Parcel-Level Microsimulation of Land Use and Transportation: The Walking Scale of Urban Sustainability

Paul Waddell
University of California Berkeley


1 Introduction

A key motivation for advancing the state of the art in integrated microsimulation modeling is the rising urgency finding combinations of land use, transportation and other policies that will dramatically reduce greenhouse gas emissions and fossil fuel consumption. Cities and metropolitan regions are the laboratories within which pathways to sustainability will be created, and we are unlikely to traverse these paths by single-occupancy vehicle.

The scale of the pedestrian is becoming a focal point for creating sustainable strategies. Neo-traditional neighborhood designs, transit-oriented development, and substantial research into the interactions between travel choices and neighborhood urban form highlight the importance of the walking-scale in promoting walking and transit use. In this paper we make a case for integrated parcel-level microsimulation of land use and travel behavior, explore recent advances that make it more feasible, and lay out some of the research opportunities (challenges, if one prefers their glass half-empty) that lie ahead.

This paper was drafted as a resource paper for a workshop on Computational Algorithms and Procedures for Integrated Microsimulation Models at the 2009 IATBR Conference. It does not pretend to provide a comprehensive, or perhaps even thorough, review of advances in integrated microsimulation modeling of land use and transport. It instead focuses on developing a coherent assessment of the case for such modeling, focusing on the particular case of microsimulation at the parcel level of detail. The citations draw heavily on a cluster of related, new and ongoing projects, and are intended to provide a sampling sufficient to sketch outlines of research that will likely unfold over the next several years.

2 The Emergence of Integrated Microsimulation Modeling, and a Reality Check

Early work in developing integrated models aggregated agents and used very coarse geographic units of analysis, typically by aggregating the zones used in travel models into large districts. Examples of this aggregated modeling approach include DRAM/EMPAL (Putman, 1983), TRANUS (de la Barra, 1990), MEPLAN (Eschenique et al, 1990), METROSIM (Anas, 1994), and MUSSA (Martinez, 1996). In fact, virtually all land use and all integrated models prior to the mid 1990’s were very spatially aggregate, and static equilibrium, in nature.
There were early, arguably heroic, efforts to begin moving land use models into the realm of microsimulation. These include MASTER (Macket, 1990), the Dortmund model (Wegener, 1985), and the NBER HUDS Model (Kain and Apgar, 1985). Following these pioneering efforts, more ambitious and comprehensive integrated microsimulation models began to emerge, including UrbanSim (Waddell, 2000, 2002; Waddell et al., 2003), ILUTE (Salvini and Miller, 2003), ILUMASS (Moeckel and Wegener, 2007), and PECAS (Hunt and Abraham, 2003).

By some measures, the emergence of integrated microsimulation modeling has gone extremely well. Academic research in this area is emerging at a growing rate, and the visibility of this kind of modeling has even begun to factor into legislation such as the State of California SB 375 bill, passed in 2008, that fairly explicitly directs metropolitan regions to undertake integrated land use and transportation modeling. By other measures, however, there are warning signs that expectations may exceed capacity to deliver.

Wagner and Wegener (2007) state in their assessment of the ILUMASS project, an ambitious effort to undertake a large-scale integrated microsimulation model system in the European context:

“The ILUMASS project, although quite successful in terms of the progress made in the individual subprojects and from the perspective of bringing different disciplines together, has not met all of its initial goals. It was not possible to run more than a few test scenarios and the results from these scenarios have to be taken with some caution, because only very basic checks for plausibility and consistency could be made. This experience is not uncommon with large urban microsimulation projects in recent years. Many of these projects had to readjust their plans when the project targets proved to be too ambitious.”

Similar stories arise on the American side of the pond. The Oregon Statewide Modeling effort, perhaps the most ambitious, long-running and expensive effort in the U.S., has been in continuous development since 1995. After 15 years of investment, it is in its third incarnation (based on the PECAS model), its developers and users conclude the following about its performance (Weidner et al., 2009):

“Model runtimes are longer than desired. A 19-year run takes 3.5 days to complete and outputs consume 65GB of disk space. An additional three hours are required to process the model output into summary tables and graphs. Convergence became more time consuming in the later years, revealing the potential for longer runtimes for some analyses, and the need for a convergence criteria that expands with increased economic activity.”

Wegener (2009) claims that none of these integrated microsimulation models have yet to be brought fully calibrated and brought into operational use, though there is room for dissent on this point. If all of this sounds vaguely reminiscent of the concerns raised by Lee (1973, 1994) about early large-scale urban models, it should. The title of a recent presentation by Michael Wegener, one of the pioneers in urban microsimulation, was “From macro to micro: how much micro is too much?” (Wegener, 2009). The key points of his bleak assessment of the state of integrated microsimulation modeling raised two
different kinds of concerns: excessive run times that make it impossible to properly calibrate or evaluate the models, and concerns about stability and stochastic variation make it exceedingly hard to assess the validity of these models – what some might call the signal to noise ratio.

