Modeling Urban Form in City Simulations

Natalie Wiseman

A Thesis

In

The Department

Of

Geography, Planning and Environment

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science (Geography, Urban and Environmental Studies)

at

Concordia University

Montreal, Quebec, Canada

April 2016

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Natalie Wiseman

Entitled: Modeling urban form in city simulations

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Geography, Urban and Environmental Studies)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________________ Chair
Monica Mulrennan

______________________________________ Internal Examiner
Jochen Jaeger

______________________________________ External Examiner
Yann Roche

______________________________________ Supervisor
Zachary Patterson

Approved by ____
Chair of Department or Graduate Program Director

____
Dean of Faculty

Date ____

Abstract

Modeling Urban Form in City Simulations

Natalie Wiseman

One way of planning for a region's future transportation infrastructure and capacity needs is to forecast travel demand. Integrated land use-transportation modeling tools do this in a way that is sensitive to how the performance of a transportation system affects peoples' decisions of where to live and firms' decisions of where to locate. As such, they are helpful tools for analyzing different transportation and land use policy scenarios. The transportation system is a factor in how new urban areas develop. Policies attempt to regulate dispersed urban development, known as urban sprawl, however integrated modeling frameworks can only evaluate those policies that affect the extent of non-urban land legislated as developable, or those within urban areas.

Incorporating other policies relating to sprawl into integrated models is limited by their ability to represent geometric changes to the landscape; those associated with the transition of non-urban land to residential use. Building on current methods for representing geometric landscape changes, this thesis is about models and algorithms for representing the specific forms these changes can take. There are a number of algorithms, taking distinct approaches to subdividing blocks into parcels and generating roads, suggesting different algorithms are better for generating different forms. There is little guidance on when to use which algorithm, potentially resulting in sub-optimal geometric representation of future urban areas.

This thesis outlines a process for representing the spatial distribution of urban form in future urban areas within integrated models. To this end, it has estimated a model for predicting the spatial distribution of road network patterns in future residential neighborhoods and identified the block subdivision algorithm most suited to subdividing each of the possible road network patterns. Results can serve as guidelines for deciding which algorithms to use on which road network types. They also present a possible way of estimating the type of future road network in a local area as a function of slope of terrain, period of development, proximity to a river and adjacency to a road network of

the same type, among others. This knowledge could help improve the accuracy of population predictions and potentially be implemented within the modeling process of integrated models once these are better able to represent geometric changes to the landscape.

Dedicated to

Ola

Acknowledgements

I am most thankful to my supervisor Zachary Patterson for his constant encouragement and support with this research. Thanks to the people of the TRIP lab for making this experience fun and collaborative.

Thank you to Craig Townsend, whose comments on an early draft of this thesis improved it very much, and to Jochen Jaeger for his helpful suggestions on the proposal. I am grateful to Yann Roche for imparting a curiosity and healthy skepticism about the potential of Geographic Information Systems.

I am especially grateful to my grandmother for reading through my entire proposal, though she understood very little.

I am privileged to have had funding from the Agence métropolitaine de transport, the Canadian Foundation for Innovation and the Canadian Research Chairs program as well as a Concordia Merit Scholarship.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Acronyms

CGA (Computer Generated Architecture)

DRAM (Disaggregate Residential Allocation Model)

EMPAL (Employment Allocation Model)

GO (Grid Overlay)

GPD1 (Generalized Parcel Divider 1)

ITLUP (Integrated Transportation and Land Use Package)

MAUP (Modifiable Areal Unit Problem)

MRC (Municipalité régionale de comté)

MUSSA (Modelo de Uso de Suelo de SAntiago)

NYMTC-LUM (New York Metropolitan Transit Commission Land Use Model)

OBB (Oriented Bounding Box)

PECAS (Production Exchange Consumption Allocation System)

SI (shape index)

SS (Straight Skeleton)

TAZ (Traffic Analysis Zone)

WP (willingness to pay)

Introduction

Providing transportation infrastructure to meet growing demand is a key concern of regional planning agencies as well as government and transit organizations. Sustainable planning can provide usable alternatives to car ridership and encourage a shift from automobile to transit modes thereby reducing congestion on highways and greenhouse gas emissions. Transportation forecasting is the process of estimating the number of vehicles or people that will use a given transportation facility in the future and helps determine demand for increased transportation capacity and infrastructure. Traditionally, this has been done using a 4-stage travel demand model that takes exogenous predictions about the spatial distribution of population and jobs as inputs, and transforms these into trips assigned to routes along a network.

However, travel demand models can't predict the effect that the performance of the transportation network has on future distributions of population and jobs. For example, increased congestion to and from a specific area might deter individuals or firms from locating there. Conversely, new transportation infrastructure in a certain area might provide incentives for people and firms to locate in its vicinity. In this way, the new infrastructure induces demand, leading more people to utilize it than the models anticipated. A traditional 4-stage travel demand model, that doesn't account for the effect of transportation network performance on population and job distributions, will tend to overestimate the congestion relief benefits of new infrastructure by underestimating additional travel time and congestion associated with induced demand (Borning et al., 2008; Waddell et al., 2007).

Though not always part of travel demand models, the feedback loop between transportation infrastructure and land use has long been recognized. Vancouver's 1975 integrated transport and land use plan (GVRD, 1975) tapped into this feedback by planning lively and diverse centers away from downtown and providing efficient transit services to them. The pockets of activity would attract ridership that would relieve the burden on governments to subsidize this transit. In 1989, Newman and Kenworthy proved the link between low population density and gasoline consumption, validating future planning based on this (Newman & Kenworthy, 1989). In 1994, the Regional Alliance for Transit (RAFT) criticized the land use transportation modelling system of the Metropolitan Transportation Commission (MTC) in California (Lewis, 1998). RAFT pointed out the modelling system assumed transportation demand was fixed and couldn't affect land use and that the only other alternative to the one being proposed was a do-nothing scenario. The group argued that under these conditions, any proposed project would seem desirable. RAFT created an alternative scenario to compare against the proposed one that included, among other things, a shift in land use from low density dispersed development to denser development with mixed uses around transit stops. While the modelling outcomes favoured the RAFT scenario, it became clear that the policies they proposed had outgrown the modelling system's ability to fully evaluate their effect on travel demand.

Meanwhile, since the 1960s, integrated modelling tools were being developed to capture these feedbacks between transportation and land use not captured by traditional 4-stage models. The first integrated modeling systems weren't useful as policy analysis tools because the computational capabilities, modelling methods and available data of the time weren't advanced enough to support their complexity (Miller 2003). However, as computing grew in sophistication, it became capable of supporting integrated models to the point where they became useful in policy analysis and planning (Miller 2003). By this time, integrated modeling systems had been developing incrementally for over three decades, which resulted in a diversity of approaches that were capable of being implemented.

There are currently many different integrated modeling systems with real-world implementations as described later in the Introduction. First, a description of a famous legal case involving the Legacy Parkway project in Utah is given. While it refers to one particular planning project in which one particular integrated modeling system was used, it serves as an example of how integrated modelling had drastically changed the field of transportation planning more generally. This case is described by Adler (2006) and Waddell et al. (2007).

The outcome of the case was in large part determined by an integrated land use-transportation demand modeling system that combined the traditional travel demand method with a land use-forecasting component, thereby accounting for the feedbacks between the two. Mandated with meeting projected demand for increased auto ridership, in 1996 the government of Utah proposed a mega highway project that would be built over the Great Salt Lake wetland, a legally protected critical area home to many ecologically sensitive animal species. The Sierra Club, UBET (Utahns for Better Transportation) and the Mayor of Salt Lake City took the state to court under the National Environmental Policy Act (NEPA), Clean Water Act and Clean Air Act, which found they had failed to consider alternatives to the proposed project. In the context of this case, and in anticipation of similar cases in the future, Paul Waddell at the University of Washington developed an UrbanSim application for the Greater Wasatch Area, in Utah. UrbanSim is an integrated land use-transportation model that accounts for the feedbacks between the performance of the current transportation network and future transportation demand. This modelling tool demonstrated that the Legacy Parkway project wouldn't solve Utah's congestion problems, but would likely worsen them by dispersing the regional population even farther. The model also directed its users toward an alternative proposal that would better meet future demand. The settlement resulted in a much narrower parkway, more integrated with the existing transportation network and with meanders to minimize impacts on wetland (Adler, 2006). As part of the settlement, the Wasatch Front Regional Council was mandated to use UrbanSim in its future transportation planning activities, including transportation demand forecasting and developing a long-range transportation plan (Waddell et al., 2007).

UrbanSim's ability to account for transportation system performance in population and employment projections is shared by a number of similar systems that form the basis for modern integrated land-use transportation modelling. A number of integrated modelling frameworks exist such as ITLUP, MEPLAN, TRANUS, MUSSA, NYMTC-LUM and PECAS and will be discussed in detail in the literature review. These vary in their approaches to different aspects of the modeling process, such as the scale at which space is represented, the way markets are represented, the way trip distribution is estimated, and the extent to which behaviours of individuals can vary (Clay, 2010; Hunt et al., 2005; Miller, 2003).

Trends in integrated modelling

Within the field of integrated modelling, there has been a trend toward representing phenomena at increasingly smaller spatial scales. This has been made possible by technological developments in Geographic Information Systems and Computer Science that make complex computations on this data possible. Increasingly fine spatial representation has also been motivated by understandings in a number of problem areas such as organized complexity, rational choice theory and statistical aggregation bias. Organized complexity applies to problems that deal with a large number of factors that are interrelated into an organic whole and whose interrelationships are correlated rather than random (Weaver, 1948). Rational choice theory is a framework for understanding and modeling social and economic behavior of individuals (McFadden, 1980). Statistical aggregation bias refers to the tendency for the spatial units chosen for the study of a phenomenon to affect the results of the study (Waller & Gotway, 2004). This increasingly disaggregated spatial representation presents many opportunities but also many challenges within the field of integrated modelling.

