Development of Prototype UrbanSim Models

Zachary Patterson *   Michel Bierlaire *

18 August 2008

Abstract

UrbanSim is an integrated transportation land-use model that has been under development since the late 1990s. It has received a fair bit of attention in the integrated modeling community. It is well-known primarily due to its disaggregated approach. A number of papers describing the application of UrbanSim have appeared in the formal and grey literatures. Some of these papers report on successful applications of UrbanSim with little description of the amount of effort required to develop an operational model. Those that do report on the effort and challenges of using UrbanSim suggest that substantial data and human resources are required. One recent report quantifies the human resource requirements as an interdisciplinary team of four for four years. This reputation makes many potential users think twice before developing an UrbanSim model. We believe the only way to evaluate the potential of UrbanSim is by having a good understanding of what it can do, and how much effort is required. Understanding UrbanSim, however, does not require having a fully operational model. This paper is aimed at researchers and institutions that would like to evaluate UrbanSim, but are concerned about the effort required. Based on two UrbanSim applications (Brussels in Belgium and Lausanne in Switzerland), this paper describes a procedure to develop a prototype UrbanSim model and how to use it to evaluate UrbanSim for application to a new region.

*École Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland. E-mail: {zachary.patterson, michel.bierlaire}@epfl.ch

1 Introduction

UrbanSim is a rapidly evolving integrated transportation land-use model ("integrated model" hereafter) that has been under development since 1996 by the research team of Paul Waddell at the University of Washington. It has received a fair bit of attention in the academic and grey literatures. A number of characteristics of UrbanSim have led to this interest. First, it is open source and therefore freely available and its code can be changed and adapted by whomever would like to use it. Second, it is disaggregate. Geographically, it operates at the level of gridcells (normally 150 x 150 metres) or parcels. With respect to population it operates at the level of individual households. With respect to employment it operates at the level of individual jobs or establishments. Compared to most other integrated models that operate at the level of much larger traffic analysis zones (TAZ), this characteristic of UrbanSim allows a much finer-grained approach to urban modeling.

While such a fine-grained approach allows for a great deal of flexibility in analyzing many aspects of an urban system (e.g. different planning or zoning policies), this does not come without costs. In particular, the data requirements for an operational UrbanSim model are large. Moreover, the complexity of model preparation, estimation and calibration can seem very onerous.

This paper is aimed at researchers and planners considering starting an UrbanSim project, or who would like to evaluate UrbanSim, but who are wary of the investment required to do so. It is based on the experience of developing two prototype UrbanSim models for the cities of Brussels in Belgium and Lausanne in Switzerland. It begins with a literature review and a short description of how UrbanSim works. A fourth section describes the two case studies. The fifth section describes a procedure to develop prototype UrbanSim models that can be used as a basis for further model development, or for evaluation. A final section reports on the main conclusions.

2 Literature Review

Despite a healthy literature on UrbanSim, and despite a reputation for heavy data requirements, there has been relatively little research that evaluates the difficulty of using UrbanSim. The literature that does, concentrates on the efforts required for a fully developed, operational model. In this literature there is little guidance on how to evaluate UrbanSim for a prospective application.

This literature review considers only research that has involved the use of UrbanSim directly, has been a spin-off of work on UrbanSim or that has reported on UrbanSim. The literature surrounding UrbanSim can be classified into five categories:

1. Various different descriptions of UrbanSim as it has evolved. These include: Waddell et al. (forthcoming); Waddell and Borning (2004); Davis et al. (2006); Waddell (1998); Waddell (2001) and Waddell (2000). Hunt

et al. (2005) contains a description and analysis of UrbanSim in comparison with other models.

2. Articles in the computer science literature describe UrbanSim and various aspects of the UrbanSim system in the context of software and user interface development: Noth et al. (2003); Freeman-Benson and Borning (2003); Schwartzman and Borning (2007) and Waddell et al. (2003).

3. A literature on methodological developments has used data relating to, or resulting from, UrbanSim to investigate improvements in two broad areas. The largest number of these articles have looked at discrete choice innovations relating to household location choice (de Palma et al., forthcoming; de Palma et al., 2007 and de Palma et al., 2005), and joint household location choice and mode or workplace choice (Waddell, Bhat, Eluru, Wang and Pendyala, 2007 and Pinjari et al., 2007). The other articles have looked at sensitivity analysis of variation in UrbanSim results (Pradhan and Kockelman, 2002) and methods to quantify the amount of uncertainty in UrbanSim results (Sevcikova et al., 2007).

4. A number of UrbanSim applications have been reported. Half of these have been written by the developers of UrbanSim. They show that UrbanSim can be used successfully (Waddell, 2002) and that the integration of land-use can have an important impact on transportation system performance results (Waddell, Wang and Charlton, 2007). Waddell (2002) shows that for the case of Eugene, Oregon, UrbanSim produced good results for predicting land-use evolution (e.g. where households and jobs will locate in the future). This is demonstrated using correlations of UrbanSim predictions against actual developmentin 1995. Waddell, Wang and Charlton (2007) provide a detailed description of an application for the region of Salt Lake City, Utah. Among other things, it shows that compared to a system analysis using a traditional transportation modeling approach that total vehicle miles traveled (VMT) are 5% higher and that total congestion delays are 16% higher. A more recent, although not final model (Waddell, Ulfarsson, Franklin and Lobb, 2007) of San Francisco shows how a recent version of UrbanSim has been used with an activity based model. This paper reports that it took 1 year to develop this model.

5. There have also been three independent reports of UrbanSim applications. Joshi et al. (2006) report on the application of UrbanSim to analyze the effects of a planned light rail system in Phoenix, concentrating primarily on land-use implications. Nothing is mentioned about the effort required for model implementation. The number of authors in the paper suggest the resources required were significant. Zhao and Chung (2006) is the most detailed independent analysis of an application of UrbanSim in Volusia County, Florida. They report success in implementing UrbanSim and that it is feasible, although there is not much detail about the resources required. They report that the main challenges were related to

data collection and preparation and parameter estimation. They also conclude that the expertise required to develop the model is outside of the scope of what some MPOs may have and therefore that they would require external consultant services, and that the user manual is insufficient.

Loechl et al. (2007) describe the results of modeling efforts for the region of Zürich. The model was not fully implemented (there was no interaction with a transport model) and the paper describes problems encountered in data collection and how these were overcome. It also reports on the simulation results obtained in this effort.