In spite of these quite serious concerns, we take a more optimistic view of the prospects for integrated microsimulation. In the balance of this paper, we chart the case for and the issues arising from moving to a logical endpoint of spatial disaggregation within integrated microsimulation models: the parcel and building level. We then return to the challenges raised by Wegener and others, of how to achieve sufficient computational performance to make these models usable in a practical sense, and how to address concerns of model instability and calibration.

3 How Much Micro is Enough?

In microsimulation models, the question of how to disaggregate geographic units of analysis from the earlier use of large districts had to be addressed. The problems of aggregation bias are well understood, as are the side effects of the Modifiable Areal Unit Problem (MAUP). Large units of spatial analysis also prevent the assessment of localized, walking-scale accessibility that is increasingly important to address the importance of walking access for transit mode use, and for walking to shopping and other activities that are important from a health perspective.

Although travel model zones offered a logical increase in spatial resolution from the earlier use of highly aggregated districts, the zones used in most current travel models are still generally too coarse to provide reasonable walking scale measures. The most common unit of analysis for microscopic land use models is the gridcell, using resolutions ranging from that available in Landsat imagery (30 meters) to 150 meters, as was common in early versions of UrbanSim (Waddell, 2000, 2002). Most land cover change models, and Cellular Automata approaches to modeling land use change, also adopted gridcells as the unit of analysis.

The advantages of gridcell (raster) data structures are well known. There are efficient for storing and manipulating large volumes of spatial data. Most GIS systems provide excellent support for efficiently manipulating raster datasets with a form of map algebra that formed the basis of most early GIS systems. Spatial queries on a raster are very fast. Aggregation of rasters to coarser resolution is trivial. These and other advantages provide substantial motivation for the raster unit of analysis.

The grid, or raster, data structure, of course, has limitations as well as benefits. The fact that gridcells are uniform units of analysis means that they ultimately do not conform to other underlying geographies, unless the resolution of the gridcell is exceedingly high (exceedingly small cells). At 150 meters, for example, a single cell represents more than 5.5 acres, and approximates the size of a suburban residential block. Parcels, streets, buildings, and every other natural and built element will cross boundaries of coarse gridcells, broadening the aggregation problem into a messier measurement problem due to confounding mixtures of underlying features. Often the spatial distribution of parcels
and buildings is quite irregular, and higher level geographies such as traffic analysis zones or gridcells do not reflect the underlying data well, as shown in Figure 1. One could adopt a very high resolution grid for the modeling, for example 10 meters, but the problem does not disappear, it changes to a different one. With very high-resolution cells, one typically would observe a fragment of an underlying feature – a portion of a lawn, or a rooftop, or a street. High resolution cells make no behavioral sense as a unit of analysis. Workarounds have been developed to handle some of these problems, but ultimately the data structure remains problematic for microsimulation modeling of urban activities and built form. The raster data structure is better used as a means of data representation or integration, but not for behaviorally realistic modeling of the built environment.

Figure 1. Traffic Analysis Zones, Parcels and building footprints from Honolulu

The most logical candidate from a behavioral perspective is the parcel of land, and individual buildings on parcels. Land use regulations such as municipal comprehensive plans, zoning, and building codes, apply directly to parcels. Setback and height regulations apply to buildings and their physical placement on parcels. These are important tools of local policy, and if we are to make integrated models responsive to the land use policies available to local communities that control most land use regulation, it is extremely valuable to use geographic units of analysis relevant to those policies.

The advantages of using parcel geography are straightforward. They represent objects that are behaviorally meaningful: parcels are bought and sold, regulated, and developed.
Parcels are also increasingly inventoried, mapped, and monitored by local agencies charged with property tax assessment. Parcels are behaviorally meaningful in the context of integrated land use and transportation since they also ultimately represent the locations of activities. Hence, accessibility measures can take advantage of the resolution provided by parcel geography, for example, to determine whether transit stops or grocery stores are within walking distance from parcels, and might substitute walking trips for those by car.

4 Making Parcel Data Usable for Integrated Microsimulation Models

As should be expected, the benefits of parcel geography do not come without associated costs. The list of costs begins with gaps in the availability of parcel data. Although parcel data is increasingly available from local property tax assessment agencies, there are many parts of the United States that still do not have digitized parcel maps, and many other parts of the world that have even less availability of such data. Worldwide mapping and monitoring of parcel level data is advancing rapidly, and this will become less of an issue in the future – but it remains a significant constraint today. Even for localities that have digital parcel maps available, there are often significant gaps in the attribute data available. Valuable information on buildings is often missing, such as square footage, number of stories, year built, or building type. This problem is particularly acute for the parcels that are exempt from property tax collection, since there is little motivation for tax assessors to spend time creating useful data for parcels that will not be assessed taxes. There is a significant need, therefore, for data imputation and cleaning algorithms for these incomplete data sources, and there may be a need for synthesis in the case of more extensive gaps.

4.1 Data Imputation and Cleaning

Two alternative pathways may be useful to consider, depending on the details of data availability and modeling objectives. The first approach applies where the source data is generally available, but is subject to missing values and errors. The goal in this approach is to leave the data as close to the original data as possible, and augment the subjective reasoning capacity of users of the data to be able to make their effort much more efficient in the arduous task of data cleaning and integration.

