Econometric guidance for developing UrbanSim models.
First lessons from the SustainCity project.

Nathalie Picard[1] *and Constantinos Antoniou*[2]

[1]*Université de Cergy-Pontoise and INED, Paris, France (email: Nathalie.Picard@u-cergy.fr)*

[2]*National Technical University of Athens, Greece (email: antoniou@central.ntua.gr)*

Abstract

In the context of the SustainCity project (http://www.sustaincity.eu), three European cities (Brussels, Paris and Zurich) will be modelled using the land use microsimulation platform UrbanSim. This platform relies on various models interacting with each other, to predict long-term urban development. The aim of this paper is to provide some econometric insight into this process.

A common set of notation and assumptions are first defined, and the more common model structures (linear regression, multinomial logit, nested logit, mixed MNL and latent variable models) are described in a consistent way.

Special treatments and approaches that are required due to the specific nature of the data in this type of applications (i.e. involving very large number of alternatives, and often exhibiting endogeneity, correlation, and (pseudo-)panel data properties) are discussed. For example, importance sampling, spatial econometrics, Geographically Weighted Regression (GWR) and endogeneity issues are covered.

Specific options of the following models: (i) household location choice model, (ii) jobs location/firmography, (iii) real estate price model, and (iv) land development model, are demonstrated in the context of the on-going case studies in Brussels, Paris and Zurich. Finally, lessons learnt in relation to the econometric models from these on-going case studies are summarized.

1 Introduction

Typically, urban development models have been based on aggregate principles. UrbanSim is among a new breed of models that use microsimulation (Waddell et al., 2003) in an effort to overcome the limitations of earlier models and provide a more dynamic and detailed paradigm. The advantages and disadvantages of using microsimulation are not within the scope of this paper, but the main implication is that more and more detailed data are required.

In the context of the SustainCity project (http://www.sustaincity.eu), three European cities (Brussels, Paris and Zurich) will be modelled using the land use microsimulation platform UrbanSim. This platform relies on various models interacting with each other, to predict long-term urban development. The aim of this paper is to provide some econometric insight into this process.

A common set of notation and assumptions are first defined, and the more common model structures (linear regression, multinomial logit, nested logit, mixed MNL and latent variable models) are described in a consistent way.

Special treatments and approaches that are required due to the specific nature of the data in this type of applications (i.e. involving very large number of alternatives, and often exhibiting endogeneity, correlation, and (pseudo-)panel data properties) are discussed. For example, importance sampling, spatial econometrics, Geographically Weighted Regression (GWR) and endogeneity issues are covered.

Specific options of the following models: (i) household location choice model, (ii) jobs location/firmography, (iii) real estate price model, and (iv) land development model, are demonstrated in the context oft he on-going case studies in Brussels, Paris and Zurich. Finally, lessons learnt in relation to the econometric models from these on-going case studies are summarized.

1.1 Data availability and limitations

The following sections provide an overview of the available data for the three case studies considered within the SustainCity project (http://sustaincity.org): Brussels, Paris and Zurich. UrbanSim has very large data requirements, making data collection a long and complicated effort. Data collected from various sources need to be processed, matched and homogenized, before they can be used. Besides these practical issues, however, there are further challenges to be dealt with. For example, some of the collected data imply further restrictions (e.g. those related to data protection) or are not public and therefore their use is limited. Finally, there are also privacy issues that can limit the usability of data, at least in their more disaggregate forms, forcing again for aggregation (resulting in loss of data) or other forms of anonymization. Such restrictions are particularly stringent for Brussels and Paris case studies, which had important consequences on data and econometric methods used.

1.2 Objectives and policy implications

The policy objectives of the SustainCity project can be summarized in the following three points:

a) Define objectives (sustainability and others) of policy makers: what are the components (economic, environmental, social, etc.), what is the horizon (5 years or 50 years), valuation of each component (monetary and/or categorical) as well as the level of aggregation;

b) Translate the model outputs into objectives for policy makers: this includes developing output reports for the model and suggesting feedbacks of some elements for the model development (local environmental quality has a clear feedback on housing demand and prices);

c) Define alternative sustainability policy packages, translate them into model inputs and discuss expected outcomes.

The policy objectives need to be defined by type, level of aggregation and level of quantification. The impacts of standard policies have been studied in various projects. For example, much attention has been devoted to the impact of road pricing. Road pricing has a positive aggregate impact, but implementation costs are not trivial, and acceptability is an issue since some agents gain, while others loose. The transfers needed to improve acceptability and preserve equity are well understood, but the land use impacts and the implementation are still not clear. Parking policies, traffic restrain, pedestrian areas in city centres, lanes restricted to bicycles, provide other types of "soft" policies, whose short run and long run impacts are likely to be non-negligible.

2 Suitable techniques

2.1 Notations and assumptions

The majority of the models that will be estimated fall under two general categories: Linear regression models and discrete choice models. The objective of this subsection is to provide a basic set of notations and assumptions, in order to ensure that the model development work will be presented in a consistent manner.