Integrated modelling, city generation engines & block subdivision algorithms

There is a functional disconnect between the fields of geometric and behavioural modelling that stems from the fact that the two have developed to serve different purposes. Vanegas et al. (2009b) have clarified this problem as part of an effort to develop a simulation tool that spans the two fields. Behavioural simulation refers to modelling long-term behaviours in an urban space based on socioeconomic factors (Vanegas et al., 2009b). This modelling helps to inform decision-making regarding policies in urban areas, both present and future. In contrast, geometric modelling, or simulating changes in the structure or character of a space, has been undertaken in the fields of computer graphics for developing videogames and more recently for natural resource management (Kelly & McCabe, 2007; Parish & Muller, 2001; Vanegas et al., 2010; Weber et al., 2009). In consequence, frameworks for modelling these two types of processes have taken on very different forms. For example, utility or ‘happiness’-maximizing models are used to predict behaviour that leads to mobility. In contrast, rule based heuristics and neighbourhood analyses have been used to model changes to the physical character of space (Xie & Batty, 2003).

In addition to the differing functionalities of the two modelling frameworks, there are also computational limits to their integration. Behavioural simulation tools generally assume fixed two-dimensional geometric features. Dynamically computing these geometric features would require integration with geometric or spatial simulation tools. However, this integration is limited by the magnitude of dynamic spatial computations that behavioural simulations would require. These are too great for most current spatial simulation frameworks to compute efficiently. According to Vanegas et al. (2009b) there is a disconnect between the two fields, "…because behavioural modelling is not concerned with generating the geometry of the urban space and because geometric modelling usually does not consider behavioural properties of real-world cities." This disconnect concerns the field of integrated modelling, which sets out to model the dynamics between the transportation system and the way land is used for the following reason: what happens when the behavioural processes that take place within a space change the structure or boundaries of the space itself?

This is the type of dynamic that occurs with urban sprawl. Demand for housing grows along the rural urban fringe, increasing land values and putting pressure on farmers to sell their large tracts of agricultural land to developers (Irwin et al., 2003; Newburn & Berck, 2006; Pocewicz et al., 2007). The developers then subdivide the land and may build houses on the new subdivisions. This changes the population density, amount of green space and other variables that the models then use in making new predictions. By not modelling these processes explicitly, simulation tools preclude assessing policies that relate to them and introduce bias in model predictions (Schirmer, 2010).

Integrated modelling has experienced a sort of disconnect with respect to representing this kind of land cover change, namely, the transition from agricultural land to residential use. This is important since major population growth occurs at the rural-urban fringe, and urban sprawl is a serious environmental concern that is closely linked to the transportation system. This type of land use change not only involves low density population growth and building construction at the fringe, things that integrated models simulate, but it also necessarily involves major restructuring of space and transfers of land ownership. Integrated modelling frameworks can interpret or 'read' data that represents spatial features, such as traffic analysis zones and, less commonly, parcels, since these are the units based on which behavioural data and the flow of travel is represented (UrbanSim, PECAS, MUSSA). However, only the PECAS system can make changes to these spatial objects in response to, say, predicted increases in housing demand along the rural urban fringe, that would lead to new housing developments but first to the subdivision of agricultural land into residential lots (Hunt & Abraham, 2007). This would require not only 'reading' but also 'writing' new spatial data, namely, parcels. Integrated modelling frameworks and Geographic Information Systems have been developed to perform different functions, and so fully integrating the two such that each can read from and dynamically compute an input to the other, which is what is required to accurately simulate geometric changes to the landscape, is in its beginning stages (Schirmer, 2010; Vanegas et al., 2010; Waddell, 2009). To date, the only operational integrated land use transportation model that has this capability is PECAS.

Visualizing changes in space is something that Geographic Information Systems do well. And so simulating and visualizing land cover changes is something that has been undertaken extensively in this field within Cellular Automata models, Markov Chain Transition matrices and City Generation Engines, discussed in the literature review. However, simulating the decision-making processes and complex dynamics that lead to these changes is something that integrated modelling frameworks were conceived to do (Vanegas et al., 2009b; Xie & Batty, 2003). In fact it is more than a little surprising that more of these systems don't explicitly represent processes relating to urban sprawl and is a testament to the extent of the disconnect the field has experienced.

The main contribution to integrating behavioural and geometric modelling within the field of GIS includes City Generation Engines. City Generation Engines can read the forecasts of integrated land use transportation models and structure this data on population, jobs and buildings into plausible 3D cities. They also support manual editing of an urban system, and update the rest of the elements to maintain plausibility and internal consistency (ESRI; Synthicity; Vanegas et al., 2009b). These engines use automatic parcel subdivision and road generation algorithms to subdivide large blocks into smaller parcels to accommodate growing population and jobs there in a spatially realistic way. One contribution to spanning the behavioural-geometric disconnect within the field of integrated modeling includes the idea of modeling spatial arrangements of urban objects within a behavioural modeling framework, namely the multinomial logit model, to predict building types among other spatial characteristics of objects. Multinomial logit models are used in modeling discrete dependent variables that have multiple outcomes (Schirmer, 2010). Recently, an integrated modeling framework called PECAS was developed that models behaviours leading to land use change explicitly and the resulting parcel subdivision processes this implies (Hunt & Abraham, 2007). It also uses an automatic parcel subdivision algorithm to do this.

This thesis contributes to spanning the behavioural-geometric disconnect from the GIS side of the problem. It deals with automatic generation of both the parcel patterns within street blocks and the road networks that shape these blocks. Such processes are useful for representing the transition of non-urban land to residential use that often manifests in urban sprawl.

Rather than write new algorithms for generating landscape features, the work undertaken here aims to make best use of algorithms that have already been developed to capture how different urban forms are distributed in space. In doing so, three different block subdivision algorithms are compared to see if certain ones perform better overall, or on specific block types. In addition, a model is developed for estimating which road network, or block type, is likely to appear where.

Since integrated modeling is heading in the direction of increasingly fine spatial representation, explicit representations of urban sprawl will be possible. This is likely to make integrated land use-transportation modelling more directly relevant in evaluating parcel-level policies relating to urban sprawl. And sprawl would be represented more accurately with the understanding of which block subdivision algorithms work better for subdividing which block shapes and which of these block shapes is likely to appear where. More realistic depictions of urban form could also lead to more accurate population and demographic forecasts in this context.

Moreover, missing data, and especially missing parcel data, is a common barrier to implementing parcel-level integrated land-use transportation models and is another application to which block-subdivision algorithms can be put (Patterson & Bierlaire, 2010). With guidance on which algorithm to use for which block type, users may be better equipped to complete their baseyear parcel datasets in as realistic a way possible.

Literature Review

Integrated Land Use-Transportation Modelling Frameworks

Integrated land use transportation models exist within a variety of frameworks that approach the different dimensions of the modeling process in a variety of ways. A number of articles reviewing these different integrated land use-transportation modelling systems have been written. Iacono, Levinson and El-Geneidy (2008) discuss the different types of modelling frameworks and provide operational examples of each. Wegener (2004) reviews recent developments in operational integrated land-use transportation models focusing on their ability to test land use and transportation policies and assess their impacts. Hunt, Kriger and Miller (2005) conducted a review of the six most widely used frameworks. The following summary of their review gives a sense of the relative benefits and drawbacks of each framework's approach as well as the trends within the field. Another framework called PECAS has since been developed and will also be reviewed here.

ITLUP (Integrated Transportation and Land Use Package, also known as DRAM/EMPAL) has been the most widely used framework in the US, in development since the 1960s by Professor Stephen Putnam at the University of Pennsylvania (Putnam, 1996). It is a fully integrated model, containing a variety of sub-models each performing a different step of the traditional 4-stage travel demand models. The sub-models DRAM (Disaggregate Residential Allocation Model) and EMPAL (Employment Allocation Model) predict household and employment location. In addition, the DRAM model also uses the results to estimate trip generation (number of trips) and trip distribution (pairing origins and destinations). A multinomial logit modal split sub-model (to predict preference for certain modes) and a trip assignment sub-model are used to assign trips to routes along a transportation network. Usually households and jobs are classified into four types, but more classifications are possible. The system represents space using large zones, with associated spatially aggregated data. Prices and land markets are absent from the model. The data required by DRAM/EMPAL is modest and readily available, making it an easily implemented framework.

MEPLAN was developed in the UK by a private consulting firm called Marcial Echenique and Partners Ltd (Hunt, 1994). It has been implemented in over 25 regions throughout the world. Space is represented in the model as large zones, to which quantities of households and economic activities are allocated. Production (supply) is allocated to zones to balance consumption (demand) using discrete choice models, in which production prices are a variable. Interactions among these economic activities between zones give rise to trips and transportation demand. Degrees of interactions, or economic flows, are represented as interdependencies between different branches of a regional economy using metrics called technical coefficients. These are defined as standardized inputs per dollar of output for each industry, structured in a form called an input-output matrix. The system solves for equilibrium market conditions at each time step, and prices for consuming space are set such that supply equals demand. Though equilibrium prices aren't realistic, they are helpful in estimating actual prices. Travel demand arising from economic flows is assigned to a network using logit functions that estimate mode and route choice, accounting for congestion. Resulting transport disutilities become inputs to the mode and route choice logit functions in the next time period. This system is behaviourally and spatially aggregate.

