential to the way the rest of the data was disaggregated but that were not available, such as household and employment relocation rates. Assumptions, often based on the experience of Stratec, were used for these types of data.

A summary of available data can be found in Table 2. As can be seen, a great deal of data normally required for UrbanSim was not available. The most obvious and important missing data related to buildings and their characteristics. There were no individual building data and therefore no information on improvement values. Equally crucial, there were no data on historical development (i.e. buildings built in the past).

6.2 Available data in Lyon

In Lyon, there are 777 TAZs roughly comparable to municipalities, whose population usually does not exceed 4000 inhabitants. Estimated public transit travel times for the morning peak (2007) were obtained for the current study from the MOSART transportation model for Lyon. Public transport, which is well-developed in the urban area, includes buses, trolleybuses, tramways, metro, and commuter trains.

For each zone, the available data includes centroid coordinates, number of residential units, distances to highway and arterial, percent coverage for water, open space, residential land, and industrial land for 1999 (population refers to 1999; income to 2006). For the majority of zones, average price per square meter for a combination of apartments and houses was also available.

The synthetic households in Lyon distributed among zones were imputed from the INSEE[3] data. Compared with the Eugene households table, there is no information about the number of children or household race. The absence of data about racial composition is a peculiarity of French statistics. Instead, there are data on home ownership status and job status of head of household. Though there are some data about relocation rates for households, there is no link with income level.

The available data on jobs in Lyon in 1999 contained the total number of jobs in each zone and the percentages of jobs in four employment sectors in 304 communes, which consist of one or several zones. The sectors are agriculture, industry, construction, and the tertiary sector.

Another peculiarity of French statistics is that commercial employment is not considered a separate employment sector. Instead, there is the tertiary sector, which includes commercial jobs as well as jobs in education, medicine, transport, public services, etc.

Although different temporal references in the data imply inconsistency, it was not possible to collect all the data for the same reference year. As for Brussels, many important data issues for Lyon are missing (e.g. buildings, historical construction, development constraints, etc.). The summary of available data is in Table 2.

[3] Institut national de la statistique et des études économiques (National Institute for Statistics and Economic Studies).

Table 2: Summary of available data.

Data	Brussels	Lyon
Vacancy rates; household relocation rates	Assumptions from Stratec	From INSEE
Employment relocation rates	Assumptions from Stratec	n/a[1]
Transport measures	TRANUS interzonal logsums	MOSART interzonal travel times by public transport
Jobs	From TRANUS base data: ● zonal employment ● 13 sectors	From INSEE: ● zonal employment ● 4 sectors in communes
Households	From TRANUS base data: ● zonal population ● 7 household types	From INSEE: ● zonal population ● 3 income groups
Development constraints	Stratec GIS zoning data	n/a
Land and improvement values	From TRANUS base data: ● residential land values ● non-residential land values	From OTIF[2] and Perval[3]: ● residential improvement values
Buildings	n/a	n/a
Historical construction data	n/a	n/a

[1] *n/a: data not available*

[2] *Observatoire des transactions immobilières et foncières, Lyon*

[3] *Perval, France*

7 Data preparation and disaggregation

7.1 Data preparation and disaggregation in Brussels

Description of this model has been documented in various projects and technical reports (see Patterson and Bierlaire 2007; Samartzis 2007; Singh 2008; Stoitzev and Zemzemi 2008). The application for Brussels used the example of Eugene dataset.

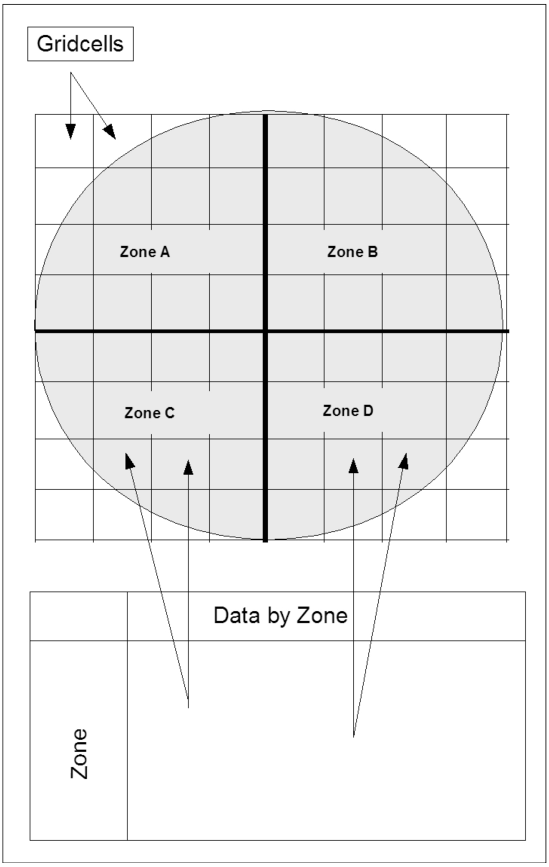


Figure 3: Zonal data disaggregation to gridcell level.

Figure 3 presents a graphical representation of the disaggregation from the zonal to the gridcell level. First, a system of gridcells was created with gridcell dimensions being the same as for Eugene (150 mby 150 m). This resulted in a grid of roughly 193 000 gridcells for the region as a whole. Geographical characteristics of gridcells were assigned to the extent that data were available. Particularly crucial for the following steps was assigning planning types to each of the gridcells using the GIS layer of zoning described above. Data on built form (residential units, surface area by building type, etc.) were included after the creation of the buildings table that first required the households and jobs tables as described below.

Second, the households table was created by disaggregating households to residential gridcells in their respective zones. Household characteristics were randomly assigned to households based on their household category and readily available public data (e.g. age distributions from the Belgian statistical agency StatBel). Some data (e.g. income) were not easily available, so coarse estimates were used.

Third, the jobs table was created by disaggregating jobs to appropriately zoned gridcells (e.g. industrial jobs of a given zone were randomly assigned to industrial gridcells in the same zone). All jobs of a given sector were assigned to the same type of building, of which there were only four (e.g. industrial jobs were assumed to be housed in industrial buildings).

After the overlay with a GIS layer of gridcells, zonal boundaries of TAZs and planning land use zones became rectangular, i.e. distorted, meaning that the criterion of data compatibility is not completely satisfied. Assigning households and jobs to gridcells through random selection means that additional random errors arise during the disaggregation process.

As mentioned above, aggregate information on building surface areas existed by zone in the TRANUS data. One option would have been disaggregating surface areas in such a way that fictional buildings were created in gridcells. At the same time, there were no data on residential units, which is required for UrbanSim. It was reasoned that it was more logical to use households and jobs and to determine the amount of residential units and surface areas, but in order to do this, it was necessary to have vacancy rates, or at least estimates of vacancy rates. No vacancy rate data were readily available, so assumptions were used. Residential vacancy rates were assumed to be highest (10 percent) in the city center and lowest (2 percent) in the extreme periphery.[4] Non-residential vacancy rates were assumed to be a constant 10 percent throughout the region.