The linear regression model is given by $Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$ where the error terms ε_i are assumed to be white noise (normally distributed with zero mean and variance σ^2). Linear regression can be applied to quantify the dependent variable *Y* as a linear combination of a number of variables (or regressors) X. The regression coefficients β denotes the increase

to the dependent variable per unit change in the corresponding explanatory variable (or regressor) *X*.

In the discrete choice framework the entity of reference is the individual decision-maker, described by a number of socio-economic characteristics, e.g. age, gender and income. These decision makers choose among a set of available (discrete) alternatives. The identification of the choice set among all available alternatives is one important aspect, which becomes particularly relevant when a huge number of possible choices may be available. A decision-maker *n* selects one and only one alternative from a choice set $C_n = \{ 1, 2, ..., i, ..., J_n \}$ with J_n alternatives.

The specification of a random utility model uses the following utility specification (for a decision maker *n* choosing alternative *j* from a choice set of *J* alternatives) $U_{jn} = X_{jn}\beta + \nu_{jn}$ where X_{jn} are observable variables that relate to the alternative *j* and decision maker *n*, β is a vector of coefficients of these variables, and n_{jn} is a zero-mean, random term that is iid extreme value. Several assumptions can be made about the distribution and the variance/covariance structure of the error term. The most common assumptions lead to the logit model (i.i.d. Gumbel error terms) and probit model (Normal error terms).

2.2 Model structure

The main types of models that are being considered in this project are outlined in this section, starting from the more straightforward and moving to the more advanced.

2.2.1 Linear regression

Simple linear regression

The linear regression model is an attractive and simple method that is being used extensively. While the linear regression model is simple (to run and interpret), elegant and efficient, it is subject to the fairly stringent Gauss-Markov assumptions (Washington et al., 2003). If these assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. Best Linear Unbiased Estimator. In other words, it is unbiased and has the lowest total variance among all unbiased linear estimators.

The basic Gauss-Markov assumptions require: Linearity (in the parameters; nonlinearity in the variables is acceptable); Homoscedasticity; Exogenous independent variables; Uncorrelated disturbances; and Normally distributed disturbances.

Interval regression

It is often the case, especially concerning income or price data, that information is missing on the exact value of the explained variable. Instead, the only available information is that the explained variable lies in some interval. In that case, a maximum likelihood estimator can be used. The likelihood then corresponds to the probability that the explained variable lies in the observed interval. The statistical structure of the model and the assumptions are similar to simple linear regression.

2.2.2 Multinomial Logit (MNL)

The most common discrete choice model is the linear in parameters, utility maximizing, multinomial logit model (MNL), developed by McFadden (1974). One of the most noteworthy aspects of the multinomial logit model is its property known as Independence from Irrelevant Alternatives (or IIA), which is a result of the i.i.d. disturbances. The IIA property states that, for a given individual, the ratio of the choice probabilities of any two alternatives is unaffected by other alternatives. This property was first stated by Luce (1959) as the foundation for his probabilistic choice model, and was a catalyst for McFadden's development of the tractable multinomial logit model. There are some key advantages to IIA, for example the ability to estimate a choice model using a sample of alternatives, developed by McFadden (1978). However, as Debreu (1960) pointed out, IIA also has the distinct disadvantage that the model will perform poorly when there are some alternatives that are very similar to others (for example, the now famous red bus – blue bus problem); this can be a significant concern when dealing with the models in software such as UrbanSim where a large number of rather similar alternatives may be available.

2.2.3 Nested Logit (NL)

There are many ways to relax the IIA assumption, and many variations of discrete choice models aim at doing just that. Nested logit (NL), introduced by Ben-Akiva (1973) and derived as a random utility model as a special case of GEV by McFadden (1978, 1981), partially addresses this issue by explicitly allowing correlation within sets of mutually exclusive groups of alternatives. The nested logit is widely used in practice due to its extremely tractable closed form solution.

Multinomial and nested logit are the workhorses of discrete choice modeling, and form the foundation of models in areas such as travel demand modeling and marketing. This is because they are extremely tractable and fairly robust models that are widely described in textbooks (for example, Ben-Akiva and Lerman, 1985; Greene, 2000; Louviere et al., 2000; Ortuzar and

Willumsen, 1994) and can be easily estimated by numerous estimation software packages (for example, biogeme, Bierlaire, 2003). Nested logit models have been used to estimate extremely complex decision processes, for example, detailed representations of individual activity and travel patterns (see Ben-Akiva and Bowman, 1998).

Beyond MNL and NL, there are many directions for enhancements that are pursued by discrete choice modelers. Two of these categories of models (mixed MNL and latent variable models) are outlines in the next two sections.

2.2.4 Mixed MNL (MMNL)

Mixed logit is a highly flexible model that can approximate any random utility model (McFadden and Train, 2000). It obviates the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Unlike probit, it is not restricted to normal distributions. Its derivation is straightforward, and simulation of its choice probabilities is computationally simple. Like probit, the mixed logit model has been known for many years but has only become fully applicable since the advent of simulation.