TRANUS is proprietary software developed by Modelistica in Venezuela (Modelistica, 1995). It is very similar to MEPLAN but with a more restricted set of functional forms and modelling options within the framework. It has been applied in Central and South America, Europe and some parts of the US, namely, Sacramento, Baltimore and Oregon.

MUSSA (Modelo de Uso de Suelo de SAntiago) was developed by Professor Francisco Martínez as an operational model of urban land and floor space markets for Santiago, Chile (Martínez, 1996). MUSSA is another equilibrium model in which building supply is balanced with demand but in which both buyers and sellers are profit-maximizing agents. Developers supply building stock to meet demand derived from consumers' willingness to pay (WP) for the buildings (for both households and firms). The building supply and utilities consumers derive from their buildings adjust iteratively to each other until equilibrium is established. The model represents land at the spatial unit of the traffic analysis zone: a variably sized unit analogous to the Census Tract. The model can accommodate fully disaggregated household data. Furthermore, the land use component is connected to the travel demand model in a modular fashion rather than being fully integrated.

NYMTC-LUM (New York Metropolitan Transit Commission Land Use Model) was developed by Professor Alex Anas for New York , but has also been used in Chicago and Illinois (Anas & Arnott, 1993). Also based on microeconomic theory, it models interactions between residential housing, commercial floor space, labour and other travel demand. Demand and supply processes are defined for each type of interaction. The model determines housing prices, floor space rents and workers' wages such that demand and supply balance in their respective markets. The model uses traffic analysis zones as the spatial unit of analysis. It is also behaviourally aggregated, meaning households, jobs and buildings generally have uniform characteristics. The land use component is also connected to the travel demand model in a modular fashion rather than being fully integrated.

As previously mentioned, UrbanSim was developed by Paul Waddell at the University of Washington. It is currently being used in the United States in Hawaii, Oregon, Utah, Washington, New Mexico, Houston, Phoenix, San Francisco, and California, among other places (Borning & Davis, 2005; Lee et al., 2015; Waddell, 2010; Waddell et al., 1998). In Europe, UrbanSim has been applied in Paris, Lyon, Brussels, Zurich, Lausanne, Amsterdam, Tel Aviv and Youngsan-gu, Korea (Borning & Davis, 2005; De Palma et al., 2014; Hassan et al., 2010; Kryvobokov et al., 2013; Patterson & Bierlaire, 2010; Schirmer, 2010). UrbanSim is another system that is based on microeconomic theory, but in contrast to MUSSA and NYMTC-LUM, it doesn’t assume equilibrium market conditions. Both buyers and sellers are profit-maximizing agents and demand is estimated based on willingness to pay (WP). Market clearing, the process in real markets by which prices adjust in such a way that brings supply closer to demand, occurs at the less aggregated level of the traffic analysis zone and property type. In a given year, quantities of land are developed based on expected profits (expected revenue less costs). This expected revenue is based on the previous year's prices and the land is only available for occupancy in the following year. Demand is based on the previous year's prices and on current supply. This lag in response time between price adjustment and new building availability creates a dynamic equilibrium situation in which supply is constantly adjusting to be equal to demand, but never quite reaches that point. In consequence, the model end state is path dependent and requires a solution for intermediate years. Demand is predicted in the model for the spatial resolution of the traffic analysis zone. On the supply side, the model uses the individual land parcel as the unit of land development and redevelopment. The model can be used with behaviourally disaggregate data on both the supply and demand sides. The framework permits the analysis of policy scenarios that include comprehensive land-use plans, urban growth boundaries, density thresholds, mixed-use development, redevelopment, environmental restrictions on development, and development pricing policies, in addition to the transportation infrastructure pricing policies handled by the linked travel demand models.

PECAS (Production Exchange Consumption Allocation System) has been implemented in several areas of the US, namely, Sacramento, San Diego, California and Baltimore. It is also under continual development in Calgary, Oregon and Atlanta (Clay, 2010). The framework uses an input-output table similar to MEPLAN to account for amounts of production (supply) and consumption (demand) in each sector in each zone. This in turn influences where households and other activities are allocated at an aggregate scale (Activity Allocation module), which in turn informs the development of land from one time period to the next in a microsimulation model of developer behaviour (Spatial Development module). The exchange zones simulate equilibrium markets, adjusting the exchange prices in the exchange locations until all markets clear. Travel demand emerges from the movement of commodities between zones and is informed by technical coefficients that determine dependencies between industries for goods and services. On the supply side, space is represented at the parcel level. A unique feature of PECAS is that it models land use change explicitly, and spatially represents the parcel subdivision processes it implies (Hunt & Abraham, 2007). As such, it has overcome the behavioural-geometric disconnect in integrated modeling. How it does this will be discussed in a later section. However, its structure is complex and difficult to implement. Its demand side is less behaviourally and spatially detailed than its supply side and it assumes equilibrium market conditions.

Several distinctions in the above models are worth further discussion. These include representation of markets and assumptions of equilibrium as well as the levels of spatial and behavioural representation. Whether or not models assume equilibrium depends on the way they represent markets. Models that assume supply is balanced with demand are solved for equilibrium in a future year of interest, or the solutions are updated at 5-year intervals. This is the case for ITLUP, MUSSA, NYMTC-LUM, MEPLAN and TRANUS. While PECAS assumes equilibrium, it runs at 1-year time steps to account for information lags from the previous year. The only exception to the equilibrium of markets assumption is UrbanSim, which runs at 1-year time steps to simulate the system's dynamic rather than equilibrium state.

Another variation in the frameworks is the extent to which variation in behaviours can be represented. At one end, behaviourally aggregate models classify households, employment, developers, and government actors into types and the number of different types defines the degree of behavioural disaggregation (ITLUP, MEPLAN, TRANUS, MUSSA, NYMTC-LUM). At the other extreme, agent-based models can represent the set of households and employers within a region (when available) so that no aggregation need occur at all (UrbanSim). PECAS represents disaggregate behaviours on the supply side but aggregate behaviours on the demand side.

The degree to which the input data is aggregated (or disaggregated) in the models is important since it determines which processes the models are able to capture and the extent of aggregation bias in their results. Behavioural units are aggregated to the basic unit of space represented in the model. Different spatial resolutions can represent processes of land development (supply side) and land consumption (demand side). ITLUP, MEPLAN and TRANUS use large zones to represent land available for development and consumption. MUSSA and NYMTC-LUM use smaller traffic analysis zones, units similar in size to census tracts. UrbanSim can represent both the supply side and demand side of market processes at the parcel level. As such, it is capable of agent-based microsimulation: of capturing processes at the behavioural and spatial scales at which they actually occur. PECAS is also capable of microsimulation, but only on the supply side. Unlike UrbanSim, it can represent processes that change the structure of the parcel fabric such as land use change and urban sprawl.

This review of the most widely used integrated modeling frameworks indicates there is a trend toward increasingly fine levels of behavioural and spatial representation. As a result, more recently developed integrated models use parcels. While the components of a true microsimulation model exist, these are spread between different frameworks. UrbanSim models markets and behaviours at a spatially disaggregate level while PECAS models land use change at the parcel level. Parcel data is a necessary input to microsimulation models, and is fundamental to its theoretical framework.

Why parcels matter

By definition, a parcel is a geographic unit that is taxed, and the basic unit on which property taxes are assessed. They are increasingly inventoried, mapped and monitored by local agencies charged with property tax assessment (Waddell, 2009). As a result, there is incentive for governments to keep them up to date. The increasing availability of parcel level data and the increasingly computationally efficient technologies for manipulating it make it compelling to use in integrated modeling frameworks. More importantly, using parcel level data is consistent with understandings in complexity theory, behavioural theory and statistics (Waller & Gotway, 2004; Xie & Batty, 2003).

Explicit representation of parcels and their associated properties, such as population, jobs and households solves the MAUP (Modifiable Areal Unit Problem) , which introduces statistical bias and occurs when analyzing data aggregated to larger spatial scales. The MAUP occurs because the phenomena we try to analyze are continuous, but we represent them as occurring within socially constructed geographic units, such as Census Tracts. This problem is evident in that “conclusions based on data aggregated to a particular set of districts may change if one aggregates the same underlying data to a different set of districts” (Waller & Gotway, 2004). This problem is evident in conscious attempts to alter electoral districts to influence election results in favor of a political party (Griffith, 1907). Similarly, different statistical inferences are drawn from the same data set that is aggregated to different spatial resolutions. Given that many policy decisions are informed by statistical inferences derived from aggregate data, researchers are wary of this problem. It has been argued that the only real solution to the MAUP is to use individual-level data that explicitly represents each object of analysis

(Weeks, 2004). These data units are related to their spatial units, usually land use parcels individuals actually occupy.

Integrated modeling is increasingly representing individual level data in this way and in so doing, avoiding this source of bias in its predictions. Integrated modeling also has a history of representing diverse behaviours arising from individuals with a diversity of characteristics. This has its roots in Discrete Choice Analysis (Ben-Akiva & Lerman, 1987) and the field of Random Utility Theory it draws from. Discrete Choice Analysis models the utility, or happiness, an individual derives from a given option in light of all the other options available to them. In these models, the characteristics of the individuals and the aspects of the alternatives relevant to their utilities vary. In this way, integrated modelling frameworks can represent individuals' choices of where to live and firms' choices of where to locate among other behaviours. Representing both spatially and behavourally disaggregate data, as the field of integrated modeling is increasingly doing, makes fundamentally different forms of modeling possible such as microsimulation and agent-based modeling.