Buildings were created in gridcells in order to: 1) house the jobs or households present, and 2) account for vacancy rates. Building characteristics were a function of surface area or residential units. An example of industrial buildings will illustrate the process of building creation. First, gridcells are assigned their zoning type. For a given zone, industrial jobs were assigned randomly to the industrial gridcells of the zone. For each gridcell containing industrial jobs, an industrial building with surface area sufficient to house the jobs (including vacancy rates) was created. Its characteristics (i.e. surface area, improvement value, etc.) were determined as a function of the number of jobs. Buildings for all jobs and households were created in the same way. Once buildings were created, their characteristics could be used to fill in the missing information on built form in the gridcells table. This is represented graphically in Figure 4.

Historical data on jobs and employment from 1991 were used to create the development event history table. Buildings were randomly selected to house the new populations or jobs that appeared between 1991 and 2001, with each building representing one development event. For example, if there was an increase in the number of industrial jobs in a zone between 1991 and 2001, enough buildings to house this increase were randomly selected as having been built over that time period. Each of the buildings represented an entry in the development event history table.

The result of these processes was that all households and jobs in the TRANUS data were assigned individual characteristics and gridcells to which they were associated, and fictional buildings constructed to house them. It was then possible to use these synthetic households, jobs, and buildings to run UrbanSim simulations, particularly to calibrate the disaggregate loca-

[4] This pattern of high vacancy rates in the center of the city is unusual for European agglomerations that tend to have the lowest vacancy rates at the center. This is one of the reasons that Brussels has earned the reputation as the "most American city in Europe."

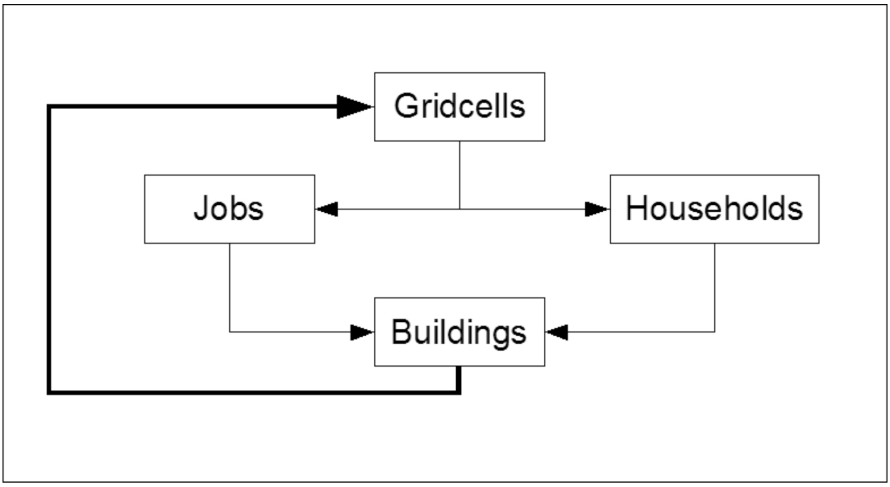


Figure 4: Representaiton of method for incorporating data on buildings.

tion choice models (based on individual households, jobs, and buildings) and the LPM (based on individual gridcells) using what was initially aggregate data.

7.2 Data preparation and disaggregation in Lyon

In Lyon, another starting point was chosen. In order to save time for data disaggregation and to avoid the additional errors connected with this process, the same number of gridcells and zones were created, with each gridcell being mapped to one zone, whereas buildings were not included in the modeling. As such, the centroids of zones are considered as centroids of gridcells and so an irregular network of gridcells was created, resulting in a more complicated problem of incompatibility in spatial dimension than with the Brussels application. The cell size was changed to 100 by 100 meters, because the shortest distance between these centroids is slightly more than this distance. In the Lyon application, the irregular network of gridcells leads to systematic errors when applying the “within walking distance” concept. The majority of gridcells have few, if any, adjacent neighbors and few neighbors in surrounding areas. To compensate for this distortion, the walking distance circle radius is increased to 1200 meters, though this compensation improves the measure only partly. Most of the other UrbanSim constants have the same values as for Eugene.

The available data in Lyon were not sufficient to fill in all the tables and feed all the models in UrbanSim. For example, consider the gridcells table in more detail. While data on some environmental attributes for Lyon were available, some other data, (e.g. square footage per job) in different sectors were used from the Eugene dataset. Moreover, some ad hoc assumptions were made for the area of premises occupied by different real estate types and for property value attributes.

The only available data on property value is the average real estate price per square meter in 657 zones. For the remaining zones, this attribute was incorporated by GIS applying a triangulated irregular network (TIN) model and converting it to a raster. Residential improvement value in each gridcell was estimated as a product of the average real estate price per square meter, the average apartment area, and the number of residential units. Possibly the coarsest assumptions are done for land value—for instance, residential land value is assumed to be 20 percent of residential improvement value.

According to the data available, the current average residential vacancy rate in the Lyon region is five percent. This figure is used as a target total residential vacancy rate. For household relocation probabilities, ad hoc assumptions were used. The target total non-residential vacancy rate, as well as probabilities of job relocation, were taken from the Eugene template.

Percentages of jobs in employment sectors in communes were disaggregated to zones considering the total number of jobs in each zone. Each fiftieth job was considered as home-based; the remaining industrial, construction, and agricultural jobs are assumed to be located in the industrial building type. Of the services jobs, three-quarters are assumed to occupy governmental buildings and one-quarter chosen randomly is assumed to be located in commercial buildings.

In the development event history table, each gridcell is presented once, with a scheduled year generated randomly from the interval of recent years, and with development size as a percentage of what exists for the base year. Thus, industrial square footage and improvement value equal to 10 percent, while residential, commercial, and governmental areas and values equal to five percent of the corresponding magnitudes from the gridcells table.

7.3 Data Disaggregation and the Briassoulis Criteria

As mentioned in the background section, it is relevant in the present context to evaluate the data and the disaggregation process in a more formal framework. Due to the applicability of Briassoulis' work, her framework is used here. For the most part, the data used are consistent with Briassoulis' criteria, with the exception of some aspects in both applications.

In terms of data availability, the most appropriate data was not available given the resources available for the two case studies; one of the objectives of the work was to determine if aggregate data could be used to overcome this problem.

Georeferencing of the disaggregate data and the ease (cost) of disaggregation can be addressed at the same time. In Lyon, since there was no disaggregation, there were no costs due to disaggregation or other issues related to georeferencing of disaggregate data. For Brussels, however, there were costs associated with disaggregation. First, considerable effort was required to disaggregate the data to the gridcell level. Second, additional random error and some spatial incompatibility were introduced into the data.