Until recently, most strategies to accomplish this task could be classified as either manual inspection and cleaning (brute force), or rule-based or heuristic methods to solve one type of problem at a time, in some user-specified sequence. The first approach is extremely expensive from a time and effort perspective, and is prohibitively intensive in most real-world cases. The second approach is more widely used, with GIS and database queries used to isolate missing or problematic data, and decision rules applied to impute data for these values. A plausible approach is to sample nearby properties of a similar type, and then draw a value from the sample in order to impute a missing or erroneous value. Unfortunately, these methods are generally ad-hoc, and there is no systematic way to assess their quality. Further, the results depend heavily on the order of the imputations, since other attributes are used to select appropriate observations from which to draw values for imputing a missing value.
Fortunately, more sophisticated methods that have been developed in the computer science fields of data mining and machine learning have been rapidly advancing, and are now being applied with some success to the task of imputing missing data and identifying and correcting errors in the data. Though there are very many algorithms within these research domains, we highlight two for expository purposes.

The K-Nearest Neighbors (KNN) algorithm uses a proximity in multi-dimensional space, including x, y coordinates for spatial proximity, and other attributes to measure similarity on other dimensions, to draw a sample of similar and nearby objects. From these, it makes a prediction based on the distribution of values on the outcome variable. A training data set is used to train the algorithm to maximize its predictive accuracy, and then the trained classifier can be applied to the full dataset. Further tuning can be performed incrementally on additional training samples.

\[ L^n(x_1, x_2) = \sqrt{\sum_{i=1}^{\#\text{dim}} |x_{1,i} - x_{2,i}|^n} \]

\[ \hat{f}(x_q) \leftarrow \frac{1}{k} \sum_{i=1}^{k} f(x_i) \]

Support Vector Machines (SVM) provide an approach to maximize the separation between classes (clusters) in multi-dimensional data. As with KNN, these dimensions can include x, y coordinate space, so this methodology can simultaneously incorporate clustering in geographic space and in attribute space. The objective is to find the multi-dimensional hyperplanes that maximally separate clusters identified in the data, as shown in Figure 2. As with KNN, training data is used to tune or train the algorithm to maximize predictive success, and it can be refined iteratively.

Figure 2: Support Vector Machines for Imputing Categorical Variables

Machine learning libraries such as WEKA (2009) and ORANGE (2009) provide rich functionality and a user interface for constructing multi-stage data cleaning and
imputation on large datasets. Preliminary testing of customized application of Support Vector Machines for categorical attribute imputation such as for building type, and K Nearest Neighbor algorithms for continuous valued attributes such as the square footage of buildings, have shown promise. Draft versions of a data imputation toolkit for use in preparing UrbanSim data at a parcel level are currently in field testing by planning agencies in Seattle and Detroit, both of which have reported very positive results compared to previous data cleaning methods available to them.

4.2 Data Synthesis Using Procedural Modeling

In many metropolitan regions, parcel data is in relatively incomplete state of development, or might not exist at all. It is common to encounter situations in which the core counties of a metropolitan area have relatively complete parcel data and attributes, but contain some counties, often more rural, for which only parcel boundaries are available, with no attributes of buildings. And there are certainly localities for which there are not even parcel boundaries available. An alternative to the data imputation strategy using machine learning is a data synthesis approach that uses procedural modeling to ensure internal coherence and consistency among geometric and other attributes. This approach focuses more on the task of achieving an integrated microsimulation database that is internally consistent and coherent.

Prior work on the use of procedural modeling algorithms on an architectural scale set the stage for the more recent work on procedural modeling of landscapes. Müller et al (2007), for example, developed an approach to image-based procedural modeling of building facades. The facade of a building can be described with rules that reflect a pattern of placement of windows and other features, repeated on each story, with deviations for the ground floor and top floor. Further rules capture texturing and 3 dimensional aspects of the buildings, as depicted in Figure 3.

Procedural modeling has been extended to the landscape scale by Aliaga et al (2008). In this application, rules describing the geometry of parcels and streets, and textures from aerial imagery, are extracted from source data. These procedures are then used to impute missing sections of the landscape. While the original purposes for the algorithm were in computer graphics, the potential for adapting algorithms such as this to the problem of synthesizing missing parcel data should be readily evident from the example in Figure 4.
Figure 3. Image-based procedural modeling of facades (Müller et al, 2007). A building facade image that is used as input to the algorithm (top). A wireframe overlayed over the original image (middle). The resulting 3D model rendered using relighting and shadow maps (bottom).

Figure 4. Example-based urban layout synthesis (Aliaga et al, 2008). A new urban layout is generated by extracting and reproducing the structural attributes of the example fragment and reusing aerial-view imagery.
5 Modeling the Evolving Geometry of Parcels and Streets

A problem for the use of parcels in integrated models is that parcels change over time. Parcels may be subdivided in the course of conversion from a large farm to a housing development, for example. Or a developer may buy multiple parcels that are adjacent, and redevelop them as a single integrated development, by dissolving the original parcels and re-platting the new development. These are much more difficult problems, since they involve modeling the geometry of parcels as well as the economic behavior of land owners and developers.