Microsimulation is a type of computerized analytical tool that models the processes in systems at the scale at which they occur. The behavioural counterpart to microsimulation is agent based modeling (ABM), a class of computational models for simulating the actions and interactions of autonomous agents (both individual or collective entities such as households or firms), with a view to assessing their effects on the system as a whole. The idea is that complex system dynamics emerge from the cumulative interactions between individual agents' behaviours and decision-making processes (Xie & Batty, 2003). This form of representation is compelling when modelling urban dynamics, which are typically highly complex and difficult to model by looking for overarching trends (Pocewicz et al., 2007; Xie & Batty, 2003).

The benefits of microsimulation in integrated models are apparent when one recognizes that land use regulations apply directly to parcels and buildings and that local policies make use of these regulations. Taken together local communities shape land use on a large scale. It therefore makes sense to use geographic units of analysis relevant to those policies to model land use.

Parcels are also important in the context of integrated modelling because they represent the locations of activities. Integrated models can use parcel-level data to represent the walking-scale of activity and to calculate walking-scale accessibility measures (Waddell, 2009). For example, accessibility measures calculated at the parcel level can determine whether transit stops or grocery stores are within walking distance of parcels, and might substitute walking trips for those by car. This can reveal new things when used in estimating models, for example, that accessibility by walking to retail is important in household location choices. This can tell us how to make places more amenable to walking and transit use.

Thus advances in technology, rational choice theory, understandings of system dynamics and resulting advances in modeling frameworks have been taken up by the field of integrated modeling with the aim of predicting transportation demand more accurately and providing services that better meet the needs of citizens and governments. Still, barriers exist to implementing integrated models within fully spatially and behaviourally disaggregate microsimulation frameworks.

An alternative to the parcel for representing local level processes, and the most common unit of analysis for micro level land use models, is the grid cell. Most land cover change models, namely Cellular Automata and Markov Chain Transition matrices, model land use at the level of the grid cell. This method uses a uniform grid, with each cell storing a different value representative of the area it covers (Alexandridis & Pijanowski, 2007; Wickramasuriya et al., 2013; Xie & Batty, 2003). Multiple superimposed grids can represent different variables of the landscape. The spatial resolution of the grid cells represents the smallest area on the earth capable of being represented, usually between 30 to 150 meters. The grid cell uses a raster data structure, which is extremely computationally efficient, because each cell stores a value coded in a number of bytes that are directly interpreted by the computer processor (Waddell, 2009). As an example, digital images are stored within raster datasets, wherein each cell represents a different color value.

For example, Markov Chain Transition matrices (Pocewicz et al., 2007) are used to estimate the probabilities with which a given land use type changes to another for a given grid cell. These are trend-based and don’t incorporate individual behaviour explicitly. Similarly, Cellular Automata models can attribute simple rule-based behaviours to cells that define the way they interact with cells in their vicinity. These frameworks only allow for modelling simple behaviours between a given cell (or object/agent) and cells in its vicinity (Xie & Batty, 2003). Such behavioural patterns are more characteristic of physical features such as land cover, rather than moving human agents. This makes them poorly adapted for simulating movement through a transportation network, and in consequence trip generation and distribution in cities. As a consequence, these systems aren't used in policy analysis.

A rare exception is a study by Alexandridis and Pijanowski (2007), which integrates a parcel generation program with an agent-based land use change model directly. Their system uses an agent-based land-bidding model, where land-use-driven acquisition of land provides the framework for determining agent actions. Agents make decisions about how to partition their parcel in two, by scanning the raster cells representing their lot according to several different algorithms, then choosing the one that best fits all personal and policy criteria (Alexandridis & Pijanowski, 2007).

Despite their computational advantages, it is widely upheld that "the raster data structure is better used as a means of data representation or integration, but not for behaviourally realistic modelling of the built environment" (Waddell, 2009). Grid cells are abstract spatial units that cut across parcel boundaries. As a result, they don't aggregate neatly to form parcels. Information is lost when assigning a single land use value to a cell that spans multiple land uses.

Without an explicit representation of parcels, integrated models can't really address policies and behaviours that vary at the parcel level. Nevertheless, there are considerable challenges of parcel level representation within integrated models. Integrated microsimulation models are likely to be unstable, display high stochasticity to the point that random variation overshadows systematic responses (Vanegas et al., 2009b; Waddell, 2009). This is due to the instability and uncertainty inherent in the behavioural processes that these systems attempt to model explicitly. More relevant to this research are two problems with the use of parcel data, namely, representing land use change and missing data from base year parcel datasets, both of which are addressed in part by block subdivision algorithms. Untaxed parcels and some in rural areas are often absent from tax assessor databases. The two approaches for filling in missing parcels are manual inspection and cleaning (brute force) and rule-based or heuristic methods. Manual inspection and cleaning is very time consuming and the effort can't be reused in other areas or datasets. This makes rule based, procedural methods preferable like the block subdivision algorithms used in City Generation Engines (Waddell, 2009).

City Generation Engines address the problem of representing land cover change within integrated models, albeit in a behaviourally unrealistic way (ESRI; Synthicity). The next section describes the efforts at spanning the disconnect between behavioural and geometric representation in the field of integrated land use-transportation modeling and the role of block subdivision algorithms within them. The following section describes the advances in block subdivision algorithms. The one after that describes research on urban form typologies. Thereafter, a methodology for testing block subdivision algorithms on different urban forms is outlined.

Difficulties simulating geometric changes in integrated models

As previously mentioned, block subdivision algorithms are necessary for representing land use change in a behaviourally realistic way. Such algorithms are most commonly implemented in City Generation Engines, in which they are triggered by population growth or road network formation. However in reality, parcel subdivision is most often a combined decision between a landowner to sell their fringe land and a land developer to buy and convert it to residential use. It can also result from the repositioning of obsolete or underutilized buildings and sites within an already urban environment, a process in urban infill. Though numerous behavioural models of parcel subdivision at the fringe exist (Irwin et al., 2003; Newburn & Berck, 2006; Pocewicz et al., 2007), such a model has only been incorporated into one integrated land use-transportation modelling system, namely, PECAS. The reason is that integrated modeling systems generally don't have the capability to compute these complex geospatial algorithms. Computation times for anything other than calculating distances between points are excessive. Furthermore, integrating behavioural simulation into a spatial simulation environment would place spatial computation requirements on the system that these aren't yet equipped to support efficiently.

A better appreciation of this comes with an understanding of the way that parcel data is structured. Parcels are stored in a vector format. A vector is a point located in space, or node, in relation to some reference point. In GIS, the locations represent geographic coordinates. Lines between the points are not stored explicitly but are rather inferred by the program. In this way a set of points can form a closed polygon: the basic structure of a parcel. The back end of spatial vector data not only stores locations of nodes, but also the adjacency relationships (or topologies) between lines connecting the nodes. This structure makes it possible to edit and spatially analyze the datasets in addition to many other common operations in geospatial analysis. This structure however makes data volume much greater than the individual objects the data represents, and operations between datasets computationally complex. Geographic Information Systems are designed to handle these large datasets and complex queries; however, most of them not in a temporally dynamic simulation environment. Integrated land use transportation systems in themselves currently can only handle point-to-point computations, as in calculating accessibility measures (Waddell, 2009).

Spanning the behavioural-geometric disconnect in integrated modelling

This section outlines the best attempts at spanning the behavioural-geometric disconnect in integrated modeling. The first is done within City Generation Engines, as discussed in the work of Vanegas et al. (2009b). The second is so far only theoretical and involves modeling spatial arrangements between objects within a multinomial logit type model (Schirmer, 2010). The most complete work in this area is the PECAS framework, which explicitly models behaviours leading to land use change and the resulting parcel subdivision process that occurs (Hunt & Abraham, 2007).

Vanegas et al. (2009b) developed a system that uses a set of superimposed grid cell layers, wherein each layer represents a different variable in the urban system. The behavioral variables represented are population count, job count, accessibility and land value. They spanned the disconnect between behavioral and geometric modelling by overlaying a set of grid cell layers onto vector datasets of parcels and roads. Each of these layers represents a different geometric variable of the urban system, namely road length, road tortuosity (ratio of road length to length of a cell) and building volume. In this way, physical properties of space are stored in grid cells in an abstract way and can interact with the behavioral properties. A change in a quantity of a behavioural variable such as population, leads to a change in a geometric variable such as road length. A changing cell value in the road length grid sends an instruction to the underlying vector based road network to extend itself by the amount of the change.

While this method of storing spatial properties of data has been used in GIS and Cellular Automata for some time (Xie & Batty, 2003), it is the mathematical models used to link the behavioral and geometric variables that make the Vanegas et al. framework useful in integrated modeling specifically. These models were developed by Waddell and Ulfrasson (2004). The basic form is that of a differential equation, wherein a change in a value of a cell is a function of all the values of the cells it superimposes. The solution of the differential equation gives the values of the other cells needed to satisfy the relationships between variables, in other words, the equilibrium conditions of the system.

Mobility and location choice are the primary ways that the system alters its state in order to reach equilibrium, or internal consistency. As an example, say the user changes the population or employment count of one or more cells. This alters the accessibility and land-values of cells, which results in a new spatial probability distribution of population and jobs. If the changes in the population or job variable are larger than the equilibrium convergence threshold of the system, mobility and location choice algorithms are executed. The population and jobs will be redistributed until the system reaches a state of equilibrium.