Another report has just been released by IAURIF in Paris that itemizes ten lessons learned after four years of modeling efforts for the Paris region (Nguyen-Luong (2008)). They report that their's is the first full implementation of UrbanSim outside of the United States and it provides practical lessons that they were able to derive from their efforts over the past four years. Among other things it reports that an interdisciplinary team of 4 people was required for four years! They also provide interesting insights into the factors that are required to develop a well-functioning, UrbanSim model.

To summarize, in the formal literature there is some evaluation of the use of UrbanSim. Loechl et al. (2007), Zhao and Chung (2006) are two independent sources that refer to the problems and challenges of using UrbanSim. Nguyen-Luong (2008) is not in the more formal literature but is a good analysis of the use of UrbanSim for a completed project. There is, however, little guidance in the literature about how to evaluate UrbanSim as an integrated model. The purpose of this paper is to describe how to develop an UrbanSim model for evaluation, without having to invest the resources required for a full-fledged model.

3   How UrbanSim Works

UrbanSim is evolving rapidly with new functionalities and advances in how it models urban environment. This description concentrates on how it has traditionally worked. UrbanSim is composed of a number of submodels that are run to predict the location of households, jobs and new real estate developments. The primary driver of UrbanSim is demographic and economic evolution. This is represented by exogenous data on households and jobs for each year of simulation. The evolution of households and jobs is modeled analogously. A simplified description of how UrbanSim models household evolution is sufficient to clarify both for a typical simulation year.

Demographic projections determine population change in the region. New households are put in a list of households to be placed later on in the simulation. At the same time, a certain proportion of households are assumed (and randomly selected) to move. They are also placed in the list of unplaced households. Households are then placed on gridcells by the household location choice model

(HLCM). New developments are created by the "development project transition model" that is influenced by the vacancy rate. The lower the vacancy rate, the more residential units will be built and vice-versa. The location for new developments is determined by the "development project location choice model." A land-price model updates gridcell land-values.

The location choice models (for people, jobs and real-estate developments) are discrete choice models that are estimated on the data for the region of interest[1]. The land-price model is a regression model estimated on data for the region of interest. It is in this way that UrbanSim is tailored to each application.

Geographically, the region being modeled is characterized by at least two definitions. Households, jobs and buildings are located in the primary division of gridcells (traditionally) or parcels (more recently). A secondary division is the traffic analysis zone (TAZ). Correspondence between gridcells and TAZs is how transportation system performance measures are assigned to gridcells. That is, all gridcells in a given TAZ use the same performance measures (travel times and logsums).

The backbone of UrbanSim is a relational database (usually in MySQL) that contains exogenous data, primary data, model coefficients and specifications, and data classifications. This is referred to as the baseyear database. The baseyear database is generally written to a baseyear cache from which UrbanSim is run. Exogenous data include overall model parameters (e.g. gridcell dimensions, units of measurement, etc.) and population and employment projections.

"Primary data" are represented by ("the six tables"): the gridcells, households, jobs, buildings, development event history and development constraints tables. The gridcells table is the central table that links all the other tables. It identifies and characterizes each gridcell in the urban system. The characteristics of a gridcell include:

- location relative to other gridcells

- political characteristics (zoning, county, city, etc.)

- TAZ correspondence

- geographical characteristics (distance to transportation infrastructure, gridcell slope, etc.)

- characterization of built form (e.g. number of residential units, surface area of office space, etc.)

Each observation of the households table represents one household. Households are characterized by socio-economic characteristics and the gridcell in which they are found. Each observation of the jobs table represents one job with jobs characterized by industrial sector, the type of building in which it is found (commercial, industrial, etc.) and gridcell. Each observation of the

---

[1]For more information on discrete choice models see e.g. Ben-Akiva and Lerman (1985) or Ben-Akiva and Bierlaire (2003)

buildings table identifies the building's location, what type of building it is and its composition (residential units, commercial surface area, etc.).

The development event history table contains information on historical developments (generally from the ten years preceding the baseyear). It characterizes new developments (residential units, surface area, etc.) and identifies the gridcell where the development took place. The development project transition model samples developments from this table to create new developments in simulation years.

The last of the six tables is the development constraints table. It 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. The development project location choice model uses this table to identify gridcells to which new developments can be placed.

The rest of the tables in the baseyear database contain information on the coefficients of the configurable models (e.g. the HLCM), various data classifications (e.g. building type 1 as residential), and other global model parameters.

The principle results of UrbanSim are the distribution of households, jobs and buildings across the study area. These data can be used to develop estimates of transportation system performance. As such, land-use and transportation system performance can be estimated for different demographic, economic, zoning and transportation planning scenarios at a fine level of detail.

This description has described the functioning of UrbanSim using regular gridcells and employment location choice models. More recent developments (flexible geographies and business location models) are now available, but have yet to become commonplace. Some description of these more recent capabilities is described below.

4 The Two Case Studies

This paper is based on the application of UrbanSim in two study regions. In both cases, UrbanSim models were developed with readily-available data and limited human resources. The purpose was to understand how difficult it is to develop an UrbanSim model that could be used to evaluate its use in a new region. The two case-studies differ considerably. In the case of Brussels, very limited data and no transport model were readily available. For Lausanne, relatively abundant and easy-to-use data were available. A well developed transportation model was also at our disposition.

4.1 Brussels, Belgium

Thanks to a partnership with Stratec, an engineering firm in Brussels, data used for the application of the integrated model (TRANUS) was available for the Greater Brussels Region. Brussels is the capital of Belgium. Data was available for an area of roughly 4,300 km^2 centered around the city of Brussels.

The study region included 139 townships in parts of Wallonia (French-speaking area to the south) as well as the Flemish Region (Flemish-speaking area to the north).

Most of data in the Brussels model came from that used in the TRANUS model. This included:

- Households (7 socio-economic classes) for 1991 and 2001 by zone;

- employment (13 sectors) for 1991 and 2001 by zone;

- land-value (3 land-uses) for 2001 by zone;

- interzonal travel times and logsums for 2001;

- zoning for the Greater Brussels Region.

GIS layers for highways and main arterials for Belgium and hypotheses for various parameters (e.g. vacancy rates) were also provided by our partner.