A related problem to the changing geometry of parcels is the creation of new local streets as a byproduct of the real estate development process. A large housing development or mixed-use planned development will create an internal layout that contains local streets or even arterials, in order to provide access to the newly subdivided parcels.

Finally, it should be noted that mixed-use development is both becoming more common in practice, and of increased interest as a tool in land use planning to promote alternatives to making trips by car. It is becoming more important to be able to characterize mixed-use development in models, as a consequence. This problem is perhaps less difficult than some of the others raised previously, since it is straightforward to create a microsimulation list of building that are tied to a parcel, allowing a many to one mapping from buildings to parcels, and making mixed use development fairly simple to represent.

The approach for synthesizing parcels, streets and imagery in Aliaga et al (2008) has been extended in Vanegas et al (2009) to incorporate aspects of the coherence between socioeconomic attributes and the geometric characteristics of terrain, streets, parcels and buildings. This development has potential for application not only on the data synthesis task, but also to the task of modeling evolution of streets, parcels and buildings over time, as an integral part of an integrated simulation system. A schematic of the general approach in this work is represented in Figure 5, which depicts the key variables in the algorithm. These are specified as a set of differential equations using an underlying raster structure, and used to synthesize coherent urban spaces.
Figure 5. Dynamic system of interactions among behavioral and geometric components

The system developed in Venegas et al (2009) can be used to sketch only skeletal information and complete the details, as shown in Figure 6. This approach has been used to generate an urban layout that is strikingly similar to the actual place it is patterned after, even though it only required a few minutes of user input. The results of the comparison of the observed to synthesized data are shown in Figure 7.

Figure 6. Completion (Vanegas et al, 2009). Taking as input a set of user-specified landmarks (downtown location (a), terrain and highways (b), and parks (c)), the system automatically completes the rest of the city. The process consists of first computing the behavioral variables (population (d) and jobs (e)) and then the geometric variables (blocks, parcels (f) and buildings (g)) while respecting the landmarks. A view of the resulting 3D model of the city is shown in (h).
Figure 7. Validation (Vanegas et al, 2009). Following the automatic completion process shown in Figure 6, the system generates a city that closely resembles a real-world city. In this case, a user is given a satellite view (c) of Pacific Grove, California, and is asked to draw the terrain and highways, click on a downtown location and highlight the parks. The system generates the geometry of the roads and buildings (f) (colors are manually adjusted for comparison). Zoomed areas of the original (a, b) and synthetic (d, e) cities show how the density and style of the real city are reproduced at different locations, including high-density (a,d) and low-density (b, e) neighborhoods.

6 Modeling Location Choices at a Parcel Level

The potential to use parcel-level data in integrated modeling has been on the horizon for some time, with early forays including a model of aggregated location choice and parcel level real estate development (Waddell, 1998). Recent work has demonstrated that residential location choices can be effectively modeled at a parcel level, for datasets as large as San Francisco (Waddell et al, 2009) and Seattle (Lee et al, 2009). These models use a parcel or a building as their unit of location choice, and are able to incorporate attributes of buildings and parcels that have not previously been possible to integrate, such as square footage of building and land, property-specific price, and the walking-scale characteristics surrounding the parcel. Lee et al (2009) find that localized walk access to retail remains important in residential location choice, even in the presence of numerous other access measures such as generalized access to employment, individual worker access to their workplace, and time-space prism-based accessibility to non-work activities on the way home from work. These kinds of findings provide important progress in developing models that are useful in informing local planning for ways of making places more amenable to walking and transit use.

6.1 Location Choice and Prices

One of the persistent questions in the literature and in urban modeling more specifically, is how to model the interaction of prices and location choices. By far the most widely used empirical tool to model real estate prices is the ubiquitous hedonic regression model.
The hedonic regression model was developed by Rosen (1974) as a two stage model to predict both prices and demand. But virtually all hedonic models of real estate prices apply only the first stage of the model, and are reduced-form models that cannot truly separate demand and supply influences on prices. The hedonic model is a workhorse model in the literature, and still provides the most robust estimates of the implicit prices of key variables of interest, such as environmental amenities or the premium for walkable neighborhoods.

One problem in integrated models is that one would like to have robust predictions of prices and at the same time a consistent set of parameters for location choice models that reflect individual consumer demands. This has proven to be an elusive combination.

Models that emphasize equilibrium assumptions impose a very strong assumption that prices adjust to clear the market, and impose a model structure that simultaneously resolve location choice and prices by allowing prices to adjust using an algorithm that is not empirically based, but is instead assumed and imposed on the model. This has been difficult to reconcile with models such as hedonic regression that have both a strong theoretical foundation and a very robust empirical presence in the literature. Further, those who would claim that the housing market meltdown in the U.S. that precipitated a global economic downturn in 2008-09 will be hard-pressed to describe this phenomenon as simple market clearing with price adjustment. Nevertheless, if one does not need realistic price predictions, and is willing to accept the assumption that prices fully clear the market, and transactions costs are negligible, and expectations do not influence prices (in order from strong to heroic assumption), then it is entirely appropriate and useful to use a simple Walrasian auctioneer to call out prices until the market clears.