A detailed description of mixed logit is available in Train (2003) and Walker (2001). The specification of a random coefficient mixed logit model uses the following utility specification (for a decision maker n choosing alternative j from a choice set of J alternatives) $U_{jn} = X_{jn}\beta_n + \sigma_j \varepsilon_{jn} + \nu_{jn}$ where X_{jn} are observed variables that relate to the alternative j and decision maker n, β is a vector of random taste parameters specific to individual n, ε_{jn} is a Gaussian, zero-mean error term, with a standard deviation σ_j, and ν_{jn} is a zero-mean, random term that is iid extreme value.

Several other approaches that allow for the explicit modeling of correlation among observations exist and could be applicable to this problem. To name a few: Normal mixing distributions (e.g. Abdel-Aty et al., 1997), Generalized Estimation Equation (GEE) models (an extension of generalized linear models) (e.g. Abdel-Aty and Abdalla, 2004), Heteroscedastic Extreme Value (HEV) model, and the multinomial probit. MMNL has several interesting properties that make it attractive. MMNL is conceptually very close to the MNL, which is arguably the most widely used discrete choice model. Furthermore, the tools to specify and estimate MMNL models have reached a level of maturity that can make them accessible to a wide range of researchers and practitioners. Finally, the MMNL is a fairly flexible model, as the additional error term may have a normal, uniform, log-normal or other distribution. The

additional term may also capture heteroscedasticity among individuals and allow correlation over alternatives and time. While each of these reasons may be relevant to some other method, the MMNL combines these arguments.

The most widely used model specification is the standard linear-in-the-parameters specification, used in the vast majority of such models. The actual choice of variables is determined based on data availability and estimation results of alternative considered models.

2.2.5 Latent variables

The nested Logit model is relevant when the upper level category is observable. This is the case, for example, for dwelling type or tenure type. In some cases, the upper level category is implicit and cannot be observed. This is the case, for example, for budget constraints, which prevent the constrained households to borrow in order to buy their dwelling, and so that they are bounded in the tenant category even though their expected utility is lower in this category than in the owner category. The modeller cannot know a priori which households are tenant because they chose so, and which households are tenant because they are budget constrained.

The latent variable model allows to model at the upper level of the nest the probability that the household is subject to binding budget constraints.

2.3 Dealing with data properties

2.3.1 Importance sampling

In a MNL model, under the IIA assumption, random sampling can be performed when the number of alternatives is too large. Extending random sampling to NL is not straightforward.

Importance sampling of a zone is equivalent to uniform sampling of dwellings located in the zone. The question is which dwellings should be taken into account.

Importance sampling should not prevent the same zone to appear twice or more in the choice set, but some econometric software does. In case the same zone cannot appear twice in a choice set, this leads to an under-representation of largest alternatives, which becomes more and more severe as the number of alternatives increases. This leads to a bias in the coefficients of all variables correlated with zone size. This bias should be corrected.

Note that the under-representations of large alternatives, and the resulting bias, become more and more severe when the number of alternatives in the individual choice sets is increased. As a result, the number of alternatives in individual choice sets should not be increased too much (10 alternatives randomly chosen for each household choice set was a reasonable figure for

household location choice in Paris case study) when the software used for estimating models does not allow for repetitions and does not correct the resulting bias.

The probability that a zone is included in a choice set is proportional to the "size" of the zone, which may be measured either as the population stock (number of dwellings existing in the zone, number of households living in the zone, or as a flow (number of movers to this zone, number of vacant dwellings in the zone).

Under the IIA assumption with importance sampling of alternatives, when the zones are large enough (say, more than 100 households each), aggregate demand can be consistently computed based on the probabilities computed in the individual choice sets. This means that, for computing aggregate demand, it is not necessary to compute the probability of each of the alternatives for each individual or household, which allows saving a lot of time when the number of alternatives is large.

On the opposite, in the nested logit model, inclusive value should be computed on the whole set of alternatives rather than only on the alternatives randomly selected in the individual choice set. A similar requirement (working on all alternatives rather than on the alternatives randomly selected in the individual choice set) holds for computing segregation effects or, more generally, when focusing on the geographical distribution of population characteristics.

2.3.2 (Pseudo-)Panel data

Random effects/fixed effects

The data that are used in the UrbanSim models come from several time periods. When dealing with such panel data it is often useful to consider the heterogeneity across individuals, often referred to as unobserved heterogeneity. In general, pooling data across individuals while ignoring heterogeneity (when it is present) will lead to biased and inconsistent estimates of the effects of pertinent variables (Hsiao, 1986). Several approaches have been developed to incorporate these effects in the model formulation.

One such approach is to estimate a constant term for each individual and each choice, which is referred to as a "fixed-effects" approach (Chamberlain, 1980). Perhaps the main drawback to this approach is the large number of parameters (and consequently large number of required observations per individual). A more tractable approach is to assume that the fixed term varies across individuals according to some probability distribution, which is referred to as a random effects specification (Heckman, 1981; Hsiao, 1986).