The compatibility of the different geographical systems (i.e. zones vs. gridcells) is not satisfied in both applications. For Brussels, there was not strict compatibility since TAZ's boundary was able to cross a gridcell. In Lyon, TAZs were "transformed" to gridcells of equal size. At the same time, the underlying data (e.g. jobs) were forced to be compatible—for example, the number of jobs in a zone was the same as the sum of all jobs in all of the gridcells of that zone.

Consistency means that the same definitions and the same temporal references are applied for the different geographical systems. There is some temporal inconsistency in the case of Lyon mentioned in Subsection 6.1. In respect to space, the disaggregation process in both case studies is consistent.

For reliability, or how likely the data are to represent real phenomena, this is the characteristic to which both cases adhere the least. This is because there is an additional random component in jobs and household distribution in Brussels, as well as for the job sectors in Lyon, where it refers to transition from communes to TAZs.

The final element of Briassoulis' framework to consider is her question, "How can you insure that the disaggregate data measure the concept of interest?" In this context, the concept of interest is the development of an urban area. Section 9 examines the question of how well the models simulate population progression in the two case studies.

8 Model estimation

Given the differences in the available data, different variables were used in the models of the two regions. The models can be divided between discrete choice models and traditional regression models (see Section 3). For the location choice models, the dependent variables are always gridcells; that is, a random sample of different alternative gridcell locations for each of the households, jobs, and real estate developments. By default, UrbanSim randomly samples 29 alternatives for a total of 30 alternatives for each locational decision. This random default value was kept for both applications. For the regression models, the dependent variable is the land value, residential land share, or real estate price of all the gridcells. Table 7 in the Appendix provides a summary of the models that were estimated with a few of their salient characteristics.

For the MNL models, the likelihood ratio test is presented, whereas for the OLS models, the adjusted R-squared is reported. The UrbanSim default value of the number of alternatives was not changed in both applications. For estimation of the HLCM, a 10 percent random selection of households was used in both cases. Using actual households and locations from survey data, as prescribed by the UrbanSim Users Guide (UrbanSim Project 2008), would have been a better sampling strategy, but the only available data was synthesized households, and information about recent movers was not available either. Thus simple random sampling was used, which was appropriate in providing sufficient information for estimating all the parameters (see Ben-Akiva and Lerman 1985; Train 2003).

The following two subsections concentrate on only a couple of models from each case study to provide a sense of the type and quality of the models to have been estimated using the aggregate data.

8.1 Model estimation in Brussels

All together, twelve different models (including submodels) were estimated for the Brussels UrbanSim application (Table 7). For the location choice models, the dependent variable for each was the location (gridcell) of an individual household, job, or real estate development and 29 random alternative gridcells. The locations of the households, jobs, and developments were determined in the disaggregation process. For example, each household was randomly assigned to one of the gridcells located in the zone where the household (actually) resided and that was zoned for residential purposes. The dependent variable for the land price model was the land price for each gridcell—this was the land price of the zone in which the gridcell was found.

In general, it was difficult to include many variables in the models. Whereas models in well-developed applications, such as that reported in Waddell et al. (2007a), contain dozens of variables, almost all of the models estimated for Brussels counted fewer than ten, for three reasons.

First, the data required to calculate many of the variables were not available. In many cases, proxies were used, but to the extent possible such variables were not included in models. One example was building or gridcell improvement values. Since all improvement values were simply functions of residential or non-residential surface area, they could not be used in the models and therefore reduced the total number of potential variables to be used.

The two other reasons for the small number of variables in the models are a direct result of the disaggregation process. One has to do with the small number of observations for some models. Models that suffered the most from this problem were the DPLCM. Since there was no data on historical construction, it had to be constructed, and the way in which the buildings data were constructed meant that far fewer synthetic buildings were created than were built in reality (see Section 7.3 on disaggregation above). Any positive changes in employment or population in a zone were considered to be housed in newly constructed buildings. In order to satisfy the construction needs of the new households and jobs, buildings in the historical development events table were sampled from the existing synthetic buildings. In other words, relatively few buildings were available to be selected for the development event history table and, as a result, there were relatively few observations of buildings built between 1991 and 2001, particularly the commercial and industrial submodels.

The last reason is based mainly on intuition, since it is difficult to prove the contrary; that is, there are fewer variables in the models because the disaggregation process added too much noise to the disaggregated, synthetic data to allow otherwise meaningful variables to test statistically significant. There are many examples of situations where this might be the case; one example is proximity to highways. In general, we would expect(all else being equal) that jobs or households would like to locate closer to highways (within reason). However, in none of the models was this variable statistically significant, despite having accurate GIS data on the location of highways and main arterials. Given that synthetic households, jobs, and buildings were randomly attributed to appropriately zoned gridcells in large TAZs, this should not be surprising.

Table 3: HLCM for Brussels.

Variable	Coefficient	t-value
1. Cost to income ratio	−0.07	−2.2
2. % high inc. households wwd if high inc.	0.03	22.3
3. % low inc. households wwd if high inc.	<−0.01	2.9
4. % low inc. households wwd if low inc.	0.06	55.4
5. Travel time to CBD	<−0.01	−4.2
6. Dummy for location in Flanders	−0.03	−3.1

Null log-likelihood: 440 982.25
Likelihood ratio test: 3479.87
Alternatives: 30
Number of observations: 129 655
Convergence statistic: 0.000 076 2

Despite these weaknesses and the relatively small number of variables, the estimated models were generally pleasantly surprising, with the most important variables (e.g. land price, accessibility measures, etc.) usually coming out significant with the right sign. This was not always the case.

An example of a typical model is the HLCM shown in Table 3. The model contains six variables and was estimated on a 10 percent sample of households in the region. According to the model, households prefer locations that are less expensive, all else being equal (Variable 1). They also prefer to live near households with similar incomes (Variables 2 and 4), although high-income families show some affinity to being near low-income households (Variable 3). Geographically, households prefer being closer to the central business district (CBD) (Variable 5) and locations in the Central Brussels Region or Wallonia (i.e. not in Flanders) (Variable 6).

Table 4: Commercial non-sedentary services to enterprises for Brussels.

Variable	Coefficient	t-value
1. Land value	−0.78	−26.17
2. Employment wwd	0.48	23.91
3. Work access to employment	0.06	2.05
4. Work access to population	2.43	19.71
5. Job of same sector	0.01	56.74

Null log-likelihood: −20 913.96
Likelihood ratio test: 8612.07
Alternatives: 30
Number of observations: 6149
Convergence statistic: 0.000 913

Another typical example is the ELCM for jobs in the "non-sedentary services to enterprises" sector in Table 4. It was estimated using a ten percent sample of jobs, as was the case for households. It has only five variables; however, the variables present have intuitive signs and high t-values. This suggests that jobs of this sector tend to locate where land prices are lower (Variable 1), there are other jobs within walking distance (Variable 2), there is high accessibility to employment and population (Variables 3 and 4), and there are jobs of the same sector (Variable 5). All together, a surprisingly acceptable model even if it does not have many variables.