The alternative of using hedonic regression to predict prices, and to use location choice models with price and budget on the right hand side, is a compelling model design in several respects. First, both models are both theoretically and empirically robust. Second, it is possible to deal with endogeneity between them using appropriate econometric methods. Guevara and Ben-Akiva (2006), for example, deal with the problem of price endogeneity arising from unobserved quality attributes that influence both price and location preferences. They develop a control function approach to correct for this endogeneity. In de Palma et al (2007), a related problem of the bias imposed by availability constraints is addressed. Failure to account for availability constraints of popular alternatives leads to a downward bias in the price effect in the location choice model, and an econometric method is developed to correct for this problem during estimation. While these advances apply to the choice side of the model, there are other advances in hedonic regression such as robust estimation, Bayesian Model Averaging, and Geographically Weighted Regression, that offer improvements in modeling prices. Unfortunately, it is not completely clear that a hedonic regression approach to modeling prices will provide the appropriate systemic responses when coupled with choice models in a simulation system.

Ideally, we would jointly estimate the location choice and price models within a unified framework that is fully internally consistent, and provides micro-foundations for general equilibrium in urban real estate markets, and which can consider expectations and uncertainty in order to deal more realistically with bubbles in the market. One step in this
direction has been recently by revisiting the work on the random bidding model of Ellickson (1981) and Lerman and Kern (1983). The random bidding model jointly estimates the maximum willingness to pay for each type of household, and the probability of each household type being the highest bidder. A Maximum Likelihood estimator for an auction simulator based on this approach of simultaneously estimating the maximum willingness to pay and the category of the winning bidder was developed by Glynn et al (2006). Wang and Waddell (2009) develop a Bayesian estimator that extends this framework to represent the location choice of individual households (or the probability that they participate in an auction) using endogenously estimated choice set filters. This approach mimics the way that households actually engage in when searching for housing, using criteria such as budget, number of bedrooms, and other attributes. This Bayesian estimator adapts the formulation of McCulloch and Rossi (1994) and McCulloch et al (2000). The approach in Wang and Waddell (2009) is a preliminary effort to bring together the concepts of individual discrete choice and price formation, into a jointly estimated model system. Much more remains to be done to fully operationalize this approach efficiently on a full-scale integrated model system, but it is an important step in this direction.

6.2 The Problem of Massive Choice Sets

An additional problem in the context of estimating models of location choice where the number of alternatives is large is how to deal with the choice set construction, since in general it is neither computationally feasible to estimate a model with the full universe of alternatives, nor is it behaviorally plausible. Most applications involving location choice with moderate to large choice sets have adopted a random sampling of alternatives approach, relying on the Independence of Irrelevant Alternatives (IIA) property of multinomial logit models to produce consistent estimates of the parameters (see Ben-Akiva and Lerman, 1987 for a full development). Two important challenges remain relatively unaddressed in the literature, however.

The first challenge is that in any spatial choice model, the assumption of IIA may be difficult to fully support. In most spatial choice situations, there is likely to be correlation in unobserved attributes that are correlated with preferences, and which would violate IIA assumption. If this is the case, then relying on the MNL model for modeling location choice may be problematic. Unfortunately, no other choice model can be shown to produce consistent parameter estimates using sampling of alternatives. Even the nested logit model (NL), which reflects an IIA assumption within each nest, cannot be used with sampling, unless an adjustment is made in scaling the logsums to account for the sampling (see Lee et al, 2009b, for a description of the adjustment and an empirical application using it). Recent work by Guevara and Ben-Akiva (2009) has begun to clarify the path to extending sampling to other non IIA-based models, which offer considerable potential for enriching location choice models using parcel-level detail.

The second remaining challenge, assuming that the sampling challenge is or will soon be satisfactorily solved for at least the GEV family of choice models, is how to deal with the massive choices sets in the application of the models in predictive, or simulation mode. Conventional wisdom (e.g. Ben-Akiva and Lerman, 1987) holds that one should estimate
a choice model such as an MNL with a random sample of alternatives, but then in simulation or predictive mode, enumerate the full set of alternatives in the universal choice set. Unfortunately, for models such as parcel level location choice models, the universal choice set is potentially a million or more in size, for a moderate sized metropolitan area. This would be both behaviorally implausible as an approach (since people cannot actually make this kind of assessment), and is likely to run into problems from a flattening of the probability distribution predicted by the model, since each of the alternatives receives a non-zero probability. The capacity to differentiate good from bad alternatives is potentially diminished then by either of the two approaches: drawing a sample of poor alternatives in constructing a sampling-based simulator, on the one hand, or flattening towards a uniform distribution the choice probabilities if the full set of alternatives are enumerated. It is now becoming clearer that it is possible to apply models using sampling with appropriate adjustments to the scale of the logsum to approximate the universe (Guevara and Ben-Akiva, 2009).

Ways forward on this dilemma include two-stage models that explicitly construct the choice set, or which provide a nesting structure that at least reduces the severity of the probability flattening problem, but with computational cost that is likely to be generally impractical for operational models. Particularly attractive is a nested location choice model with possibly full enumeration of submarkets (geographic with building type) at the top level and building-level choice with sampling of alternatives at the lower level. This approach would also reflect a behavioral process of focusing in on alternatives that are closer to the desired set of attributes, and which are closer substitutes. A step in this direction is taken by Lee and Waddell (2009), which develops a nested model of move and location choice, using the previous residence as a reference alternative and incorporating dummy variables for alternatives within the same geographic area, and a dummy for alternatives within the same type of neighborhood.