4.1.1 Data Preparation

Description of this model has been documented in various project and technical reports (Singh, 2008; Samartzis, 2007; Patterson and Bierlaire, 2007 and Stoitzev and Zemzemi, 2008). The model for Brussels used the Eugene-Springfield dataset provided with the UrbanSim distribution as a model. As such, the Brussels data was prepared so that it could be used by the same structure as the Eugene-Springfield model.

A standard 150 x 150 meter grid was used that contained roughly 193,000 gridcells for the region as a whole. Geographic characteristics of gridcells were assigned to the extent that data were available (zoning information, proximity to roads, etc.). 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.

The households table was created by disaggregating households to residential gridcells in their respective zones. Characteristics were assigned to individual households based on the characteristics of their socio-economic categories. The jobs table was created by disaggregating jobs to appropriately zoned gridcells (e.g. industrial jobs on industrial gridcells). All jobs of a given sector were assigned to the same type of building (industrial, commercial, etc.). To populate the building table, one building of each type required by the jobs and households present on the gridcell was created. For example, one residential building with enough units to house the households present was created per gridcell. The number of units and total surface area of non residential buildings was adjusted to account for vacancy rates. Other building characteristics (e.g. improvement value) were calculated as functions of the number of units and non residential surface area of the buildings.

Historical data on jobs and employment from 1991 were used to create the development event history table. First, zonal employment and population change

between 1991 and 2001 were determined. Then buildings from each affected zone were randomly selected as having been built in the ten years before the baseyear. Enough buildings were randomly selected to house the new population and jobs. Each building represented one development event. Development constraints were derived from the number of residential units and non-residential surface areas observed in the gridcells table by plan type of gridcell.

The majority of the work was done by a master's student in the context of a thesis in three and a half months. Some subsequent work incorporating better land-use data was done by undergraduate students and a postdoctoral supervisor in stops and starts over the following year (1.5 person-months). Understanding the basic data requirements of the model and preparation of the available data to meet these requirements was done within two months. In order to test that the data used respected model requirements, the first simulations were done then. These simulations used the Brussels data, but the models from the Eugene example. The following month and a half was spent estimating the location choice and land-price models using the Brussels data, fine-tuning the data and models, and running simulations. An additional 1.5 person-months was required to produce the results presented. Additional data and a usable transportation model could not be obtained without a significant investment of resources so work on the model was finalized.

4.1.2 Results

Unsurprisingly, given the use of aggregate data, results from the Brussels model are not awe-inspiring. Given coarse jobs and households data and no building data, it was difficult to estimate robust models. For the most part, the models had a limited number of variables, especially if compared to fully operational models (e.g. Waddell, Wang and Charlton, 2007). Despite this, models were generally pleasantly surprising, with the most important variables (e.g. land price, accessibility measures, etc.) normally coming out significant with the right sign. This was not always the case. The most problematic models were the real estate development models that had few observations - the industrial development project location choice model, for example had only 26 observations. An example of a typical model is the household location choice model shown in Table 1.

The model contains six variables all together. Households prefer locations that are less expensive, all else equal (Variable 1). They also prefer to live near households with similar incomes (Variables 1 and 4), although high income families show some affinity to being near to low income households (Variable 3). In geographical terms, households prefer being closer to the central business district (CBD) (Variable 5) and locations in the Central Brussels Region or Wallonia (Variable 6).

Simulation results compare surprisingly well with actual population growth by city in the Brussels region. Figure 1 shows a map of the difference between actual and simulated population growth rates between 2001 and 2007. In fact, for more than half of the cities, the difference in simulated population growth to

	Variable	Coefficient	Std. Error	t-value
1	Cost:Income	-0.0661	0.0307	-2.2
2	% High Inc. If High Inc.	0.0334	0.00150	22.3
3	% Low Inc. If High Inc.	0.00400	0.00138	2.9
4	% Low Inc. If Low Inc.	0.0603	0.00109	55.4
5	Travel Time to CBD	-0.000622	0.000148	-4.2
6	In Flanders	-0.0267	0.00856	-3.1
	Null Log-likelihood is:	-440982.247		
	Log-likelihood is:	-439242.311		
	LR Test:	3479.871		
	Number of observations:	129655		
	Convergence statistic is:	7.617E-05		

Table 1: Brussels Household Location Choice Model

actual growth was between 2% and -2%. All (except 1) were between ±10%. A pattern emerges in this map where population growth in cities along a northeast axis are under-predicted. It appears that this is the result of the household location choice model. It seems to over-emphasize the importance of land-price and under-emphasize the importance of travel time to the central business district in household location choice.

4.2 Lausanne, Switzerland

Lausanne is the capital of the Canton of Vaud. It is located in the middle of the north shore of Lake Geneva. The study region (Lausanne-Morges) covers an area of approximately 200 km^2 that includes 45 communes. It was home to roughly 277,000 people and 162,000 jobs in the baseyear of 2000. For documentation on the Lausanne model refer to Patterson and Hurtubia (2008) and Bettex (2008).

Compared to the case of Brussels, the region of Lausanne had abundant and readily-available data. Swiss censuses of households (2000) and businesses (2001) provided data by hectare for all of Switzerland. Excellent data (mostly at the 1:25000 scale) were available for zoning and other geographic characteristics. Finally, a transportation model (EMME) for the region was developed at the EPFL. Some important data were not easily available, that is: data for land prices, improvement values or surface area required by job. Moreover, the Swiss federal census does not ask any questions about revenue.