The location choice models used make several simplifying assumptions not made by traditional Discrete Choice models. One is that all agents are behaviorally uniform, with no specific attributes or types. Another is that the concept of utility can be substituted by a weighted attractiveness measure. This is measured as the weighted sums of accessibility and land value. This measure is translated into a probability by normalizing it with the sum of the attractiveness measures across all locations in the sampled set. Accessibility is calculated for each grid cell as a function of its distance from roads, slope of terrain and a distance normalized measure of activity level (population and job counts). Land value is calculated in a similar way.

The behavioural variables are connected to the geometric variables through a similar set of differential equations. Solving these equations gives the change in a geometric variable for a given change in a behavioural variable required to satisfy the relationships between variables defined by the differential function. To give a concrete example, the solution gives the target total length of roads inside a grid cell needed to accommodate its population and jobs. These values then become instructions to generate road networks and buildings. The road generation process builds on decades of algorithm development, reviewed later, and is sensitive to slope of terrain in addition to behavioural variables. To illustrate how the system functions: say a user draws a new highway and keeps the total population and jobs constant. The system determines that the new highway increases accessibility from the rural area to the downtown. As a result, population moves to now accessible lower land-value areas and new roads and buildings are adaptively generated. As closed paths of the road network form, the block subdivision algorithms are executed to subdivide the blocks into parcels. The number of parcels is generated to satisfy population and job density boundary conditions. Lastly, building envelopes are generated based on the per grid cell building volumes.

The developers tested this system by attempting to generate a real city using input data of Pacific Grove, California from a past year. While the spatial distribution of the city generated is not exactly like the real city, it is a plausible one. Furthermore, the system is stable, i.e.; it converges to similar equilibrium points when initial values of cells are changed in two different ways (though not the same because of stochasticity). Some drawbacks are the oversimplified representations of behavioural processes and equilibrium assumptions. Additionally, real estate developer decisions aren't modeled by the system, and so neither is urban sprawl. These changes rather need to be input by the designer. This system is capable of determining what a city will be like in the future, but not suitable for making quantitative predictions.

An attempt to span the behavioural geometric disconnect using the current integrated modeling framework has been proposed by Schirmer (2010). His approach would enhance the simulation results by accounting for geometry and avoiding unrealistic construction on parcels. The approach understands building types as objects that react along behavioural rules based on parcel variables. Shape grammars are rules for arranging objects in space that define the character of a street or area. A well-recognized and respected shape grammar rule is that all residential buildings should have street access. This method proposes a more elaborate set of rules, defining, for example, how much a given building type is offset from the street. For example, one can imagine that a taller more rectangular building tends to have a smaller street offset than a shorter more square structure.

Schirmer proposes that multinomial logit models can be estimated using empirical data that define the grammar rules, or typologies of an area. A high utility would be derived from putting a building on a parcel such that the street offset is realistic, ie: that it satisfies the grammar rule. While objects don't themselves experience utility from being arranged in certain ways, this might increase their usefulness or attractiveness to people. The approach might be extended to model the shapes of blocks or road network patterns that are carved out of a former agricultural parcel within the land use change process.

PECAS is the only operational integrated land use-transportation modeling system that represents land use change explicitly. When a parcel dataset is input, the system automatically recognizes parcels that are above a certain maximum area threshold, generally agricultural or undeveloped land parcels. It considers these as multiple smaller pseudo-parcels, formally subdividing them from the larger homogeneous parcel only when development events occur on the land (Hunt & Abraham, 2007). The algorithm used to create the pseudo-parcels is a binary space-partitioning algorithm that splits the parcel in half, and then the resulting two parcels in half, and so on until a maximum area threshold is reached. However, details on the exact algorithm used aren't available. PECAS also models the processes leading to land cover change, and parcel subdivision explicitly. The models it uses are in keeping with recent advances in behavioural modeling, namely estimating the utilities of each land parcel to developers in alternative uses. This informs probabilities with which land transitions from one use to another (Irwin et al., 2003; Newburn & Berck, 2006; Pocewicz et al., 2007). Based on these estimates, specific development events are randomly assigned to appropriate parcels.

Block subdivision algorithms

The idea to use computer programming to model patterns in the urban environment originates from Parish and Müller (2001). They adapted the use of L-systems, which had been successfully used to generate realistic trees in computer graphics programs, to the generation of road and highway networks. An L-system, or Lindenmayer system, is a type of formal grammar, containing an alphabet of symbols that can be used to make strings of symbols. The system contains a set of rules for expanding each symbol into a larger string, as well as a mechanism for translating these generated strings into geometric structures. Aristid Lindenmayer, a Hungarian theoretical biologist and botanist, introduced L-systems to describe the behavior of plant cells and to model the growth processes of plant development. Parish and Müller (2001) demonstrated how such systems could be extended to realistically generate road and highway networks, subdivide the land into lots, and create the appropriate geometry for buildings on the lots.

Their land subdivision algorithm involves recursively subdividing blocks along the longest pair of approximately parallel edges, until parcel sizes are under a user-specified threshold area. Parcels with no street access are deleted, leaving holes in the middle of blocks. This method also produces some irregularly shaped parcels. Later works have taken this procedural approach and applied it to improving algorithms for generating realistic building patterns and road networks.

More recently, research has also focused on improving techniques for automatically generating parcels. To Parish and Müller's recursive algorithm, Weber et al. (2009) incorporate a varying maximum area threshold depending on a parcel's land use type (Figure 1). That is, they define a maximum area below which no further splitting of parcels occurs. Furthermore, this area threshold varies with the type of land use, namely commercial, industrial or residential, that occupies the particular block being subdivided. They also rotate the split axis when necessary to ensure street access.

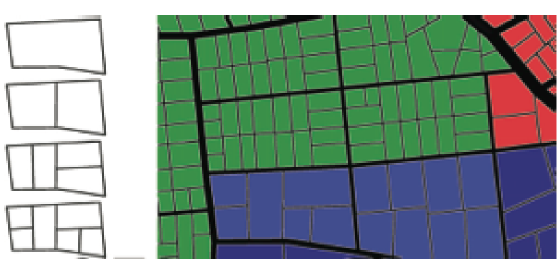


Figure 1 : *Left: The initial block (top) is recursive split (approximately orthogonal to the longest side) until all sub-polygons are below a land use dependent area threshold. Right: Resulting lots. Note the different sizes and ratios of the lots in residential (green) and industrial zones (blue).*

Figure 1: Recursive binary subdivision algorithm by Weber et al. (2009)

More recently, Vanegas et al. (2012) implemented an OBB (Oriented Bounding Box) algorithm as a method for recursively splitting street blocks into parcels (Figure 2). Use of the OBB produces more regularly shaped parcels, and ensures a maximum number are oriented parallel to an adjacent street. Their algorithm also tries to ensure street access by splitting the bounding box along either the long or widest edge. However this method also produces blocks with some patio parcels: ones with no street access but with access to an alley or similar.

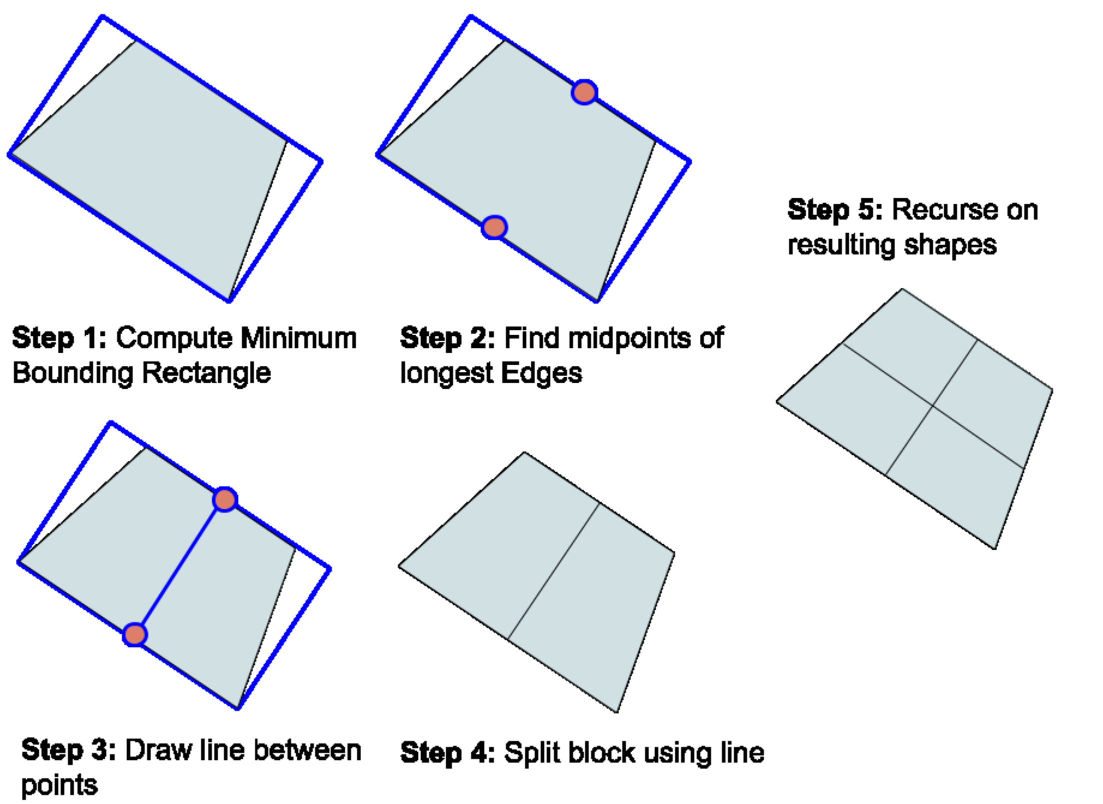


Figure 2: OBB block subdivision algorithm

Vanegas et al. also introduce another block splitting algorithm, called the SS (Straight Skeleton) algorithm, inspired by the method planners use to produce a perimeter-block. This is a block style wherein parcels surround the outer edge of a block, leaving an empty interior typically used for a school or park. Their algorithm is based on the straight-skeleton shape, which traces the outline of the initial block, at a user specified inner distance from the block boundaries (Figure 3). This skeleton shape is then split at regular intervals.