8.2 Model estimation in Lyon

In the Lyon application, twelve models (including submodels) were estimated (Table 7) for the base year of 1999. For the location choice models, the dependent variable for each was the location (gridcell/zone) of an individual household, job, or real estate development and 29 random alternative gridcells/zones. The dependent variable for the LPM was the land price for each gridcell. Price per square meter for the REPM was available for 657 gridcells/zones; data for the other 120 zones were interpolated. The dependent variable for the RLSM was the proportion of residential land for each gridcell.

The small number of variables can be explained primarily by the lack of available data. Others' experience with UrbanSim (e.g. Nguyen-Luong 2008) also shows that it is better to start with simple models containing few variables than to build very sophisticated models with plenty of variables, because all the models should be implemented in UrbanSim and used for simulation.

One model in particular, the HLCM, is described in detail as an example. The model was estimated for all of the household types, as in the Eugene example. For estimation, a 10 percent random selection of households was used. The attributes significant at the five percent level are presented in Table 5. In the estimation, 66 225 random households from 765 gridcells are used. The extraction of different random samples of the same size does not significantly change coefficients and their t-values except in the case of the travel time to CBD. This variable, which is only marginally significant in Table 5, can easily lose its significance. The log of this variable is always insignificant.

Households do not like to live near households of different income levels (Variables 1 and 2). Interestingly, the reported specification works better than that with the corresponding variables for similar income. The number of households (Variable 3) and percent of households with own accommodation (Variable 4) both increase the utility of a particular location (though the latter variable to smaller degree), whereas the log of the number of residential units (Variable 5) is negative (i.e. the locations with fewer residential units are preferable). Attempts to add the property value variables failed.

Table 5: HLCM for Lyon.

Variable	Coefficient	t-value
1. % high income households wwd if low income	−0.08	−36.08
2. % low income households wwd if high income	−0.03	−22.15
3. Log number of households	1.46	19.78
4. Log % households with own accommodation	0.02	2.52
5. Log residential units	−1.43	−19.14
6. Travel time to CBD	< 0.01	1.99

Null log-likelihood: −225 244.3
Likelihood ratio test: −223 963.45
Alternatives: 2561.7
Number of observations: 66 225
Convergence statistic: 0.000 01

The "size variables" of the number of households and residential units included in the model are highly significant. As our model estimated and then simulated at the same geographical level of gridcells, we do not aggregate alternatives; therefore, the scaling restriction for "size variables" (see Ben-Akiva and Lerman 1985) is not applied in this case.

The positive sign and low significance of travel time to the Lyon CBD can be explained by the distortion from the heterogeneity of zone sizes—outer zones are usually substantially larger and consquently many of them have larger populations despite lower densities, meaning that HLCM attempts to explain why zones with larger travel times to CBD are more attractive. The apparent attractiveness of outer zones is illustrated by the map of population (Figure 5), which should be compared with the map of population density (Figure 4). However, this fact is not taken into account in the application where each zone corresponds to one gridcell. The same is true for value attributes and other parameters.

For the ELCM estimation (Table 7), we use all the industrial and home-based jobs, but only a portion of commercial jobs. Thus, we use only commercial jobs from the gridcells, where

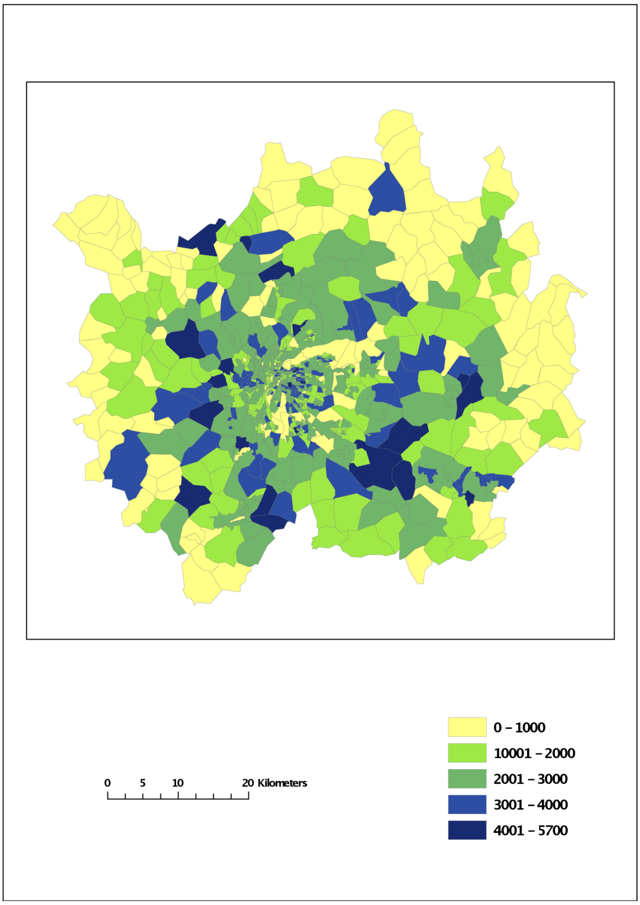


Figure 5: Zonal population in the Lyon urban area in 1999.

the number of such jobs is at least 700. There are 29 such gridcells, and the number of jobs in commercial buildings used in estimation is 34 895.

Another model example is the ELCM-industrial for industrial jobs; its variables significant at the 5% level are presented in Table 6. The industrial jobs prefer locations with higher land and improvement values (Variable 1), smaller areas of industrial premises (Variable 2), which are farther from arterial (Variable 4), but closer to highway (Variable 5), and with low accessibility to employment (Variable 5, which is calculated with travel time). The last peculiarity can be explained by the specificity of industrial sector, where jobs are often concentrated distantly from other sectors. However, Variables 1 and 2 are based on ad hoc assumptions.

The DPLCM estimation results for Lyon (Table 7) show that in the industrial specification, the small number of observations (in comparison with commercial and residential specifications) can be enough to construct a model with as few as three variables. However, the models' performance depends on assumptions made for the base year and development history. Among the OLS regression models in Lyon (Table 7), the LPM has the highest adjusted R-squared, but its dependent variable is based on assumptions, which is not the case for the REPM.

Table 6: ELCM – industrial for Lyon.