7 Integrating Parcel-Level Land Use with Dynamic Travel Modeling

In this section, we turn to the issues arising in using parcels as a unit of analysis in integrated land use and travel models. The description focuses on integration with activity-based demand and dynamic assignment models as is being developed in Pendyala et al (2009).

7.1 Parcel-based Local Accessibility

In using highly disaggregate spatial units of analysis, a major motivation is to be able to measure walking scale accessibility efficiently. These localized accessibility measures can then be used to estimate the impacts of walking-scale accessibility on outcomes such as travel destination choice, residential location choice, business location choice, and even property values. Due to the amount of computation involved in such measurements, it is imperative that the calculations be done efficiently within the simulation environment, rather than trying to use a coupled GIS system to handle these spatial queries. GIS coupling is to date simply not efficient enough for operational use in a spatial simulation environment.
The Open Platform for Urban Simulation (Waddell et al, 2005) has leveraged fast image processing algorithms embedded in the Scipy (2009) Ndimage library for Python to compute distances on a raster surface, as was used in early versions of UrbanSim that used 150 meter gridcells as the basic unit of spatial analysis. This proved to be quite efficient, and allowed computations such as summing retail employment within walking distance to be done in memory, even on fairly large regions containing approximately one million gridcells.

Unfortunately, the use of parcels as the basis for measurement of spatial relationships such as how much retail activity is within walking distance, or how far it is to the nearest bus stop, cannot use the same efficient raster-based algorithms. The irregular geometry of parcels raises complications. If one is concerned about measuring distances from the perimeter of a parcel, the computational cost of measuring proximities within an integrated simulation model are exceedingly high, since these would involve complex geometric buffering and topological queries. The problem can be reduced considerably by using parcel centroids, which simplifies the problem to one of computing Euclidian distances between points. This can be done efficiently and is roughly comparable in speed to grid-based spatial queries. It does leave a measurement problem, in that large parcels are represented by only a centroid, and this will reflect a form of aggregation error when distances are computed to or from the parcel centroid, depending on how buildings are located on the parcel. It is a much smaller problem than is the case with the use of zones or districts, however, and in any case, there is no data source generally available below the parcel (buildings are rarely mapped to locations within a parcel).

The Scipy Spatial library provides good support for doing spatial analysis on point data. Using a KD-tree multi-dimensional indexing algorithm, two types of spatial queries have been added to OPUS to provide parcel centroid-based spatial queries. The first handles queries that require summing objects that are within a specified distance of each parcel, such as computing the total number of jobs within 800 feet, as in Figure 8, for Seattle.
Figure 8: Total Jobs within 800 Feet of Each Parcel within Seattle

This type of spatial query requires traversing a relational data structure that ties businesses (or jobs) to buildings, which are linked to parcels. So the full spatial query returns an index of which parcels are within a specified distance, and then sums the jobs that are within buildings on those parcels.

Figure 9: Relationships of Agents, Buildings, Parcels and Higher-level Geographies

The OPUS Expression Language has been developed to facilitate easy coding of queries on related objects such as those depicted in Figure 9, including a very efficient and
concise means of specifying aggregation and disaggregation of queries (Borning et al, 2008). For example, computing a zone-level average household income, and assigning that average to each building in the simulation for a microsimulation model of residential location choice at the building level is a complex query. It requires connecting households, buildings, parcels and zones in order to compute the average zonal household income, and then requires moving the result back down to the building level, using a reverse navigation of the relational tree. Constructing a query that would do this in a GIS or database environment would require a very complicated SQL statement. Writing custom code to do this would also not be trivial. The OPUS Expression for this query, by contrast, is straightforward and concise:

```
building.disaggregate(zone.aggregate(household.income,
    intermediates=[building, parcel],
    function=mean)
```

The inner portion of this query generates the zonal average household income, and provides information on how to traverse the relationships among the objects. The outer portion of the query disaggregates, or assigns, the resulting values to all of the buildings that fall within each zone. Since these are all simple vector calculations, the computational cost of these types of queries is very small.

A third type of spatial query allows incorporation of spatial datasets such as bus stops, parks or other features, and computes the distance from all parcel centroids to the nearest feature in the target spatial dataset. Figure 4 demonstrates this type of query for the distance to the nearest bus stop from each parcel centroid in Seattle. Setting up these spatial queries requires adding centroid x, y coordinates to the source and target datasets, but is quite efficient. Many variations on the Euclidian distance metrics are available in the Scipy Spatial library, including Manhattan distance, and measures of multidimensional proximity and correlation which could be used for cluster analysis or classification.
Figure 10: Distance to the Nearest Bus Stop from Each parcel in Seattle

The maps in Figures 8 and 10 were generated also by the OPUS simulation platform, using the Mapnik library (Mapnik, 2009) for cartographic display.