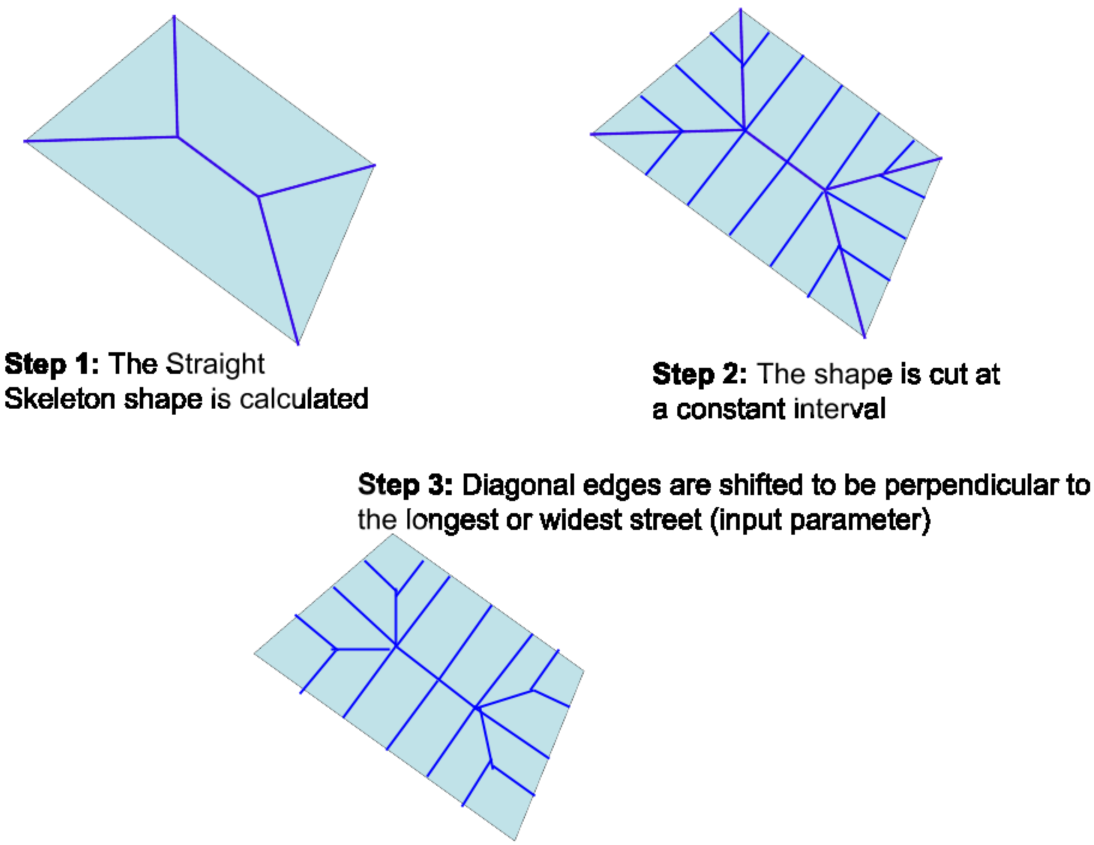


Figure 3: Straight Skeleton algorithm

One of the input parameters to the program specifies the distance to set the far edge of the parcel from the street. An infinite distance produces parcels with no interior space, but where one parcel is always abutting another as in Figure 3. Both algorithms in Vanegas et al. (2012) take as inputs high-level per block descriptive parameters specified by the user, and enforce them.

Wickramasurya et al. (2011) developed an algorithm to generate both roads and parcels within a block. Their method involves overlaying a grid onto a block, then using the block to clip the grid (Figure 4). Four different grids are tested in a single iteration with varying parcel and street orientations, and the most space maximizing one ultimately used. The selection is made so as to maximize the number of parcels or minimize the length of new roads within the block. Not only does their algorithm have built-in criteria for selecting the best grid, they also statistically test how well the parcels and roads they generated approximate the observed parcels in a given block.

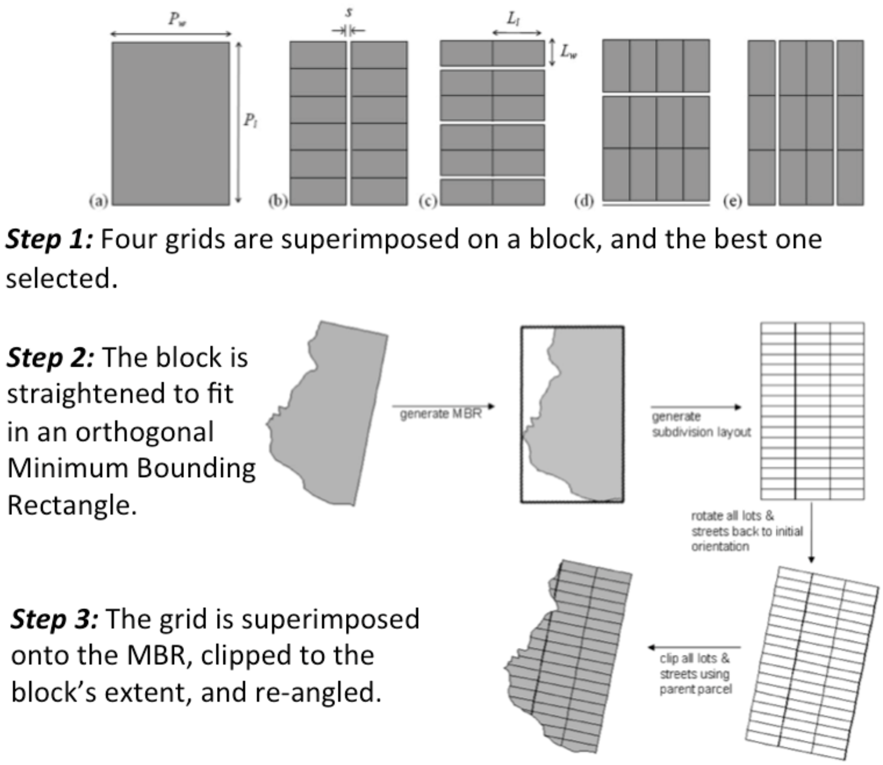


Figure 4: Algorithm by Wickramasuriya et al. (2012)

The built-in system for testing different grid structures developed by Wickramasurya et al. is an interesting one. It is this form of program to which the results of this research could be applied in the future. Namely, combining a set of block subdivision algorithms into a single program, and executing the one best suited to the particular block shape in question. However, this would be difficult to do without a better idea of which block subdivision algorithms are best suited to which block types- an understanding that the current research will hopefully contribute to.

Dahal and Chow built on the work of Wickramasuriya et al. (2011) by developing a set of tools for generating both roads and parcels. Each tool is designed to subdivide a land area of a specific shape, namely, L- or T-, or else to create blocks with a distinct subdivision style. These block styles include those with cul-de-sacs running through them, with inner looped roads, with multi-family housing or with a Manhattan style street network. Of all the tools, only one didn't specify a specific subdivision style or block shape and was considered more generally applicable on a wide variety of areas (Figure 5). Called GPD1 (Generalized Parcel Divider 1), this tool uses a combination of first, recursive binary subdivision and then grid drawing to create blocks that are adapted to an area's geometry. Once the shapes produced by the recursive binary subdivision have widths less than a user-specified value, roads are generated along their perimeter to produce blocks. A grid is then drawn within these blocks according to a user specified average parcel area and width. In this context, the user chooses from among the available tools in the toolset they deem best suited to the case at hand. However, they could also potentially be combined within a single tool that detects the block shape and then automatically executes the method best suited to subdividing it.