4.2.1 Data Preparation

The Eugene-Springfield dataset was used to provide the base structure for the Lausanne model. A hectare (100m x 100m) gridcell system corresponding to that used for Swiss censuses was used. There were roughly 21,000 of these grid-cells in the study region. Geographic characteristics of gridcells were assigned to the extent that data were available (zoning information, proximity to roads,
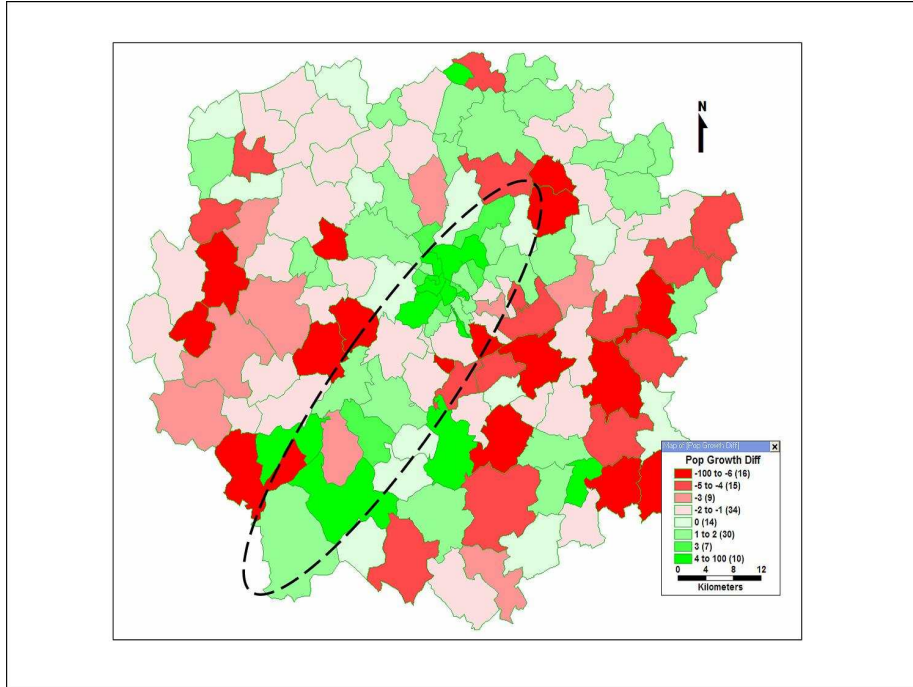
Figure 1: Difference between real and simulated population growth rates in the Brussels region (2001-2007)

etc.). Data on built form (residential units, estimated surface area by building type, etc.) were included after the creation of the buildings, households and jobs tables. As a proxy for land prices, gridcell population and job density were used for residential and non-residential land values.

Households data (except income) came directly from the census. Estimates of income for the different job types attributed to households was used as the indicator of revenue.

For the jobs table, information from the Federal enterprise census was used. Enterprises were characterized by their hectare location, industry type and number of jobs. Preparation of the jobs table required creating a record for each job for each enterprise.

The buildings table used data from the population census and jobs table. As part of the population census, information is available on all buildings with residential units. This includes the number of residential units and period of construction, but does not include information on improvement value. This was used for residential buildings, with residential improvement value being estimated as a function of the number of residential units. No information was available for non-residential buildings. Non-residential buildings were created

9

with enough surface area to house the number of jobs (accounting for vacancy rates) requiring different building types. Non-residential improvement values were a function of the building surface area.

The development event history table used data from the population and enterprise censuses. Residential development events could be directly extracted from the residential building data of the population census. Improvement values were defined as a function of residential units. For non-residential buildings the 1995 enterprise census was used to calculate the change in jobs by gridcell. A development event was created on those gridcells where there was an increase of more than 4 jobs for a given building type (i.e. industrial and commercial). Surface area was a function of the number of jobs. Improvement value was a function of surface area. Development constraints were derived from the number of residential units and non-residential surface areas observed in the gridcells table by gridcell plan type.

The vast majority of work on this model was done by one postdoctoral fellow. Initial data preparation was done in Python and with TransCAD. Part of the effort necessary was devoted to familiarization with Python. Data preparation and incorporation lasted roughly two months, after which the first simulations were run with the original models from the Eugene example. Lausanne-specific models were estimated with UrbanSim and simulations integrated with the transportation model were begun two weeks later. Preparation of data from UrbanSim for the transportation model (EMME) (and *vice-versa*) was automated, with files being transferred between different computers where the two models were housed.

4.2.2 Results

The results from the Lausanne model are more encouraging than the Brussels model. The models estimated had more significant variables than for Brussels model. Since building data were lacking, many important variables such as surface area by industrial sector and improvement values could not be used. As an example, Table 2 shows the household location choice model for the prototype Lausanne model.

The odds of choosing a location are decreased if it is expensive (Variable 1), but increased if there is retail employment nearby (Variable 2). People prefer to live near people of similar incomes (Variables 3 and 4). Young households prefer to be in high density, mixed-use locations (Variables 5 and 6), whereas households with children prefer lower densities (Variable 7). Households prefer locations that have good accessibility to other people (Variable 8) and to be closer to the central business district (Variable 9). At the same time, the odds of choosing a location decrease the closer it is to the train station (Variable 9).

While most models were more robust than those for Brussels, initial simulations did not perform as well in terms of population growth. Actual population growth by city was compared with simulated growth between the years 2000 and 2007. The results are shown in Figure 2. The difference between observed and simulated growth is much larger than in the case of Brussels. Part of this

	Variable	Coefficient	Std. Error	t-value
1	Cost:Income	-5.935	0.747	-8.0
2	Retail Employment WWD	0.0298	0.00328	9.1
3	% High Inc. If High Inc.	0.0298	0.000616	48.4
4	% Low Inc. If Low Inc.	0.0236	0.00113	21.0
5	High Density if Young	0.428	0.0177	24.1
6	Mixed Use if Young	0.454	0.0217	21.0
7	Res. Units with Children	-0.00472	0.000103	-45.6
8	Accessibility to Population	0.400	0.0455	8.8
9	Travel Time to CBD	-0.0211	0.00259	-8.1
10	Travel Time to Station	0.0320	0.00210	15.2
	Log-likelihood is:	-440830.606		
	Null Log-likelihood is:	-444383.444		
	LR Test:	7105.676		
	Number of observations:	130655		
	Convergence statistic is:	5.398E-04		

Table 2: Household Location Choice Model

is due to the small size of cities in the area - one quarter had less than 1,000 people. The model also predicts exaggerated densification in areas that are already quite dense (mostly communes at the center of the region). This has to do with the variables of the household location choice model, and that development constraints were not constraining enough. The HLCM places people in dense locations closer to the center. The same is true for the residential development project location model. Together, the fact that development constraints were not restrictive enough appears to explain the simulated exaggerated densification.

5 Developing a Prototype UrbanSim Model

This section describes the procedure used to develop both the Brussels and Lausanne UrbanSim models. It is a procedure we have found can be followed to develop a model for evaluation in 3 to 5 person-months. Figure 3 illustrates the procedure we refer to as Iterative Improvement. The procedure can be divided into three phases: familiarization, implementation and evaluation.