2.3.3 Spatial econometrics

Spatial effects represent some of the main methodological challenges that have to be tackled in first-stage hedonic regression. We may distinguish two kinds of spatial effects: spatial dependence and spatial heterogeneity.

Spatial dependence may be "considered as the existence of a functional relationship between what happens at one point in space and what happens elsewhere" (Anselin, 1988). Many recent hedonic price studies suggest that in a cross-sectional hedonic price analysis, the value of a property in one location may also be affected by the value of other properties located in its neighboring area (Yusuf, 2004). Two broad causes may lead to spatial dependence. Firstly, there is the byproduct of measurements errors for observations in contiguous spatial units. In several cases data are collected only at aggregate scale. This often implies a poor correspondence between the spatial scope of the phenomenon under scrutiny and the delineation of the spatial units of observations and thus potential measurement errors. Those errors will tend to spill over across the frontiers of spatial entities as one may expect that errors for observations in one spatial unit are likely to be correlated with errors of neighboring geographical entities (Anselin, 1988). A more fundamental cause of spatial dependence is due to varieties of interdependencies across space. Location and distance do matter and formal frameworks proposed by spatial interaction theories, diffusion processes, and spatial hierarchies structure the dependence between phenomena at different locations in space (Anselin, 1988).

Spatial heterogeneity is related to the lack of stability over space of the behavioral or other relationships under scrutiny. It implies that functional forms and parameters vary with location and are not homogenous across the dataset. Several factors, such as central place hierarchies, the existence of leading and lagging regions, vintage effects in urban growth, etc., suggest modeling strategies considering the particular characteristics of each location or spatial entity (Anselin, 1988).

It has been amply demonstrated that the neglect of spatial considerations in econometric models not only affects the magnitudes of the estimates and their significance, but may also lead to serious errors in the interpretation of standard regression diagnostics such as tests for heteroskedasticity (Kim et al., 2003).

Several contributions have attempted to control for spatial effects in first stage hedonic price estimation. They mostly use two kinds of frameworks: Spatial econometrics models or Geographically Weighted Regression. There is no consensus about the variety of solutions pro-

posed in the literature. The best modeling strategy often depends on the specificity of the case study investigated.

Spatial econometrics models capture spatial dependency in econometrics models, avoiding statistical issues such as inconsistent or inefficient parameters estimates. In those models, spatial dependency can be handled in several ways. Indeed, in the spatial econometrics toolbox we distinguish: the Spatial Autoregressive Model (SAR), the Spatial Error Model (SEM), a mix of the SAR and the SEM – the Spatial Mixed Model (SMM) – and the Spatial Durbin Model.

In a SAR model, both the direct and indirect effects of a neighborhood's housing characteristics are captured through a spatial multiplier. This model is particularly appropriate when there is structural spatial interaction in the market and the modeler is interested in measuring the strength of that relationship. It is also relevant when the modeler is interested in measuring the "true" effect of the explanatory variables, after the spatial autocorrelation has been removed.

A contrario, in a SEM model, spatial autocorrelation is assumed to arise from omitted variables that follow a spatial pattern (Kim et al., 2003). Conversely to the SAR model, the SEM is appropriate when there is no theoretical or apparent spatial interaction and the modeler is interested only in the correction of spatial autocorrelation (Anselin, 2001).

The Spatial Durbin Model includes a spatial lag of the dependent variable as well as spatial lags of the explanatory variables. This model is an extension of the SAR that allows the structural characteristics of neighboring houses to influence the price of each house. It also captures how the price of houses in one area depends on the characteristics of neighboring areas (Brasington and Hite, 2005).

Besides spatial econometrics models, Geographically weighted regression (GWR) is a local version of spatial regression that generates parameters disaggregated by the spatial units of analysis. This allows assessment of the spatial heterogeneity in the estimated relationships between the independent and dependent variables.

Most of the contributions using those models assume that the dependent variable, house price or dwelling rent, is continuous. In the Brussels case study, we have to handle an issue: the information about our dependent variable, dwelling rent, is collected through a categorical variable. Each modality of this discrete variable refers to a unique interval of dwelling rent.

Therefore, we have to resort on techniques designed to estimate spatially dependent discrete choice models. Lesage and Pace (2009) provide a detailed overview of spatially dependent discrete choice models. From all those models, the ordered spatial probit model is the one that proposes the modeling strategy that is the closest to the one we have to implement. However, there are important differences between our "Spatial Interval Regression" model and the ordered spatial probit model. In the ordered spatial probit model, the cut points separating interval of the latent variable are unknown. Therefore, there is an identification issue and the variance has to be normalized to one so that regression coefficients as well as cut points may be estimated. In our model the vector of boundaries of the dependent variable is known. Hence, regression coefficients as well as the variance may be jointly estimated. A similar analysis has already been undertaken by Goffette-Nagot *et al.* (2010). They explore the spatial variation of land prices in Belgium. While they also account for spatial autocorrelation, their analysis differs since they consider land prices rather than rents as their dependent variable. Moreover, land price information is collected at the level of the municipality rather than at an individual level.