7.2 Parcel-based Transit Accessibility

Moving beyond the kinds of localized accessibility measures that can be computed using parcels and relationships as depicted in Figure 3, in order to integrate land use and travel behavior microsimulation, we also need to represent how easy it is for people to move by motorized modes, to destinations that are important to the daily activity schedule, such as work, school, and shopping. Transit is particularly important from a planning perspective to address greenhouse gas emissions, and from a social equity perspective.

Transit accessibility has been difficult to measure well in most travel model applications, since the measures rely on aggregated zones that do a poor job of reflecting the spatial pattern of origins and destinations at a walking scale, and impose coarse representations of the transit network. More useful and realistic measures would not only reflect transit access points such as bus stops, but also the time schedules of transit service, including transfers between routes, and the walk time on both ends of the trip.

An example of this kind of more detailed measurement is developed in Lee (2005). This approach used GIS pre-processing of parcel centroids to bus stops, the local street network, a transit time schedule, and the OPUS infrastructure augmented with the NetworkX Python Library, to construct time-dependent topological graphs of the access of parcels to specific destination types, such as grocery stores. Results, shown in Figure 11, highlight vulnerable neighborhoods that are poor and lack good transit access to basic needs such as groceries.
Figure 11: Transit Travel Time From Parcels to Nearest Grocery Store (Lee, 2005)

This initial application of time-based transit access was too computationally expensive to deploy in an operational simulation, but progress has been made since on ways of addressing the efficiency of this kind of computation.

7.3 **Parcel-based Accessibility and Activity-based Travel Demand**

The choice to walk rather than drive for short trips can potentially be influenced by the spatial configuration of parcels and streets, in addition to the distance to destinations for activities such as grocery shopping and restaurants. The choice between transit and driving for longer-distance trips to work and other activities may also be influenced by the spatial configuration of parcels and local streets, and the accessibility of transit stops. To fully develop integrated microsimulation models of land use and transportation, the representation of walking-scale measures of connection between parcels and the street network will be important. Further, as activity-based travel models continue to advance in research and in practice, there is increasing interest in moving away from zones to represent origins and destinations, and towards parcels and street segments.

The framework for activity-based travel modeling is well suited to capitalize on disaggregate accessibility measures at the parcel and street segment level. Activities occur on parcels, and the opportunity to engage in further activities can be effectively represented by a Time-Space Prism (TSP), measuring the opportunities that are within the time and spatial constraints for an individual, given their current parcel location, and the location and time of their next mandatory activity, as depicted in Figure 12.
7.4 Parcel-based Accessibility, Activity Scheduling and Dynamic Assignment

The tendency towards a fully disaggregate representation of the landscape poses significant opportunities (challenges) for modeling travel behavior. Clearly, parcel-to-parcel travel time skims by mode, purpose and time of day would be neither practical nor desirable. One possible approximation would be to use street segments as the counterpart for zones, and to associate parcels with the street segment they are connected to for access and egress. The problem then becomes smaller in dimension, but still large. If one intends to use a dynamic traffic assignment approach, representing time-dependent skims, potentially by minute, then this accessibility map becomes truly massive. Consider a network of 20,000 links, 5 trip purposes and 5 modes. With a modest time segmentation of 5 time periods per day, and storing only 2 bytes per cell, this results in a data volume of 100 Gigabytes per year, or 3 Terabytes for a 30-year simulation. If dynamic assignment is used with one-minute intervals, the results would be 28.8 Terabytes per year, or 864 Terabytes for a 30-year simulation. By any standards, this would be a very large volume of data, and implausible using current dense matrix approaches to representing origin-destination skims. Sparse representation of the skims would be essential, in the form of an adjacency list with attributes, or in its topological graph representation. Either of these forms would provide a substantially more compact representation of the accessibility map, connecting parcels to street segments, and
traversing the network to reach accessible destinations. Assuming a moderately connected network of the same dimensions as before, with each link connected to two others, the resulting storage requirement would reduce from 864 terabytes for a 30-year simulation to 130 gigabytes for the same simulation. This quantity of data is still very large, but much more plausible given current computing technology, and with some additional modifications, this could be further reduced.

This disaggregate approach to measuring accessibility using a connected graph of parcels and streets is currently being developed in a new integrated land use, activity-based travel, and dynamic assignment model system (Pendyala et al., 2009). In this model design, activities schedules are updated each minute, and as activities are committed that require travel on the network, the person and vehicle trips are packaged and submitted to a dynamic assignment model, which then releases the vehicle trips to the network on six second intervals and models their movement on the network using a mesoscopic approach with a Wardrop equilibrium route choice algorithm. Once vehicles arrive at their destinations, the vehicle trips (for cars) and person trips (for transit) are sent back to the activity-based travel demand model for agents to update their arrival times and determine whether to engage in further activities within the current time-space prism. This approach is a significant departure from other activity-based travel modeling efforts that produce a full daily travel demand and subsequently aggregate the results for traffic assignment. By contrast, this approach integrates tightly with dynamic assignment on a minute by minute basis, making within-day activity adjustments based on changing travel conditions a natural outcome. A diagram of this interaction is shown in Figure 13.