5.1 Familiarization

This step has two components: Familiarization with UrbanSim and its data requirements, and familiarization with local data and how it can be used in UrbanSim. The stage of familiarization should take between two weeks and a month.
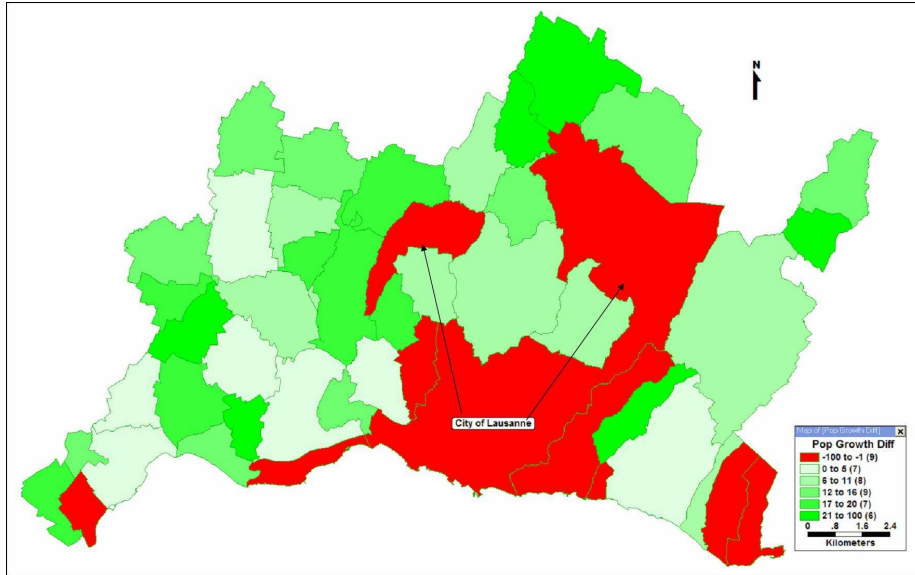
Figure 2: Difference between real and simulated population growth rates in Lausanne region (00-07)

The most efficient way to familiarize oneself with UrbanSim and its data requirements is to *learn by doing*. UrbanSim is not particularly well documented. The documentation that does exist (e.g. the user's guide) is technical and is written more for computer programmers than general end-users. A general understanding of UrbanSim from documentation is possible but not a good understanding of the difficulty of using UrbanSim or an intimate knowledge of how it works or what it can do. The best way to do this is to begin with the Eugene tutorial that can be downloaded with UrbanSim.

With the Eugene tutorial it is possible to run simulations and understand the kind of results that UrbanSim can produce. From this basic tutorial, it is then possible to understand UrbanSim data requirements by exploring the baseyear database. This is most easily done by exporting the baseyear cache to a more user-friendly format (e.g. MySQL). It is only by perusing the baseyear data and its component tables that it is possible to understand the connection between the many related tables.

Once familiar with UrbanSim and its data requirements, the developer needs to think about the fit between UrbanSim's data requirements and locally available data. In particular, the developer needs to have a sense of what data are available and what would need to be done to them so that they could fit into the structure required by UrbanSim. For data that are not available, the developer needs to think about how simulated or proxy data could be used. In the case of Brussels, no data on buildings were available. It was reasoned, however, that
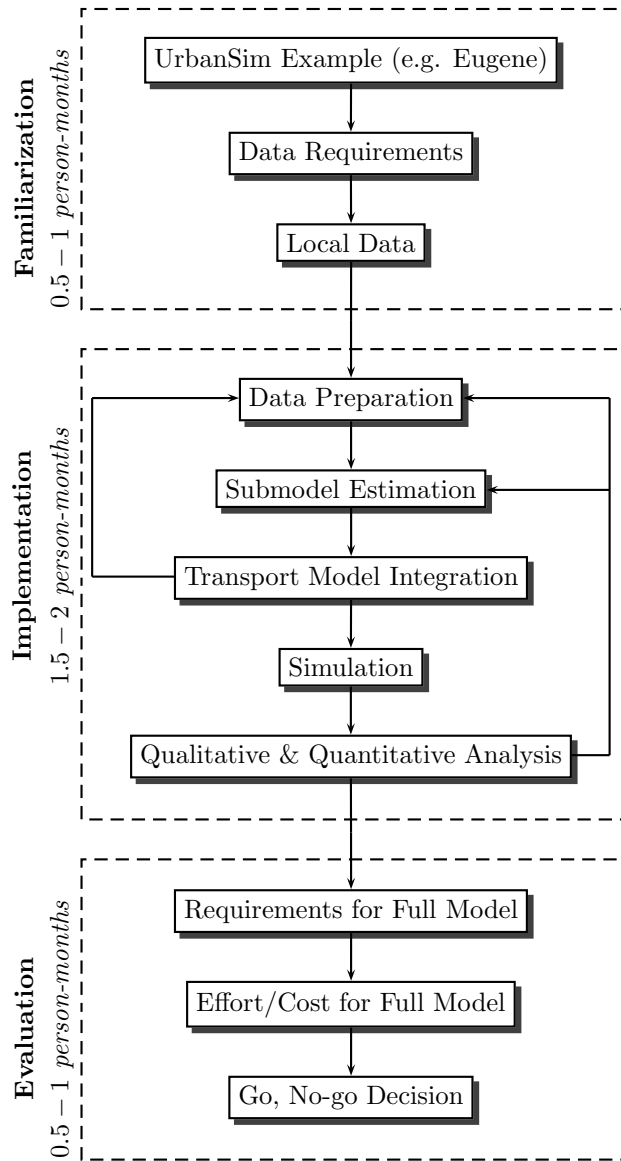
12

Figure 3: The iterative improvement procedure to develop prototype UrbanSim models

buildings could be "created" to house the jobs and households for which we did have data. This was indeed the approach taken and it allowed the development of a prototype model by using data we had at our disposition.

5.2 Implementation

Implementation is the most involved step. In this stage tables are prepared to be put in the database, the various location choice and land price models are estimated and simulations are performed and analyzed. This stage should require between 1.5 and 2 person-months of work.

The most efficient way to develop a prototype is to *develop a model for which there is an example.* UrbanSim has been evolving to provide greater model flexibility in how it models employment location and the geography at which it operates. It is now possible to model the location of businesses as opposed to jobs as has traditionally been the case. Geographically, it is now possible to run UrbanSim at the parcel (or even TAZ) level as opposed to just the gridcell level (Waddell, Wang and Charlton, 2007).