2.3.4 Endogeneity of variables and selection bias

Endogeneity is a serious problem commonly faced in LUTI models interested in interactions between modules. A typical example is given by the prices in the household location choice model, which is correlated with the error term. This problem is caused either by the simultaneous determination of the supply and the demand for dwelling units, or by omitted attributes that are correlated with price. Indeed, empirical residential location choice models have often reported estimated coefficients of dwelling-unit price that are small, statistically insignificant, or even positive. This would imply that households are insensitive to changes in dwelling unit prices, which is not only counter-intuitive, but also makes the models useless for policy analysis. See de Palma et al. (2005, 2007) or Guevara and Ben-Akiva (2005) for examples and discussions.

When endogeneity results from omitted attributes, the best solution is to include enough explanatory variables in the model of interest. Instrumental variables technique can be used to correct for endogeneity, provided that at least one instrument is available for each endogenous variable. It often proves to be difficult to find such instruments. In the case of household location, if it can be reasonably assumed that dwellings and offices compete for land, then variables measuring local business tax can be used to instrument dwelling prices. In their application on Paris case study, de Palma et al. (2005) used such instruments and found that en-

dogeneity bias becomes negligible when the household location choice model is rich enough (i.e. when enough explanatory variables are included). Note that a rich enough model can be estimated precisely enough only when sample size is large enough, which typically means at least 50,000 households.

2.4 Diagnostics

Model diagnostics are a key tool in developing appropriate models. In general there are two families of diagnostics: statistical and graphical. In order to ensure that the output of the various case studies within SustainCity are consistent and comparable, we need to ensure that the same diagnostics are provided. Each table of results should contain, for each explanatory variable, the following four pieces of information: Estimated coefficient, Standard error, T-statistic, p-value. For summary tables comparing multiple models, it is sufficient to present the estimated coefficient value and t-statistic. In terms of summary statistics, regression results should report corrected R^2 for linear regression. For MNL/NL/MMNL/latent models that are estimated using maximum likelihood, the null log likelihood and the final log likelihood should be reported, along with the AIC. Degrees of freedom should also be reported. Likelihood ratio test values should be performed to determine whether model restrictions should be retained or whether the more general models should be used. Similarly to reporting corrected R^2 for linear regression, it is recommended that corrected likelihood ratio test values be reported.

The econometric models described in this document have some explicit underlying assumptions that need to be satisfied by the data, in order to be valid. A number of violations may often occur, however, resulting in residuals that are not independently and identically distributed. In order to ensure that these assumptions are not violated (or, to be able to resolve them, or at least consider their implications), it is important to perform a series of tests, e.g. for normality, autocorrelation, endogeneity and heteroscedasticity.

One of the most effective ways to determine and visualize violations, such as autocorrelation and heteroscedasticity, is through the use of (partial) autocorrelation functions (sometimes called “corellograms”) and residual plots. Residual plots over time or against the magnitude of the dependent variable can help identify heteroscedasticity. QQ normal scores plots can be used to identify deviations from the normality assumption. These visual tests should also be accompanied and further supported by formal statistical tests. The Shapiro-Wilk test can be used to test the normality assumption. The computation of the skewness and kurtosis also provide additional information.

The Box-Ljung test should be used to for autocorrelation for various lags. A different way to test this type of lack-of-fit of a model is to consider the first few autocorrelations as a whole, using a so-called "portmanteau" test. It should be noted that the number of autocorrelations to use depends on the data and while a lag of 4 or 5 might be sufficient, using a lower lag might not illustrate the dependency. Larger lags do not add to the inference, but are also rather harmless in this context. A popular test for checking the heteroscedasticity assumption is the Breusch-Pagan test (Breusch and Pagan 1979). A usual way to test for endogeneity is to use a Hausman test.

3 Models to be estimated

Table 1 outlines the types of models that will be estimated for each model type and case study. These models are explained in the next subsections.

Table 1 Models by case study

Model	Paris	Brussels	Zurich
House-hold location	Nested: relocation/ dwelling type/ tenure status/ location	Multinomial Logit structure. Besides this, nested structures of choice will be tested in order to account for correlation of attributes across alternatives.	MNL with explaining variables of domains: life style, dwelling type, location (Household relocation: Probabilities for relocation of HH according to income and age)
Job location	Matching workplace/ business	Nested logit; sampling of alternatives	Hierachical NL of firm location choice (Bodenmann & Axhausen, 2010)
Real estate price	Simultaneous equation (5 types), spatial correlation, Dwelling level	Hedonic model; estimated using "interval regression". A spatial autoregressive model will be considered	Spatial error model (Löchl and Axhausen, 2010)
Land developt	Matching project location/land use transition	2-step model: Supply by building type per zone: linear regression/Choice of zone: Multinomial logit	NL with explaining variables of domains: project, developer and development constraints

3.1 Household Location Choice Model (HLCM)

The model is estimated using MNL with importance sampling. Extensions such as NL, MMNL or latent variables were estimated, but are discussed here because they cannot be implemented yet in the current version of UrbanSim. In case of NL, stratified sampling is an option to be discussed. Household location choice model could be estimated on the whole sam-

ple, irrespectively of tenure type and dwelling type. However, when possible, we recommend that tenure type and dwelling type are considered separately, with coefficients specific to each tenure type and dwelling type, and that the decision to move (relocation choice) is estimated together with location choice. In this case, we recommend the following nested structure: 1) decision to move; 2) tenure choice; 3) dwelling type; 4) Location.