Figure 13: Interaction of Activity-Based Travel Demand and Dynamic Assignment (Pendyala et al., 2009)
8 Conclusion: Implications for Computational Algorithms and Procedures

In this paper, we have made a case for pursuing disaggregated models to their logical limit spatially: the parcel and building level. The case is compelling on many levels, and opens significant potential for focusing on interactions that occur on the scale of the pedestrian, a scale that has heretofore been broadly ignored in integrated models, with an inevitable consequence that such models are of little help in addressing policies and behaviors that are very at the parcel and walking scale.

We return now to the cautions raised in the early part of the paper, as these have direct implications for developing software frameworks. Computational performance and inability to validate integrated microsimulation models due to stochastic variation and instability, were raised as very legitimate concerns.

8.1 Computational Performance

Run times that are excessive make models useless from a practical perspective, and one could also argue, from a behavioral perspective, since they prevent suitable refinement of the models through proper calibration and validation, as noted by Wagner and Wegener (2007) and Weidner et al (2009). What is the source of poor computational performance in microsimulation models? The choice of language? Java and Python are interpreted languages, and known to be slower than compiled languages such as C, C++, and Fortran. But the differences begin to blur when one considers that languages like Python use lower level languages to implement computationally expensive tasks and to replace loops with vector calculations.

To make the case that language is not the impediment, we use two examples. An UrbanSim application developed for the San Antonio metropolitan area in Texas, was developed using a zonal level of detail for location choices. It is a microsimulation model with a full enumeration of individual households and jobs, but the choice of locations used buildings located within zones. The full model system was run, including household and employment transition models to reconcile micro-level agents with macroeconomic totals, relocation models for households and jobs, location choice models for households and jobs, real estate price models for buildings, and real estate development models to simulate new development. The model system is implemented in Python using the Open Platform for Urban Simulation (Waddell et al, 2005). Run times per year on the fully estimated model system range from 14 seconds to 30 seconds on a Macbook Pro laptop.

Adding parcel detail to the model system does significantly slow the model system, as it is currently implemented. A San Francisco parcel level model application runs at approximately 7 to 10 minutes per simulation year. On a relatively large metropolitan area such as the Puget Sound, the run times for a parcel-level model system are 30 to 40 minutes per simulation year. There is definitely room for improvement in these run times, particularly considering that they are run on standard desktop or laptop computers, and on only one processor. Parallelization of these models is straightforward, due to many computations on sets of agents that do not directly influence each other.
In the proposed model integration of Pendyala et al (2009), heavy use of parallel computation is already implemented in portions of the model system, and will be propagated to the remainder of the model system to take advantage of hardware configurations that range from multi-core CPUs that are the current standards on desktop and laptop computers, and to be able to scale up to cluster computers, grids, or run on cloud computing configurations.

8.2 Instability and Stochastic Variation

The concerns raised by Wagner (2009) that integrated microsimulation models are likely to be unstable and suffer from excessive stochasticity, to the point that random variation overshadows systematic responses, is particularly troubling. Faster computation will not solve this problem. It is difficult to diagnose a model system that is large and complex, and this may be particularly a challenge for some recent models that have reached immense levels of complexity.

One general recommendation is to design the model system with careful attention to feedback between behaviors across the model system, and avoid a tendency to accumulate models without assessing how their behavior interacts with the rest of the system. The interdependencies among models in a complex model system that is built around interacting model components are inherently complex, likely to be highly non-linear, and may be subject to higher-order interactions that are no anticipated by the model builders. We need to think of agents (persons, households, firms) as being in a constantly adapting state of dynamic equilibration, in which they sense changes in their environment, and react by making adaptations to restore imbalances. A dynamically adaptive agent is more likely to produce realistic behavioral responses to environmental changes, whether induced by policy or by economic or demographic or technological changes.

Finally, we need to take uncertainty in models seriously. If our models have high levels of embedded uncertainty arising from uncertainty in input data, in model parameters, and even in model specifications and structures, then we should be able to analyze and even calibrate the uncertainty in the models. Sevcikova et al (2007) provide a method, using Bayesian Melding (BM), to do just this. The BM approach requires observed data (which may be incomplete or aggregate) over a historical period, for calibrating uncertainty over the initial phase of the period, and then for validating the calibrated model over the subsequent phase. A properly calibrated model will then be suitable to generate probabilistic inferences that incorporate any inherent stochasticity in the model system, and provide much more principled inferences than the usual point estimates produced by models. In closing, there is good cause for optimism that we can build models that are both behaviorally valid, and can be used to inform policy deliberations in principled ways that encapsulate to the maximum extent possible, all that we know, and all that we do not know.

Acknowledgments

We wish to acknowledge the generous support of FHWA Grant BAA DTFH61-07-R-00117, NSF grants IIS-0705898 and IIS-0705898.
References


Guevara, Cristian, and Moshe Ben-Akiva (2009) Evaluating methodologies to address non-iid errors in discrete choice models of residential choice. International Association of Travel Behavior Research, Jaipur, India.


Lee, Brian and Paul Waddell (2009) Incorporating residential mobility into residence location choice model in a microsimulation context. International Association of Travel Behavior Research, Jaipur, India.