Variable	Coefficient	t-value
1. Log total value	1.59	53.77
2. Log industrial square footage	−0.86	−68.76
3. Dummy for location near arterial	−0.46	−25.65
4. Dummy for location near highway	0.29	10.59
5. Log home access to employment	−0.13	−3.39

Null log-likelihood: −398 749.58
Likelihood ratio test: −78 171.2
Alternatives: 641 156.76
Number of observations: 117 238
Convergence statistic: 0.000 583

9 Simulation results

The estimated models are limited in the amount of information they contain that can affect the evolution of the urban systems. At the same time, the models are pleasantly surprising for the most part. The true test of the models, of course, is how all of them perform together, and the following validation exercises examine how UrbanSim simulations performed. In the reported results, population is the focus because it was the only indicator for which relatively current "real" data existed.

9.1 Simulation results in Brussels

Simulation results compare surprisingly well with actual population growth in the Brussels region. Figure 6 shows a map of the difference between actual and simulated population growth rates between 2001 and 2007—for more than half of the zones (most of which represent a municipality), the difference in simulated population growth to actual growth was between 2 percent and −2 percent. All except one were within the range of ±10 percent.

More can be said about this result than simply the differences between actual and simulated growth. In particular, there is a discernible pattern of under-prediction of population relative to actual growth along a northwest axis that extends from the center of the region to the southwest (see dashed ellipse). The reason for this under-prediction appears to be the HLCM. These zones tend to correspond with zones with relatively high land values and relatively lower travel times to the CBD. As such, it appears that since population is actually higher in these zones than predicted by the simulation, the coefficient suggests exaggerated sensitivity to land prices. It also suggests that the coefficient for travel time to CBD is not sensitive enough. In further evidence, in the east of the region there is a band of zones where population has been over-predicted. These zones correspond to areas with lower land prices and higher travel times to the CBD.

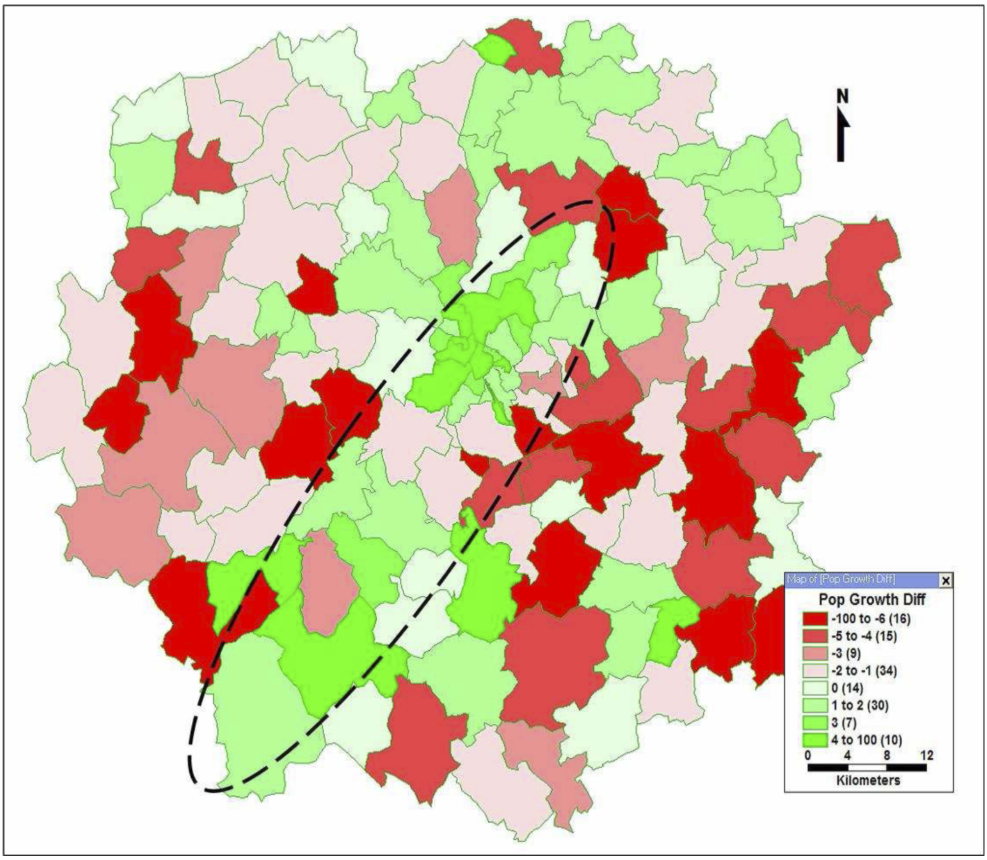


Figure 6: Difference in simulated and actual population growth, 2001–2007 (Brussels).

9.2 Simulation results in Lyon

Due to the smaller size of zones and therefore the smaller zonal populations in Lyon, it is more reasonable to focus on population itself instead of population growth. There are two problematic aspects, however. First, due to the changed methods at INSEE, only estimated population is available after 1999—we will refer to this as actual population. Second, after 1999 there were many changes in zonal boundaries and several partitions and amalgamations of zones. Of 777 zones used in 1999, only 721 had the same identifier in 2005. In the terms of Briassoulis (2001), the definitions of population and zones (in spatial dimension) do not remain constant over time, which decreases the reliability of comparison.

The comparison of predicted population with actual population in 2005 is shown in Figure 7 (for the whole territory) and in Figure 8 (for its central part). Only 721 zones are compared; the other areas are shown as white spots. Among the comparable zones, 11 percent have the difference within the interval of ±2 percent; 29 percent within the interval of ±5 percent, and 53 percent within the interval of ±10 percent. Small populations in some zones can partly explain the high differences—if there are a few people, then even negligible absolute difference between predicted and actual population leads to high relative difference. This is the case of, for example, industrial areas in the south and sparsely populated areas in the northeast in Figure 8.