An extension to latent variables was successfully estimated for Paris case study, but it will probably not be included in UrbanSim in the near future. In this experimental latent variable model, two cases are considered for step 2) tenure choice: under credit constraint, the only option available to the household is to rent, unconstrained households are free to choose either renting or buying a dwelling. The probability of credit constraint is estimated simultaneously with the other parts of the model, as an upper level conditional on moving. Additional extensions are scheduled at the bottom level, for dealing with geographical nests and Scalability.

Endogeneity of prices is a serious problem in HLCM. It can be solved by instrumenting dwelling prices. Instruments are not obvious in this context, and the choice of instruments is guided by assumptions concerning the real estate markets. In case dwellings and offices are competing for land, instruments can be found in the list of variables influencing the demand for offices. In Paris case study, variables related to local business tax (French *Taxe professionnelle*) appeared to be valid instruments.

3.2 Jobs location/Firmography

A distinction is operated between firms and plants. The way plants can be related to firms depends mainly on data availability. When the identifier (Id) of the plant is not maintained because of the move, this induces fake deaths and births, since the available data does not allow to distinguish between a move and a death & birth when the plant Id changes. All models estimated are sector-specific, since the dynamics of the job market significantly varies across activity sectors. Given the stability of activity sector either from the plant point of view or even from the worker point of view, no model is estimated for transition between sectors.

Three options may be used to study employment location: jobs location, either by itself or together with household location, and firmography. Each of these models uses Multinomial Logit (MNL) or Nested Logit (NL). In the simplest option, each job is located independently from the other jobs in the same firm or plant and from Household location, using a Multino-

mial Logit (MNL) model. This simplest option should be considered as a second best, less relevant than the other ones.

The second option, relevant from the point of view of the worker, builds a more elaborate job location choice model. It is a Nested Logit (NL) for workplace and Household location, in either order. In such a model, commuting time is a key variable explaining the location at the lower level of the nest, which happens to be by far more significant than any variable measuring either accessibility or expected time typically used in location choice models.

In the third option, firmography, relevant from the point of view of the firm, all workers working in the same plant are located simultaneously, at the same place. In addition to the location of new plants, firmography estimates the "death" of the plants using a binary logit model, as well as growing/shrinking of stable plants, using a Linear Regression model.

Note that the "birth" of plants, which is implemented in UrbanSim is not estimated. In the simulation process, newly born plants are randomly selected from the distribution of existing plants.

3.3 Real Estate Price Model

Real Estate Price Models corresponds to Hedonic price models, including simultaneous regressions, which are relevant in the case of imperfect real estate markets. Two distinctions should be operated in the model: a first distinction between renting and selling, and a second distinction between houses and flats. The relevance of these distinctions depends on: (i) correlations between various prices in the same location, (ii) market shares of each category and (iii) turnover in each category.

The Real Estate Price Model (REPM) used the following methods: Simple linear regression, Interval regression, and Spatial correlation (SAR, GWR). These methods can be applied either on Individual prices or on average local prices. In UrbanSim, parcel version uses data on individual buildings. In this case, the determinants of real estate prices include both individual and local attributes). Linear regression models are estimated on the log of total selling price, including surface in the regressors list. Aggregate data are on average prices per m². In this case, only the local attributes will be included in the regressors list. Potential determinants of the prices are: (i) Individual attributes with building characteristics (surface, age, view) and (ii) Local attributes with accessibility (to jobs, households, activities), neighborhood (households, jobs, land use) and distance to stations.

3.4 Land Development Model

This is the less advanced model in the 3 case studies (and in UrbanSim). UrbanSim currently proposes two options, which are substitutes for the moment. It is desirable that UrbanSim can evolve so that these two options are complements, and describe respectively the supply and demand for land, in relation to the politicians or stake holder versus investor points of view:

Stake holder point of view: MOS (MOS=Land Use Type) transition model, in which there is choice between the different land use types and transition between land use types for a given parcel. Formulated as a MNL with a relatively limited choice set (e.g. there are 83 land use types in Paris region, which can be grouped in 9 homogenous aggregated types). Trade-off between competing land uses can be considered.