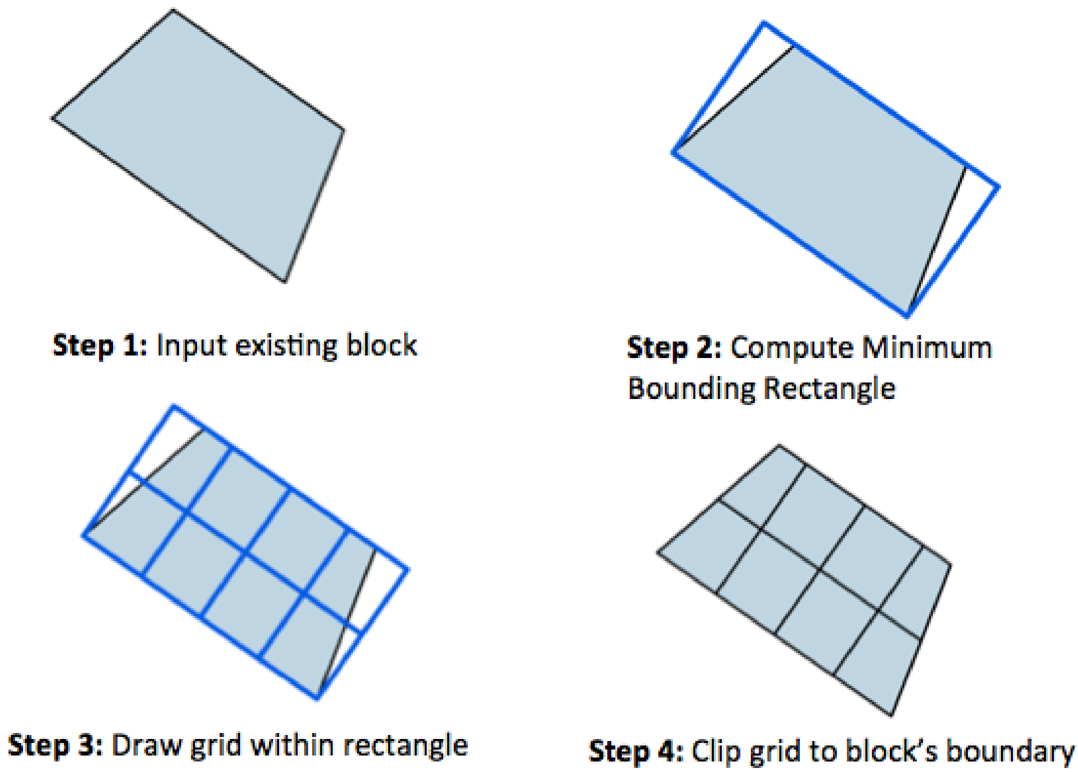


Figure 5: Generalized Parcel Divider 1 algorithm

Criteria for evaluating algorithms

Automatic block subdivision algorithms are designed to recreate several observed characteristics of real parcels. One of these characteristics is parcel shape, which tends to be regular or uniform, typically a deep rectangle, wide rectangle, approximate square, quadrilateral, sometimes polygonal (Vanegas et al., 2012). Other characteristics are that parcels have access to a street, are aligned with a street segment and have sufficient area to accommodate a building.

The three studies that evaluate block subdivision algorithms take slightly different approaches in their methods of evaluating these criteria. Each involves subdividing a block into parcels and then comparing these to their observed counterparts in a parcel dataset. Wickramasurya et al. (2011) have the most rigorous method of testing their algorithm's performance against actual parcels. They compared distributions of several metrics of similarity of observed and automatically generated parcels for two types of blocks. The comparisons were made using t-tests and correlation coefficients of number of lots and mean lots size. They also compared the standard deviations of the two distributions and their mean shape index (MSI), which is a measure of shape complexity. MSI is a variation on the SI (Shape Index) measure used in FRAGSTATS (McGarigal and Marks 1994) to measure how elongated a shape is (described later in the methodology section).

In the Dahal and Chow (2014) study, different tools are prescribed for different block types and so the tools were tested on their associated block types. An exception was Generalized Parcel Divider 1 (GPD1), which was tested on irregular shaped land areas but wasn't prescribed for them specifically. Visual comparisons were made for each of the tools in the toolset, whereas quantitative comparisons were made for GPD1, Cul-de-sac Creator and Divider with Inner Roads. Here, metrics of comparison included mean and standard deviation of width as well as length and size of lots. Number of lots and two shape metrics were also compared, namely, Regularity Index (RI), a measure of how square or rectangular a lot is, and Shape Index (SI), as mentioned above. Standard Errors were calculated for metrics of total lots and mean lot size only.

Vanegas et al. (2012) designed their algorithms to be used in City Generation Engines such as CityEngine (Esri) and UrbanCanvas (Synthicity). Users of CityEngine must manually choose which subdivision algorithm to use to subdivide a given block. To reproduce the results a user is likely to obtain, Vanegas et al. (2012) use each of the algorithms on different blocks that they suspect will perform well on these types. The metrics of similarity they use to compare the automatically generated parcels with the observed parcels are average area, average perimeter, length to width ratio and number of nearest neighbours. Short of performing statistical comparisons, they overlap the frequency histograms of the automatically generated and observed parcels to see if the distributions are similar. They also display the two sets of parcels in color-coded maps, where color represents the value of a parcel-level descriptive statistic. This allows for visual comparison of the results, which makes it possible to identify areas where the algorithm performed well and areas where it didn’t. However, it can’t help evaluate which algorithms worked better or say anything about the quality of one of their algorithms in relation to the other.

Vanegas et al. used their algorithms on blocks in three different types of neighbourhoods. The first is a mixed-use suburban area, composed of rectangular blocks with both straight and warped edges. The parcels within them are highly variable with respect to area, aspect ratio and minimum width. They subdivided each block using one of the two algorithms, favouring the straight skeleton method. Not only were the frequency histograms of the different metrics similar, but also the spatial distributions of their values as depicted in colour coded maps. Furthermore all generated parcels had dimensions and aspect ratios that were adequate for containing buildings.

They also subdivided blocks within a low-density suburban area with highly irregular shapes resulting from warped roads with frequent loops and cul-de-sacs. These cul-de-sacs resulted in blocks with large holes in the center, created by circular road segments. Here they chose to divide all parcels using the straight skeleton algorithm. The number of generated parcels was greater than the number of observed parcels, as reflected in the larger area under the histogram of the generated parcels. One of the observed parcels had a large area value that was a statistical outlier. This wasn't captured by the extracted per-block descriptive statistics and so it wasn't reproduced by their method. Other than these aspects the generated parcels had realistic shapes and all had street access.

The third site they tested had nearly rectangular parcels with some curved edges. They again used the straight skeleton method and found strong similarity between the frequency histograms of the metrics as well as their spatial variations within a block. To illustrate an extreme case, they subdivided part of the ancient Roman town of Pompeii, which had developed organically and displayed chaotic street and subdivision patterns.

They used the Oriented Bounding Box method and were able to achieve visual similarity by adjusting the split irregularity parameter, which displaces the split axis by a certain distance to account for variability in parcel widths. However they limited their tests of similarity to different street patterns of planned cities.

In general, the Straight Skeleton algorithm developed by Vanegas et al. performed well on highly irregularly shaped blocks with both straight and curved edges, composed of highly variable parcels. The Oriented Bounding Box algorithm also performed well on blocks it was used on, but because the analyses were grouped together and qualitative rather than quantitative, it is difficult to tell from the study how well it performed in relation to the straight skeleton.

Wickramasuyra et al. (2011) tested their program on two types of sites: a) with parcels oriented mostly in one direction and with all streets running mostly parallel to each other, and b) with regular or irregular shaped parcels that are oriented in many directions and with new streets oriented in several directions, including some cul-de-sacs.

The per-block parameters they used as inputs to the program were mode of parcel length and width. In contrast to Vanegas et al. (2012), they used both visual and statistical tests to evaluate their program. For blocks of the first type, that is ones with orthogonal parcels and roads, the tool produced similar results to the observed parcels. The number of parcels was the same as those observed (correlation coefficient of 0.99, p-value of 0.8762 for t-test). In this case, the null hypothesis of the t-test was that the two sets of parcels were drawn from identical distributions, so a high p-value suggests the observed and procedural parcels are the same. Except for parcels adjacent to the shortest length of the block, all generated parcels and streets were oriented in the same direction as those observed. Their algorithm generated exactly the same number of new streets oriented in exactly the same direction, and with 91.4% overlap with the observed streets. Furthermore parcels generated by the tool are usually rectangular, except areas clipped by the block boundary, in which case parcels were much smaller than the rest and highly irregular in shape. The correlation coefficient of mean parcel area was 0.92 and had a p-value of 0.4616 for the t-test, indicating the parcels have statistically similar areas. In addition, the standard deviation of parcel area and mean shape index were similar between the observed parcels and the generated ones, though no test of statistical similarity was performed. Their program generally performed well on grid type blocks, except for some highly irregular parcels adjacent to curved streets and some parcels oriented in the wrong direction.

However their program didn't perform well on irregularly shaped blocks, with variable parcel sizes and orientations as well as roads with variable lengths and orientations. Only the number of parcels yielded statistical results that suggest this metric is the same for observed and procedurally generated parcels (0.98, p-value = 0.483 for the t-test). Neither the number nor the length nor orientation of streets matched the observed layouts, and had only 22.4% overlap with the observed layouts. The tool doesn't have the ability to reproduce blocks with variably sized parcels. In addition, the observed parcels had a large variation in shape, as illustrated by a much higher MSI, which the model failed to reproduce. Their algorithm works well for subdividing approximately rectangular blocks but produces parcels of highly unrealistic sizes and shapes in irregularly shaped blocks.

For the visual comparison, the Dahal & Chow (2014) study found that for an irregular site subdivided by the GPD1 algorithm, the observed parcels are smaller, with larger size variation and different overall parcel orientations from the generated ones. For the rectangular site subdivided by the Divider with Inner Roads, there were a similar number of lots of equal sizes but a slightly different road configuration. Likewise, the Cul-de-sac Creator produced a similar number of lots of similar sizes to those observed within a rectangular site. Results were consistent for all tools and block types. Analysis of the standard errors indicate that for all sites, for the generated parcels, there are slightly higher numbers of parcels over an area but slightly lower numbers within each block and the parcel sizes are slightly larger. This is in comparison to the observed parcels for the same site. By qualitatively comparing the numbers, the average widths for both generated and observed lots were found to be equal while the lengths varied conspicuously, with generated parcels being more elongated. In addition, the generated parcels had more regular shapes and less size variation than those observed. All lots generated by the tools were guaranteed an access to a road. Total length of modelled streets for each of the compared subdivisions is shorter than length of observed counterparts.