These additions increase the realism of the model itself, but they have been made available before the release of examples implementing them. There is no example provided of a model that uses non-regular geographies or the business location model. Examination of the underlying code can be done to understand these features although it is not recommended to try to develop a model without an example. In both the Brussels and Lausanne models, the Eugene, gridcell-based example was used as the model. Its data tables were used as the structure in which to place the data for the two new study regions. If one wanted to develop models for which there is not an example, the best approach would be to: a) examine the underlying code to understand how these new tools work; and b) use simplified artificial data tables to try to make the tools work.

5.2.1 Data Preparation

At the stage of data preparation there are two things to keep in mind. The first is to *make do with what you've got.* If the goal is to put together a model to understand and evaluate UrbanSim, it is important to realize that not *all* the data represented in the example databases is required. Once an initial model is up and running, the importance of the different types of data can be evaluated. In order to save time in implementing a preliminary model, effort should be placed on preparing *readily-available* data. Obtaining all the data required before implementing the model can take a lot of time. In fact, preparation of a complete dataset takes around two years on average. For data that is not readily available, one should not be afraid to use simulated data or make simplifying assumptions. This is *not* the case for a fully operational model. The recent report Nguyen-Luong (2008) emphasizes the importance of using real data in an operational model.

As an example, in the case of Lausanne, there was no readily-available land-price data so gridcell population and employment density were used as proxies.
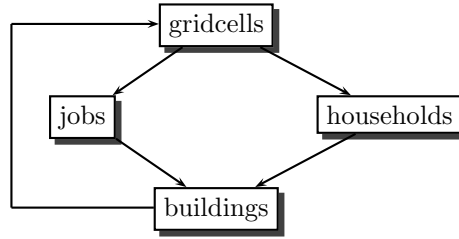
Figure 4: Iterative construction of gridcells table

This seems to have worked quite well with "land-price" consistently being statistically significant with the "correct" sign in all of the models to be estimated. While actual land-price, building, disaggregated population and employment, etc. data would have been ideal, obtaining these data would have greatly lengthened the time required to develop a preliminary model.

The second thing to keep in mind during data preparation is to *concentrate on the "six tables"*: the gridcells, households, jobs, buildings, development_event_history and development_constraints tables. These are also the tables which require the most data and the most preparation. The other tables are:

- Derivatives of these tables (e.g. the model specification tables);

- require relatively little additional data (e.g. the urbansim_constants table);

- relatively easily obtained from the travel model (the travel_data and zones tables);

- can be approximated using simplifying assumptions and data from the "six" tables (e.g. annual control totals).

So, most of the effort applied to preparing data for an UrbanSim model should go towards producing the highest quality data for these tables as possible.

In both the Brussels and Lausanne models, the first step was determination of the gridcell system and construction of gridcells table. Political characteristics of the gridcells (i.e. zoning) are particularly important for accurately predicting future development. The intimate relationship between the gridcells and development_constraints tables means accurate determination of the zoning characteristics (i.e. plan_type_id) and the zoning constraints (e.g. maximum residential units) by zoning type is particularly important. Other geographical characteristics (e.g. distance to highway, etc.) are less critical but should be included to the extent that the data are easily available. The whole procedure, however, need not be held up because one of these other, less critical, variables is missing.

Land-use variables (e.g. industrial surface area) make up the rest of the variables in the gridcells table. These data are closely linked to the buildings

table that includes information on the surface area of the different types of buildings. As such, the gridcells table needs to be constructed iteratively with the buildings table. That is, the gridcells need to be created and the buildings attributed to gridcells. The buildings data can then be used to determine the land-use variables.

In both models, some building data needed to be simulated. For Brussels, all buildings data needed to be simulated and for Lausanne, non-residential data needed to be simulated. This was accomplished by using simple rules relating residential units and non-residential surface area to jobs and households in a given gridcell. Since buildings were a function of households and jobs, these two tables were the next most important after gridcells. Thus, efforts were concentrated on the jobs and households tables so that they could be used to create the buildings table (see Figure 4).

The development_event_history is extremely important. It is a record of historical developments from which UrbanSim samples to create future developments. In the case of Lausanne, residential data for this table were readily available. For non-residential (and residential for Brussels) development events needed to be inferred from population and employment changes.

5.2.2   Submodel Estimation and Transport Model Integration

Once baseyear data has been prepared, the location choice and land-price models must be tailored to the new region. Estimating these models with UrbanSim using the baseyear database is done relatively easily. Properly developing models requires some analysis on things such as the distribution of jobs and households in the region. Much of this analysis can be done directly in UrbanSim which uses Matplotlib to produce maps for most variables of interest. The quality of the models is difficult to evaluate without seeing simulation results so it is good estimate these models relatively quickly knowing that they can be improved after analysis of simulation results.

With the land-use part of UrbanSim ready to be implemented, interaction with the transport model needs to be considered. It is commonly assumed that UrbanSim cannot function without continual interaction with a transport model. In fact *continual interaction with a transport model is not necessary* and it runs by default without continual interaction. At the same time, transport system performance data is fundamental to the workings of UrbanSim (e.g. the travel_data and zones tables). As such, it is difficult to run UrbanSim without a system of TAZs and basic performance measures (e.g. interzonal travel times) for the baseyear.

For Brussels and Lausanne the UrbanSim models were initially developed using baseyear transport performance measures. In the case of Brussels, the model was not integrated on an ongoing basis with a travel model. For Lausanne, after the sub-models had been developed using static transport system performance measures, UrbanSim was then integrated with the travel model after every five years.

Although continual interaction is not necessary, it is important to *know*

your transport model well when developing an UrbanSim application. There are two aspects to this. First, it is important to know what input data the transport model requires. This is useful to plan how you will be able to use UrbanSim results to produce inputs for the transport model. For the case of Lausanne, the transport model required zonal population and employment. Initial zonal population and employment figures were inconsistent with those of the transport model. It was, therefore, necessary to analyze the gridcell system (total hectares included and zonal attribution). The gridcell system was then changed to correspond that of the transport model. Basing the gridcell system on that of the transport model initially would have saved time.