Investor point of view: captures the choice between the different locations and location choice model for a given project. Formulated as a MNL with importance sampling (trade-off between competing locations for a given project). The list of potential alternatives depends on the land use type attached to the project and on the surface of the project: the project can be located only in parcels (e.g. communes or IRIS) for which the surface available for this land use type is larger than the surface of the project.

A **Common initial model** is used for the project definition. A project is defined as a parcel which changed detailed land use type. It is characterized by location (parcel ID, geocoded), size, former land user type, and new land use type. This is the upper level of a NL model for either option (=either point of view).

4 Conclusions and recommendations

This section summarizes some of the first lessons that have been obtained from the on-going case studies within the SustainCity project.

4.1 Lessons from case studies

4.1.1 Depending on data availability, find the best econometric strategy for each model

The need for detailed and reliable data has been motivated in this document, along with the difficulties in obtaining and using such data. However, data availability, quality and restrictions are in general location and application-specific. Therefore, it is not practical to try to develop general econometric strategies that would be applicable (let alone effective) in all settings. Instead, the modelers should be able to find the "best" econometric strategy for each model, based on the data availability. When detailed data of good quality are not available, a

more parsimonious or aggregate model might provide more reliable data than a more detailed and elaborate model (that may not be adequately supported by the data).

4.1.2 Compare estimation results obtained with an econometric software and with UrbanSim until you get exactly the same results

UrbanSim provides facilities for the estimation of econometric models, the results of which can be then used for modeling purposes. It is also possible to estimate models outside of UrbanSim and then use the estimated coefficients within UrbanSim for simulation purposes. Consistency between the model estimation and simulation is of paramount importance for the credibility of the results. As there are multiple secondary reasons that might obfuscate the model estimation process, it is recommended that UrbanSim model estimation results are compared against standard econometric software (that the modeler is familiar with) to make sure that the data and underlying assumptions made by UrbanSim are indeed understood correctly.

One practical way to verify that the results obtained by the two models are consistent is to use a systematic test "a la Hausman" (Hausman, 1978).

4.1.3 Endogeneity issue and order for running models

UrbanSim involves the running of a sequence of models in cycle. This type of models is known to suffer from endogeneity, which can have significant implications in the model results. In order to minimize the impact of this problem, it is important to ensure that the model coefficient estimates and the order in which the simulations are run are consistent.

4.2 "Standardized views"

The objective of this subsection is to provide some practical guidance towards the development of uniform and "standardized" views. This is particularly important in order to be able to develop some composite insight from the output of the models developed for the various cities (but also among different model forms for the same type of model within applications). The identified suggestions reflect already identified items and it is expected that during the further development of the models further similar suggestions will need to be made to ensure uniformity.

4.2.1 Vocabulary and units

When dealing with buildings/apartments, specific distinction should be made between floor space and land area (cover), in order to avoid confusion. The unit that should be used is (square) meters. Floor space indicates the total area of the property (in m^2); for example if an

apartment is spread in two floors and has 100m^2 per floor, then the total floor space should be reported as 200 m^2. [The fact that this is a two-story apartment, which could be considered an advantage e.g. for the real-estate price model, could be e.g. captured by an additional explanatory parameter in the hedonic regression model].

On the other hand, land area (cover) reflects the physical space that a property occupies on the land. For example, a 6-story commercial building with 200 m^2/floor would have floor space of 1200 m^2 and land area of 200m^2.

The unit for monetary measures (e.g. income/cost/rent) should be Euro (€) in all cases, in order to provide uniformity and more direct comparisons. When the original data is in another currency (e.g. Swiss Franc, CHF), then they should be converted to Euro.

When values are considered in logarithm, the neperian logarithm should be used, and denoted by ln, in order to avoid confusion.

4.2.2 Results presentation

As it has been presented earlier in this document, UrbanSim applications for different cities within SustainCity involve different types of models, at different granularities. Therefore, it is expected that the results may vary substantially. However, there are some measures that can be taken towards providing coherent results that can be used to perform some meta-analysis. One such aspect that relates to the price levels is the recommendation to use a log-transform (both for the model estimations and presentation of results).

Another price-related aspect that can have a large impact on the reported summary statistics is the way that property prices are used. In particular, using total property prices leads to summary statistics (e.g. R^2) with much higher values (than if prices per m^2 are used). Therefore, the results will be more easily supported if they are accompanied by these statistics and it is recommended that total property prices are used in the model estimations.

Acknowledgements

This research has been partly funded through the SustainCity project, co-financed by the European Commission within the FP7 program. Furthermore, the authors would like to acknowledge the contribution of the following persons (alphabetical order) in the work that led to this publication: B. R. Bodenmann, L. Chauveau, D. Efthymiou, P. Fastre, S. Gayda, R. Hurtubia, J. Jones, K. Motamedi, K. Müller, H. Ouaras, D. Peeters, A. Pholo-Bala, P. Schirmer, I. Thomas, and C. Zöllig.

References

Abdel-Aty, M. and Abdalla, M. F. (2004). Modeling drivers' diversion from normal routes under ATIS using generalized estimating equations and binomial probit link function, Transportation, 31(3), 327-348(22).