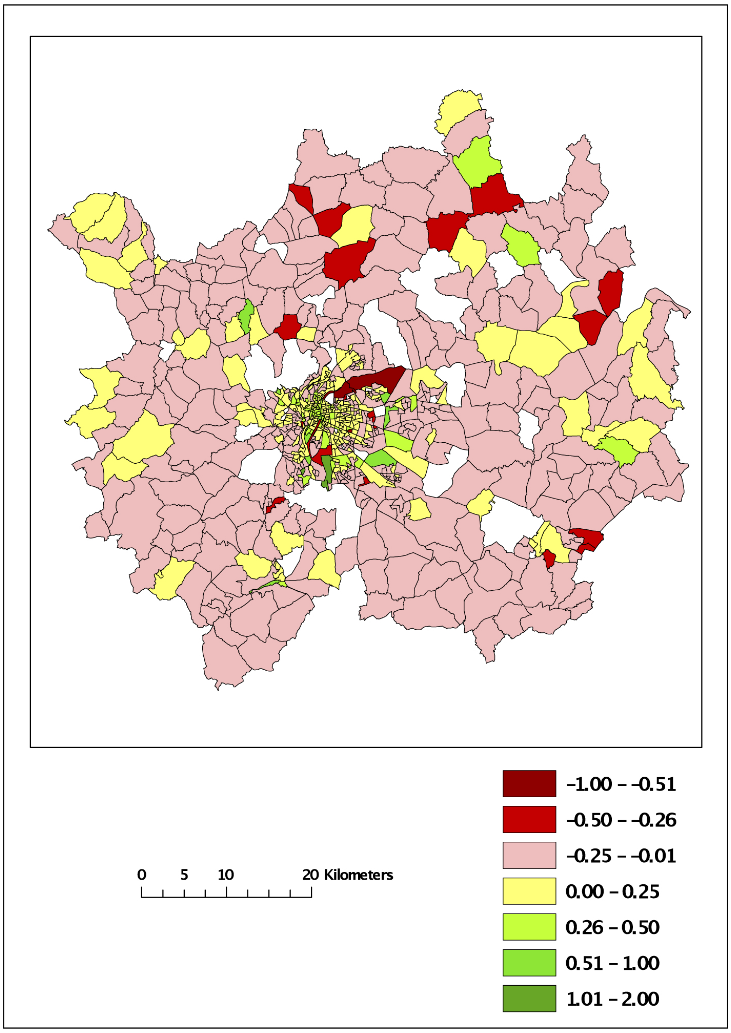


Figure 7: Difference between simulated and actual growth in Lyon urban area 2005.

Generally, population is under-predicted in outskirts and over-predicted in the central cities of Lyon and Villeurbanne. Intuitively, this can be explained by distortion arising from the heterogeneity of zone sizes, which is not reflected in the irregular network of gridcells (see Section 7.2). Another explanation is an insufficient number of variables in the HLCM (e.g. adding environmental variables would increase the attractiveness of remote areas). Figure 7 and Figure 8 are also good illustrations of the difficulty of applying the same model to both urbanized and agricultural areas.

10 Effort required and aggregate approach adopted

One critical element of evaluating the advantages and disadvantages of using UrbanSim with aggregate data is the amount of effort required to do so—a full-fledged UrbanSim application can take years to develop. A recent analysis of UrbanSim (Nguyen-Luong 2008) reports that it

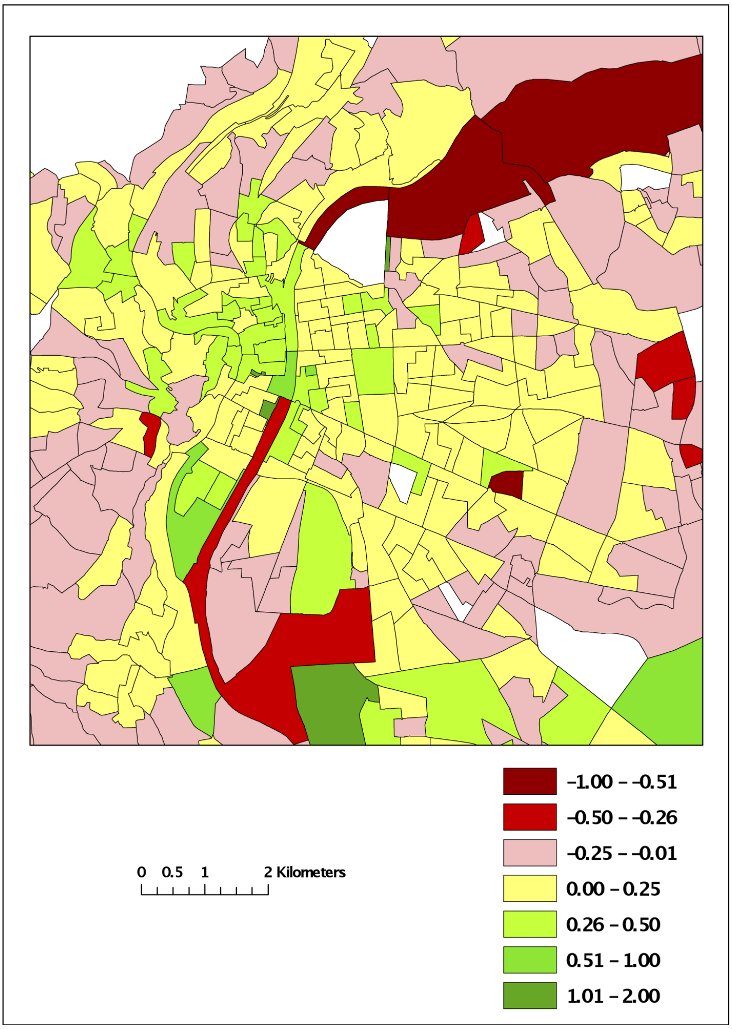


Figure 8: Difference between simulated and actual population, Lyon 2005.

took three years with four full-time staff to develop their application for Paris. This is perhaps an extreme in terms of time required, since the goal for this model was a system (transportation and land use models that run together with an integrated interface) that would run "...with the push of a button..." It does, however, give a sense of how much effort can be required to develop a full-functioning model. As such, an evaluation of UrbanSim with aggregate data must be seen in this context.

For Lyon, understanding the basic data requirements of the model, then obtaining and preparing the available data took more than half a year. The first simulation results were obtained after about two person-months of additional work. Several additional person-months were devoted to improving model performance and incorporating new data used in the reported results, but the total amount of time is difficult to measure.

For Brussels, understanding the basic data requirements and preparing available data took about two person-months of work. Another month and one half person-months were required for initial model estimation and preparation. The first simulations with the Brussels data were completed after three and a half person-months, and another one and a half person-months were spent improving the original dataset, including incorporating zoning data, to produce the results presented here. In total, around five person-months of work were required to get these results.

The effort required is partly dependent on previous experience of the people involved in model development. A wide range of knowledge and skills are used in the development of an UrbanSim model. These include knowledge of statistical modeling, data treatment, GIS, and general computer programming with emphasis on Python. Familiarity with transportation and transportation and land use modeling, as well as data availability for the region of interest, are also important ingredients. While the Brussels application was largely based on the existing TRANUS base data, the database for Lyon was created from scratch.

Based on these two case studies, an UrbanSim model with aggregate data can be developed between four and eight person-months of work—significantly less than what was reported for a fully functional application in Paris (see Nguyen-Luong 2008). While differences in sizes of the regions should be accounted for in such a comparison, it still seems that using aggregate data is a relatively low-cost solution to starting an UrbanSim project.