Second, knowledge of the transport model can help prioritize the effort put in to data preparation. Traditionally, UrbanSim accessibility measures have been based on logsums from a mode choice model differentiated by household vehicle ownership level. E.g. one logsum measure for households with 0, 1 or 2 cars. Lausanne census data did not include information about automobile ownership. At the same time, the Lausanne transport model has an aggregate, zonal mode choice model and therefore does not include household level information. While it would have been possible (and indeed it is planned) to estimate vehicle ownership levels per household, this information was not required for the transport model. As such, in the prototype Lausanne model, household ownership level was not included in the households table without causing problems for the transport model. This saved time in the prototype implementation.

Once data for the UrbanSim model have been prepared and although likely incomplete or lacking, one can and *should not be afraid to run simulations with incomplete data*. When developing an UrbanSim model, there is a tendency to be reluctant to run simulations until the *best* data are ready and incorporated in the model. The problem with this is that obtaining this data can take a long time (see Section 5.2.1). Another problem is that running simulations is critical to understanding how UrbanSim works, what results it produces and how to interpret them.

In both applications described here, most effort was concentrated on the "six tables." Once these tables were completed, the other critical but less involved tables were assembled and simulations were run using the same model parameters as in the Eugene example. In both cases, these first simulations were run after 2-3 months. This was useful to ensure that the model was working from a computational standpoint. Once these initial simulations were run, the various sub-models were estimated to tailor them to the application region.

It is when simulations begin that issues surrounding UrbanSim version emerge. The development version of UrbanSim (with continual modifications) is freely available. From time to time, stable releases are made available. The development version can provide access to features not available in the stable releases, but it can also have unresolved bugs. The result is that time can be spent trying to ascertain whether problems in UrbanSim are because of user error or a bug in the development release. Using a stable release generally avoids such problems. When beginning with the Lausanne model, we used the latest development version. While it seemed promising, there were sufficient glitches in using it that

we reverted to the latest stable release and were able to concentrate on model development.

While using a stable release facilitates the job of model development, it is also a good idea to use the *latest* stable release. Changes made between versions can be such that the same dataset may not work perfectly well between versions. While some previous stable releases are made available, eventually they are removed. This can cause problems when trying to debug when upgrading between releases. We encountered such a problem with the Brussels model. It was originally developed using version 4.0. Work continued with version 4.0 followed by a pause in development. With the start of the Lausanne model, we upgraded all versions to 4.1.2 but there were some problems using the Brussels data. It would have been ideal to test the data with version 4.0 to ensure that there were no problems with the data, but this was not possible because it was no longer available. Time could have been saved if new versions had been used as they were released.

5.2.3   Simulation and Analysis

Once simulations have been run, it is time to analyze the results. A qualitative analysis can be done first. The modeler should ask if the model *seems* to be working sensibly. Is development happening in places one might expect? Next, to the extent possible, the model should be tested against actual data. In both the Lausanne and Brussels models, actual population growth was compared to simulated population growth. This is extremely important since it provides the first hints at diagnosing model weaknesses.

There are a few potential sources of model problems that can be identified. The first is with data. There are two aspects of this: the lack of particularly critical data, and inadequate use of available data. Based on the case of Brussels more generally, it can be concluded that the use of aggregate data poses many problems for the development of a robust model. The lack of historical development data made the estimation of real estate development models difficult with unconvincing results. In the case of Lausanne, one of the main factors driving results was insufficiently binding development constraints. The result was exaggerated densification in already dense parts of the region.

The second source of problems is with the submodels. In the initial development of the Brussels model, a very clear (and too strong) pattern of location of population in the outskirts of the region resulted from simulations. These results were driven primarily from problems with the household location choice and residential development models that tended to prefer locations far from the region's center. Along with data improvements, these models were re-estimated to try to develop models that led to a better recreation of actual trends. From a methodological perspective, UrbanSim uses a traditional logit model for its discrete choice models. Use of more sophisticated models that account for spatial autocorrelation could also improve estimation results. Estimation of such models would however involve the use of other software (e.g. BIOGEME Bierlaire, 2003, Bierlaire, 2008).

5.3 Evaluation

Once a prototype model has been developed it can be used for evaluation purposes. A prototype model should not be used for planning purposes for obvious reasons. After the experience of implementation, the developer will be in a very good position to consider three factors. The first factor to consider is what a complete UrbanSim model would look like - particularly, what data would need to be incorporated, obtained or adapted and how could sub-models be made more precise and/or further improved?

The second factor to consider is the effort and/or cost that would be required to make the improvements deemed necessary for a complete model. In the case of Brussels, gathering disaggregate data for the entire region (especially from Lausanne) would be a tremendous challenge. This is not to mention the fact that it would also require a great deal of work to adapt existing transport models for use with UrbanSim. For Lausanne, overcoming model weaknesses seems decidedly easier. In the case of development constraints, it requires a better analysis of existing data. Obtaining better land price data and information on buildings, surface areas, etc. would be demanding, but the data is obtainable.

The third factor is the identification of priorities. The first consideration at this point is what the model will be used for. If spatially disaggregated projections are desired (or required) this would go in favor of UrbanSim. If only aggregate, coarse land-use projections are required, other models (e.g. TRANUS) might be considered that do not have the same data requirements as UrbanSim.

Finally, the modelers are confronted with the "go, no-go" decision. In the case of Brussels, the effort required to improve the model further are large. As such, it is unlikely that at this stage further development will continue on that model. In the case of Lausanne, however, results are promising enough and the efforts for model improvement more restrained - development of a full-fledged model is more likely. One thing is for certain though. It is not possible to make an informed decision about developing an UrbanSim model without having developed a prototype.

6 Conclusion

The purpose of this paper was to describe a procedure to develop prototype UrbanSim models and describe how to use these models to evaluate UrbanSim for planning or research purposes. Our conclusions come in three parts. First, the best way to evaluate UrbanSim is to develop a prototype model. This is the best way to understand how the model works, what is required to run it, what it can do and to estimate the effort required to develop a full-fledged model. Second, developing a prototype model is achievable within three to five months of one person's effort. This is a reasonable investment when compared to the costs reported for development of a full-fledged model. As well, if a full-fledged model is developed, this is not a sunk cost and will reduce overall development effort, costs and time. Finally, there are a number of things to keep in mind

if the goal is to develop a prototype model for analysis. These are to: Learn by doing, develop a model for which there is an example, make do with what you've got, concentrate on the "six tables," realize continual interaction with a transport model is not necessary, know your transport model well, not be afraid to run simulations with incomplete data and to use the latest stable release.