Abdel-Aty, M., Kitamura R. and Jovanis, P. (1997). Using stated preference data for studying the effect of advanced traffic information on drivers' route choice, Transportation Research C, 5.

Anselin, L. (1988) Spatial Econometrics: Methods and Models. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (2001) Spatial econometrics. In: Baltagi B. (Ed.) A Companion to Theoretical Econometrics. Oxford : Basil Blackwell, pp. 310 - 330.

Ben-Akiva, M. (1973) Stucture of Travel Passenger Demand Models, Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Ben-Akiva, M. and J.L. Bowman (1998) "Activity Based Travel Demand Model Systems", Equilibrium and Advanced Transportation Models, P. Marcotte and S. Nguyen, Eds., Kluwer Academic Publishers.

Ben-Akiva, M. and S. Lerman (1985) Discrete Choice Analysis: Theory and Application to Travel Demand, The MIT Press, Cambridge, MA.

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

Bodenmann, B.R. und K.W. Axhausen (2010) Synthesis report on the state of the art on firmographics, SustainCity Working Paper, 2.3, IVT, ETH Zürich.

Brasington, D. M., and D. Hite (2005) Demand for environmental quality: a spatial hedonic analysis, Regional Science and Urban Economics 35, 57-82.

Breusch, T.S. and Pagan, A.R. (1979). "Simple test for heteroscedasticity and random coefficient variation". Econometrica, 47 (5): 1287–1294.

Chamberlain, G. (1980). Analysis of covariance with qualitative data, Review of Economic Studies, 47, 225-238.

Debreu, G. (1960) Review of D. Luce Individual Choice Behavior: A Theoretical Analysis, American Economic Review 50, 186-188.

de Palma, A., K. Motamedi, N. Picard, and P. Waddell (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., K. Motamedi, N. Picard, and P. Waddell (2007). Accessibility and Environmental Quality: Inequality in the Paris Housing Market.

Goffette-Nagot, F., I. Reginster, and I. Thomas (2010) Spatial Analysis of Residential Land Prices in Belgium: Accessibility, Linguistic Border, and Environmental Amenities Regional Studies First published on: 17 August 2010 (iFirst).

Greene, W.H. (2000) Econometric Analysis Fourth Edition, Prentice Hall, Upper Saddle River, New Jersey.

Guevara, E. and Ben-Akiva, M. (2005) Endogeneity in residential location choice models, paper presented at TRB meeting.

Hausman, J.A. (1978). Specification Tests in Econometrics, Econometrica, 46 (6), 1251–1271.

Heckman, J. (1981). The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process. In Structural Analysis of Discrete Data with Econometric Applications. C. Manski and D. McFadden, editors. MIT Press, Cambridge, MA.

Hsiao, C. (1986). Analysis of Panel Data. Cambridge University Press, Cambridge, U.K.

Kim, C.W., T.T. Phipps, and L.A. Anselin (2003) Measuring the benefits of air quality improvement: a spatial hedonic approach, Journal of Environmental Economics and Management 45, 24 - 39.

LeSage, J.P. and R.K. Pace (2009) Introduction to spatial econometrics. Boca Raton: Chapman & Hall/CRC.

Löchl, M. and K.W. Axhausen (2010) Modelling hedonic residential rents for land use and transport simulation while considering spatial effects, Journal of Transport and Land Use, 3 (2) 39–63.

Louviere, J.J., D.A. Hensher and J.D. Swait (2000) Stated Choice Methods: Analysis and Application, Cambridge University Press.

Luce, R.D. (1959) Individual Choice Behavior, Wiley, New York.

McFadden, D. (1974) "Conditional Logit Analysis of Qualitative Choice Behavior", Frontiers of Econometrics, P. Zarembka, Ed., Academic Press.

McFadden, D. (1978) Modelling the choice of residential location, in A. Karlquist, L. Lundqvist, F. Snickars, and J. Weibull (eds.), Spatial Interaction Theory and Residential Location, 75-96, North-Holland, Amsterdam.

McFadden, D. (1981) "Econometric Models for Probabilistic Choice", Structural Analysis of Discrete Data with Econometric Applications, C. Manski and D. McFadden, Eds., Harvard University Press.

McFadden, D. and Train, K. (2000). Mixed MNL models of discrete response, Journal of Applied Econometrics, 15, 447–470.

Ortuzar, J. de. D. and L.G. Willumsen (1994) Modelling Transport, Second Edition, John Wiley and Sons Ltd.

Train, K. (2003). Discrete Choice Methods with Simulation. Cambridge University Press.

Yusuf, A.A. (2004) Does Air Pollution Affect Property Value? A Hedonic Price Analysis in Jakarta. Mimeo.

Walker, J. L. (2001). Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables. Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Washington, S. P., Karlaftis, M. G., and F. L. Mannering (2003). Statistical and Econometric Models for Transportation Data Analysis. Chapman & Hall/CRC.

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), 2003, 43-67.