Assuming a research team were interested in developing an UrbanSim model with aggregate data, either approach could be suitable in different circumstances. Based on the experience of the two teams, the approach depends on the type of model the developers are interested in. In terms of overall effort, the two approaches are comparable, but particular aspects of data preparation and model estimation required different amounts of effort. The Brussels approach required more effort in order to create gridcells and buildings, whereas the Lyon approach required more effort in certain data preparation and especially in model estimation and simulation. This heightened effort was due to the small number of observations and the correction of distortions caused by the irregular coverage of the area. As such, it is not so much the effort that should guide the choice between aggregate approaches, but rather which type of model the developers are interested in.

Developers should choose the aggregate model that conforms most with the model they would like to eventually develop. If a disaggregate gridcell (or parcel level) model is the objective, then the Brussels approach would be more suitable; if a zonal model is envisaged, it would make sense to use the Lyon approach.

11 Conclusions

A number of things can be learned by comparing these two approaches of the use of UrbanSim. The first is that it is possible to start working with UrbanSim with aggregate data. The two case studies show that this can be done in at least two ways. In Lyon, aggregate data were applied directly to the TAZs while each zone corresponded to one gridcell. The approach for Brussels was to create a gridcell system and disaggregate the aggregate data to the gridcells. In both cases, not all important information was lost through disaggregation. Despite the use of aggregate data, many ad hoc assumptions, and the use of some parameters directly from the Eugene template, the applications produce surprisingly good models as well as simulation results.

Despite the surprisingly good results, there is a limit to how much disaggregate information can be drawn from aggregate data. Moreover, it does not seem possible to develop a reliable and fully operational UrbanSim application with the use of only aggregate data. The main disadvantages of data disaggregation manifest themselves in two ways. First, aggregation does not allow for the successful production of sufficient data to allow the estimation of robust models. This was seen in the case of buildings data in the Brussels context and of gridcells in the Lyon context. In the latter case, the main limitation is in the irregular network of gridcells that does not cover the whole analysed territory and leads to distortions while applying the "within walking distance" concept or ignoring density measures.[5] For this, the Brussels approach better satisfies the Briassoulis' criterion of spatial compatibility. On the other hand, the Lyon approach allows avoiding the random errors of data assignment to gridcells and in this respect better adheres to the reliability criterion. Second, the use of aggregate data introduces sufficient noise that only the most robust relationships will manifest themselves in analyses.

Finally, a particularly positive finding is that given a lack of disaggregate data, developing an application with aggregate data can be extremely useful for understanding how UrbanSim works and what it requires. The applications reported here were begun with available data as quickly as possible and were not stopped by the lack of disaggregate data. In fact, in the absence of readily available disaggregate data, it seemed a better way to proceed with application development than by initially collecting the required disaggregate data. Moreover, it is easier to understand the peculiarities of UrbanSim using limited aggregate data than to try to grasp its principles by working with large detailed datasets. As a result, we conclude that if researchers or planners were interested in evaluating the creation of a new UrbanSim application for a given region, the use of aggregate data is worthwhile and a relatively low-cost exercise to evaluate UrbanSim requirements for a fully operational model. These two UrbanSim case studies can be considered as important first steps in developing more robust applications.

Acknowledgements

In the case of Brussels, the authors would like to thank the various students who worked on the project: Lefteris Samartzis, Iordanka Stoitzev, Fatima Zemzemi, and Yash Singh. Without the assistance of Sylvie Gayda, Nathalie Schaillée, Françoise Boon, and Hugues Duchateau from Stratec, the work would have been impossible.

In the case of Lyon, this work was developed on project SMA LOCMOB partly sponsored through French ANR (project JC05_46807) and project SIMBAD sponsored by French PREDIT and ADEME. The work was also sponsored by Fondation Maison des Sciences de l'Homme through bourse Diderot. The authors are grateful in particular to Nicolas Ovtracht and Valérie Thiebaut for their important help with preparing the travel time datasets. We also thank Jean-Pierre Nicolas for sharing data from the SIMBAD project and Philippe Zuccarello and Simon Roberge for technical assistance.

References

Ben-Akiva, M. and M. Bierlaire. 2003. Discrete choice models with applications to departure time and route choice. In R. Hall, ed., *Handbook of Transportation Science*, pp. 7–38. Dordrecht: Kluwer, 2 edition.

[5] A parcel-based UrbanSim version (see Waddell *et al.* 2007b) would better suit such an approach improving spatial consistency.

Ben-Akiva, M. E. and S. R. Lerman. 1985. *Discrete choice analysis: Theory and application to travel demand*. Number 9 in MIT Press series in transportation studies. Cambridge, Mass.: MIT Press.

Benenson, I. and P. M. Torrens. 2004. Geosimulation: Object-based modeling of urban phenomena. *Computers, Environment and Urban Systems*, 28(1-2):1–8. doi: 10.1016/S0198-9715(02)00067-4.

Briassoulis, H. 2001. Policy-oriented integrated analysis of land-use change: An analysis of data needs. *Environmental Management*, 27(1):1–11. doi: 10.1007/s002670010129.

Couclelis, H. 1985. Cellular worlds: A framework for modeling micro-macro dynamics. *Environment and Planning A*, 17(5):585–596. doi: 10.1068/a170585.

Couclelis, H. 1997. From cellular automata to urban models: New principles for model development and implementation. *Environment and Planning B: Planning and Design*, 24(2):165–174. doi: 10.1068/b240165.

de Palma, A., N. Picard, and P. Waddell. 2007. Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater Paris region. *Journal of Urban Economics*, 62(2):204–230. doi: 10.1016/j.jue.2007.02.007. Essays in Honor of Kenneth A. Small.

Duthie, J., K. Kockelman, V. Valsaraj, and B. B. Zhou. 2007. Applications of integrated models of land use and transport: A comparison of ITLUP and UrbanSim land use models. In *54th North American Meetings of the Regional Science Association International*. Savannah, Georgia.

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

Irwin, E. G. and J. Geoghegan. 2001. Theory, data, methods: Developing spatially explicit economic models of land use change. *Agriculture, Ecosystems & Environment*, 85(1-3):7–24. doi: 10.1016/S0167-8809(01)00200-6.

Liu, X. and C. Andersson. 2004. Assessing the impact of temporal dynamics on land-use change modeling. *Computers, Environment and Urban Systems*, 28(1-2):107–124. doi: 10.1016/S0198-9715(02)00045-5. Geosimulation.

Nguyen-Luong, D. 2008. An integrated land use-transport model for the Paris region (SIMAURIF): Ten lessons learned after four years of development. CD-ROM of the 88th Annual Meeting of the Transportation Research Board, reference no. 09-0024.

Noth, M., A. Borning, and P. Waddell. 2003. An extensible, modular architecture for simulating urban development, transportation, and environmental impacts. *Computers, Environment and Urban Systems*, 27(2):181–203. doi: 10.1016/S0198-9715(01)00030-8.

Patterson, Z. and M. Bierlaire. 2007. An UrbanSim model of Brussels within a short timeline. Presented at the Seventh Swiss Transport Research Conference, Monte Verità, Switzerland. URL http://infoscience.epfl.ch/record/117164.