7 Acknowledgements

References

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice, *in* R. Hall (ed.), *Handbook of Transportation Science, 2nd edition*, Operations Research and Management Science, Kluwer, pp. 7–38. ISBN:1-4020-7246-5.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis*, MIT Press, Cambridge, MA.

Bettex, L. (2008). Analyse géographique pour l'implémentation d'un prototype de modèle intégré de transport et occupation du sol pour la région lausannoise, *Technical report*, EPFL - Transport and Mobility Laboratory. Copy available from Marianne Ruegg, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland. www.strc.ch.

Bierlaire, M. (2008). Estimation of discrete choice models with BIOGEME 1.7, *Technical report*, EPFL - Transport and Mobility Laboratory. Available at: `http://transp-or2.epfl.ch/biogeme/doc/tutorial.pdf`.

Davis, J., Lin, P., Borning, A., Friedman, B., Kahn Jr., P. H. and Waddell, P. (2006). Simulations for urban planning: Designing for human values, *Computer* pp. 66–72.

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

de Palma, A., Motamedi, K., Picard, N. and Waddell, P. (forthcoming). Accessibility and environmental quality: Inequality in the Paris housing market, *European Transport* .

de Palma, A., Picard, N. and Waddell, P. (2007). Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater Paris region, *Journal of Urban Economics* **62**: 204–230.

Freeman-Benson, B. and Borning, A. (2003). Yp and urban simulation: Applying an agile programming methodology in a politically tempestuous domain. Proceedings of the 2003 Agile Development Conference, Salt Lake City, Utah, available at: `http://doi.ieeecomputersociety.org/10.1109/ADC.2003.1231447`.

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

Joshi, H., Guhathakurta, S., Konjevod, G., Crittenden, J. and Li, K. (2006). Simulating the effect of light rail on urban growth in phoenix: An application of the urbansim modeling environment, *Journal of Urban Technology* **13**(2): 91–111.

Loechl, M., Buergle, M. and Axhausen, K. (2007). Implementierung des integrierten flaechennutzungsmodells urbansim fuer den grossraum zuerich, *DISP* **168**(1): 13–25.

Nguyen-Luong, D. (2008). An integrated land use-transport model for the Paris region (SIMAURIF): Ten lessons learned after four years of development, *Technical report*, Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France (IAURIF). Available at: `http://www.urbansim.org/pipermail/users/2008-April/att-0001/` `-article_SIMAURIF_10_lessons.pdf`.

Noth, M., Borning, A. and Waddell, P. (2003). An extensible, modular architecture for simulating urban development, transportation, and environmental impacts, *Computers, Environment and Urban Systems* **27**: 181–203.

Patterson, Z. and Bierlaire, M. (2007). An UrbanSim model of Brussels within a short timeline, *Proceedings of the 7th Swiss Transport Research Conference*, Monte Verità, Switzerland. Available at: `www.strc.ch`.

Patterson, Z. and Hurtubia, R. (2008). Development of a prototype UrbanSim model for the Lausanne-Morges region of Switzerland, *Technical report*, EPFL - Transport and Mobility Laboratory. Copy available from Marianne Ruegg, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland.

Pinjari, A., Pendyala, R. M., Bhat, C. and Waddell, P. (2007). Modeling residential sorting effects to understand the impact of the built environment on commute mode choice. CD of the 87th Annual Meeting of the Transportation Research Board.

Pradhan, A. and Kockelman, K. (2002). Uncertainty propagation in an integrated land use-transportation modeling framework - output variation via urbansim, *Transportation Research Record* **1805**: 128–135.

Samartzis, L. (2007). *Modélisation de Bruxelles avec UrbanSim*, Master's thesis, EPFL - Transport and Mobility Laboratory. Available at: `http://transp-or2.epfl.ch/cours/projets.php?details=1`.

Schwartzman, Y. and Borning, A. (2007). The indicator browser: A web-based interface for visualizing urbansim simulation results. Proceedings of the 40th Hawaii International Conference on System Sciences. Available at: `http://www.urbansim.org/papers/schwartzman-hicss-2007.pdf`.

Sevcikova, H., Raferty, A. and Waddell, P. (2007). Assessing uncertainty in urban simulations using bayesian melding, *Transportation Research B* **41**: 652–669.

Singh, Y. K. (2008). Modeling of Brussels with UrbanSim, *Technical report*, EPFL - Transport and Mobility Laboratory. Copy available from Marianne Ruegg, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland.

Stoitzev, I. and Zemzemi, F. (2008). La calibration d'UrbanSim pour la ville de Bruxelles, *Technical report*, EPFL - Transport and Mobility Laboratory. Copy available from Marianne Ruegg, Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland.

Waddell, P. (1998). The Oregon prototype metropolitan land use model. Paper presented at the ACSE Conference, Portland, Oregon. Available at: `http://www.urbansim.org/papers/ASCE_Model.pdf`.

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: Planning and Design* **27**: 247–263.

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

Waddell, P. (2002). Urbansim: Modeling urban development for land use, transportation and environmental planning, *Journal of the American Planning Association* **68**(3): 297–314.

Waddell, P., Bhat, C., Eluru, N., Wang, L. and Pendyala, R. (2007). Modeling the interdependence in household residence and workplace choices. CD of the 87th Annual Meeting of the Transportation Research Board.

Waddell, P. and Borning, A. (2004). A case study in digital government - developing and applying urbansim, a system for simulating urban land use, transportation and environmental impacts, *Social Science Computer Review* **22**(1): 37–51.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M. and Ulfarsson, G. (2003). Microsimulation of urban development and location choices: Design and implementation of urbansim, *Networks and Spatial Economics* **3**(1): 43–67.

Waddell, P., Ulfarsson, G., Franklin, J. and Lobb, J. (2007). Incorporating land use in metropolitan transportation planning, *Transportation Research A* **41**: 382–410.

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

Waddell, P., Wang, L. and Liu, X. (forthcoming). Urbansim: An evolving planning support system for evolving communities, *Planning Support Systems* .

Zhao, F. and Chung, S. (2006). A study of alternative land use forecasting models - final report, *Technical Report BD015-10*, Florida Department of Transportation, Tallahassee, Florida. Available at: `http://www.dot.state.fl.us/research-Center/Completed_Proj/` `Summary_PL/FDOT_BD015_10_rpt.pdf`.