Patterson, Z. and M. Bierlaire. 2010. Development of prototype UrbanSim models. *Environment and Planning B*, 37(2):344–366.

Samartzis, L. 2007. *Modelisation de Bruxelles avec UrbanSim*. Master's thesis, EPFL – Transport and Mobility Laboratory.

Singh, Y. 2008. Modeling of Brussels with UrbanSim. Technical report, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland. URL http://transp-or2.epfl.ch/cours/projets.php?details=1.

Stoitzev, I. and F. Zemzemi. 2008. La calibration d'UrbanSim pour la ville de Bruxelles. Technical report, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland.

Takeyama, M. and H. Couclelis. 1997. Map dynamics: Integrating cellular automata and GIS through Geo-Algebra. *International Journal of Geographical Information Systems*, 11(1):73–91.

Train, K. 2003. *Discrete choice methods with simulation*. New York: Cambridge University Press.

UrbanSim Project. 2008. Opus: The open platform for urban simulation and urbansim version 4 – reference manual and users guide. Technical report, Center for Urban Simulation and Policy Analysis, University of Washinton, Seattle.

Waddell, P. 2000. A behavioral simulation model for metropolitan policy analysis and planning: Residential location and housing market components of UrbanSim. *Environment and Planning B*, 27(2):247–263. doi: 10.1068/b2627.

Waddell, P. 2001. Towards a behavioural integration of land use and transportation modeling. In D. Hensher, ed., *Travel Behaviour Research: The Leading Edge*, pp. 65–96. Paris: Pergamon.

Waddell, P. 2002. UrbanSim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297.

Waddell, P., G. F. Ulfarsson, J. P. Franklin, and J. Lobb. 2007a. Incorporating land use in metropolitan transportation planning. *Transportation Research Part A: Policy and Practice*, 41(5):382–410. doi: 10.1016/j.tra.2006.09.008. Journal Issue: Bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling.

Waddell, P., L. Wang, and B. Charlton. 2007b. Integration of a parcel-level land use model and an activity-based travel model. In *87th Annual Meeting of the Transportation Research Board*. Transportation Research Board.

Wegener, M. 1995. Current and future land use models. In *Land Use Model Conference*. Dallas.

Wegener, M. 2004. Overview of land-use transport models. In D. A. Hensher and K. Button, eds., *Transport Geography and Spatial Systems*, chapter Handbook 5 of the Handbook in Transport, pp. 127–146. Kidlington, UK: Pergamon/Elsevier Science.

Appendix: Summary of models in the two case studies

Table 7: Summary of models in two case studies.

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
HLCM	• cost/inc. ratio • % high inc. hh[2] wwd[3] if high inc. • % low inc. hh wwd if high inc. • % low inc. hh wwd if low inc. • travel time to CBD • dummy for location in Flanders	3480 22.458 (129655)	• % high inc. hh wwd if low inc. • % low inc. hh wwd if high inc. • log number of hh • log % hh with own accommodation • log residential units • travel time to CBD	2562 22.458 (66225)
		ELCM – Industrial		
Industrial	• log total value • log work access to employment • travel time to CBD • number of industrial jobs • number of heavy tertiary jobs	1017 20.515 (15176)	• is near arterial • is near highway • log home access to employment • log industrial sq.ft • log total value	641157 20.515 (117238)
Heavy Tertiary	• log total land value • log work access to population • number of industrial jobs • number of heavy industrial jobs	999 18.467 (13598)	N/A	N/A

Continued

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
Construction	N/A	N/A	● is near arterial ● is near highway ● log industrial sq.ft ● log total value	163867 18.467 (30415)
Agriculture	N/A	N/A	● is near arterial ● is near highway ● log industrial sq.ft	26341 16266 (4277)
ELCM – Commercial				
Commercial	N/A	N/A	● avg. income ● commercial sq.ft ● 3 dummies for development types ● 2 dummies for plan types ● log distance to highway ● log home access to employment ● log home access to population ● log total land value ● log travel time to CBD ● number of commercial jobs ● number of hh ● population ● number of residential units ● % of open space wwd	92889 36.123 (34895)

Continued

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
Non-sedentary services	• land value • employment wwd • work access to employment • work access to population • jobs in same sector	8612 20.515 (6749)	N/A	N/A
Sedentary services	• log total land value • log work access to employment • log work access to population • log population wwd • log employment wwd • jobs in same sector	9405 22.458 (24000)	N/A	N/A
Retail	• log total land value • log work access to population • retail jobs • in Capital Region	5473 183.467 (24000)	N/A	N/A
Local private services	• log total land value • log work access to employment • log work access to population • travel time to CBD • jobs in same sector • in Flanders	3070 20.458 19498	N/A	N/A

Continued

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
Agriculture	• log total land value • log work access to population • log total employment wwd • travel time to CBD • agricultural jobs	2508 20.515 (1511)	N/A	N/A
ELCM – Home-based	N/A	N/A	• log commercial sq.ft • log residential units • travel time to CBD	24004 16.266 (13580)
DPLCM				
Industrial	• log total land value • log basic sector employment wwd	18 13.816 (26)	• dummy for plan type • log total land value • population	98 16.266 (33)
Commercial	• in Capital Region • log work access to employment	21 13.8186 (77)	• 13 dummies for intervals of commercial sq.ft • log commercial sq.ft	328 36.123 (380)
Residential	• log total land value • travel time to CBD • in Capital Region	436 16.266 (1686)	• 3 dummies for intervals of num. res. units • 7 dummies for development types • 3 dummies for plan types • number of residential units	600 36.123 (601)

Continued

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
RLSM	N/A	N/A	• log non-residential sq.ft • home access to employment • home access to population • is outside urban growth boundary • log commercial sq.ft • log residential units • population • travel time to CBD	0.31 N/A (569)
REPM	N/A	N/A	• home access to employment • home access to population • is outside urban growth boundary • log industrial sq.ft • log number of hh • log travel time to CBD • number of commercial jobs • % low inc. hh wwd • number of residential units • total land value	0.567 N/A (777)

Continued

	Brussels		Lyon	
Model	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value[1] (obs.)	Significant Variables	LR test/adj. R^2 χ^2 Crit. Value (obs.)
LPM	• Constant • log total employment wwd • log work access to employment • log work access to population • travel time to CBD • 5 dummies for development type • log basic sector employment wwd • in Capital Region	0.517 N/A (118951)	• home access to employment • home access to population • log commercial sq.ft • log industrial sq.ft • population density • travel time to CBD	0.809 N/A (777)

[1] χ^2 Critical value at 0.1%
[2] hh: household
[3] wwd: within walking distance
[4] sq.ft: square feet
[5] inc.: income