Other goals of such programs are to minimize computational time and complexity, maximize level of automation, and optimize space within a block (Wickramasuriya et al., 2011). The programs of Vanegas et al. (2012) can generate all the parcels within a block in under three seconds. In addition, subdivision of two blocks after interactive editing can be performed at rates between 1-10 milliseconds per block. Similarly, the programs by Wickramasurya et al. can subdivide nine blocks into 92 lots in under 10 seconds. Dahal and Chow's Generalized Parcel Divider 1 tool was tested in the current research and was found to take a few minutes to subdivide a group of 9 blocks. It is possible to conclude that the Straight Skeleton algorithm performs best on irregular blocks, while the Oriented Bounding Box algorithm can be applied to blocks with mixed land uses. The algorithm by Wickramasurya et al. that I will refer to as the GO (Grid Overlay) performs well on regularly shaped blocks with little variability in parcel size but not on irregular blocks with variable parcel sizes. The Dahal and Chow algorithms produce similar numbers and sizes of lots, but that are more elongated and have lower size variation than those observed. The GPD1 also produced more regular lot shapes on square or rectangular sites.

Table 1 below summarizes the relationships for the more general algorithms that don't have a single block type or subdivision style associated with them. It also reveals the tests that have been performed and on which block types for the four such algorithms.

Based on these results, it would be difficult to formalize the relationship between block type and best subdivision algorithm. The reason being that it can't be said with confidence that one algorithm works better on a given type of block than any of the others. Such a formal relationship would be necessary for creating a larger program that captures the block type and automatically executes the tool most suited to subdividing it. Vanegas et al. (2012) concluded their study with the idea of developing machine-learning techniques to automatically capture the subdivision style of a block. Presumably such a technique would involve automatically detecting the block type from a road graph and then incorporating some knowledge of which algorithm is best suited to subdividing such a block type. A step toward this might be for a human to determine which algorithms work better for which blocks. This would involve identifying different block types and then testing each block subdivision algorithm on each type of block. Then a formal relationship between block characteristics and optimal subdivision style might be defined and incorporated into a program.

		Algorithms			
Parcel Patterns	**Block Shapes**	OBB	SS	GO	GPD1
Regular Grid	Square or Rectangular		✔	✔ H_o	
	Irregular				
Irregular Grid	Square or Rectangular	✔			✔
	Irregular				
Curvilinear Non-Grid	Irregular		✔	✗ H_o	
Farmland	Square				
	Rectangular				

Compared with actual parcels ✓ **Generated similar parcels**

✗ Generated dissimilar parcels H_0 Statistically tested

Table 1: General block subdivision algorithms and tests done

Why road networks matter

The previous section reviewed automatic block subdivision or parcel generation algorithms. Parcels exist within blocks and so the road networks that structure these blocks are also important for automatic parcel generation. Some of these parcel generation algorithms have built-in road generation functions (Dahal & Chow, 2014; Wickramasuriya et al., 2011). These are referred to as stand-alone tools because they can function independently of a larger program, while other parcel generation algorithms are part of a larger city simulation tool with separate road generation algorithms (Parish & Muller, 2001; Vanegas et al., 2012; Weber et al., 2009). To simplify the evaluation of parcel generation algorithms in this study, these are applied directly to blocks derived from the negative space of an already formed road network. In reality, these blocks must be generated automatically in new urban areas where local road networks don’t yet exist.

The sequence of road and parcel generation varies between different tools and tool settings however in all cases the road networks ultimately shape the parcel patterns. Within City Generation Engines, these road generation algorithms are executed either in sequence or in parallel to the block subdivision algorithms (Vanegas et al., 2009a; Weber et al., 2009). When a city is simulated from scratch, topography is created and then streets are generated. Subsequently, blocks are subdivided into lots and then buildings are generated on them. Thus, road generation and block subdivision operate in sequence. Here, it is the negative space of the local road network that serves as inputs for the subdivision. In contrast, with manual editing or visualizing integrated modelling outputs, a road might be generated to ensure all generated lots have street access or that the lot number per block doesn't exceed a specified maximum. Thus road and parcel generation operate in parallel. In any case, it is clear that the geometry of the local road network affects the geometries of the resulting parcels.

There are several ways in which parcel geometry and road network geometry can interact to affect population forecasts. First, the number and sizes of parcels within a block affect population density. Road networks are important in this context in light of the finding that measures of street configuration are correlated to measures of population and parcel densities (Peponis et al., 2007). Second, through typological rules, the shapes of parcels can affect the building types that are placed there. In City Generation Engines, these rules are encoded into Computer Generated Architecture rule files that specify geometric and other relationships between buildings and parcels that must be satisfied. Schirmer (2010) proposed a logit model for predicting building type within integrated modeling frameworks based partly on parcel geometry. Thus, buildings could be constrained to parcels where the resulting combinations are plausible. For example, a residential high rise is unlikely to have a large street offset, while a single-family house is more likely to. This is relevant since building type is a factor in household location decisions and in the models used to predict them (Waddell et al., 1998). In this way, a more realistic spatial distribution of parcel shapes and hence building types can lead to more realistic demographic forecasts.

A short review of some road network vocabulary can help in understanding what these road generation algorithms do. Roads can be categorized into two types: local or minor roads and major roads or access structures. Local roads typically provide access to and from housing while access structures provide access to a larger number of locations. The three primary access structures are radial, grid and circular roads (Figure 6). A structure can be regular or irregular as well as hierarchical or equal (Figure 7). In reality, road accessibility is better thought of along a spectrum of road network hierarchy, which describes the number of connections a road has with other roads in the network. Roads with a higher number of connections relative to others in the network are considered access structures and roads with a lower number are considered local roads. In a road network that has no hierarchy, where all links have equal connectivity, all roads are access structures as is the case for gridiron networks (Hillier, 1996). It is generally local roads that structure blocks, and so it is these types of road networks that this study will be concerned with. However, since the algorithms used to generate both access structures and local roads are often integrated they will both be described here.

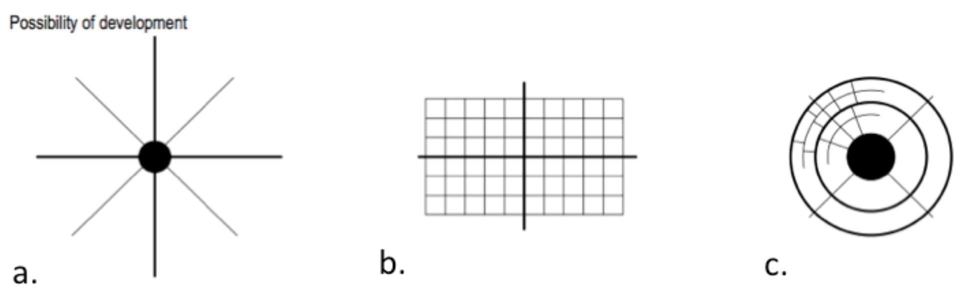


Figure 6: Primary access structures; a. radial, b. grid, and c. circular, Albert Speer & Partner (2012)

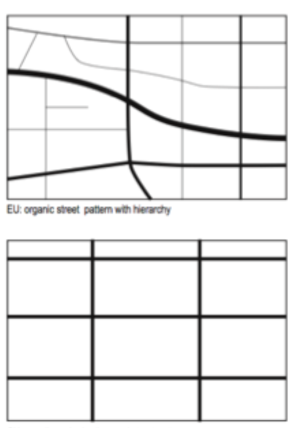


Figure 7: Top: A hierarchical, irregular road network. Bottom: an equal, regular grid road network, Albert Speer & Partner (2012)

Road Generation Algorithms

There are a number of methods for automatically generating road networks. In a review of procedural methods for terrain modelling, Smelik et al. (2007) describe several methods for generating road networks of specific patterns, among them pattern-based approaches, L-systems, agent simulations and tensor fields. The following studies draw mainly from their review. Sun et al. (2002) created a template system for generating major road networks of different types. Their system contains templates for raster, radial and mixed networks as well as one based on population centers. Minor roads can be filled in as a raster of streets. L-systems, first used to generate roads by Parish and Muller (2001), are capable of generating a wide variety of patterns, including grid and radial. Kelly and MacEwan (2007) created an interactive system for generating major roads based on elevation differences between user-specified origin and destination points. In this system, L-systems can be used to generate minor roads within areas enclosed by major roads. By adjusting input parameters, a number of minor road network types can be created. These include raster (or gridiron), industrial and organic- an approximation of road networks in North American suburbs. To reproduce informal settlement patterns in South Africa, Glass et al. (2006) used a Voronoi diagram to generate major roads. Minor roads were generated using either L-systems or regular subdivision with optional displacement noise. Lechner et al. (2006) took an agent-based approach to automating road generation adapted to residential, commercial or industrial areas and even to government buildings, squares and institutions. Chen et al. (2008) proposed the use of tensor fields to model road networks of different patterns including grid, radial and along a boundary. Aliaga et al. (2008) created an example-based approach of copying road network nodes from existing vector data or aerial images to areas where new roads are required. The nodes stored properties of the existing road network such as intersection degrees and road angles, which were then regenerated in the new areas. It is clear that the methods for generating road networks are well advanced, even for reproducing specific patterns.

Several of these approaches are integrated into City Generation Engines, which have some capability of generating road networks of specific types. The method of Aliaga et al. (2008) is used together with a stochastic sampling of nearby street nodes. In this way, streets have similar lengths and turning angles to those in their vicinity and blocks have similar aspect ratios. Weber et al. (2009) make use of L-systems to generate minor streets, whose type varies according to a user defined land use value. The possible types include organic, grid and radial. Aliaga's method assumes road network type is a function of road network types in its vicinity, while Weber et al.'s method assumes it is a function of land use type. A possible application of the type of road network model developed in this research is to estimate which road network type is likely to appear based on a number of other factors as well. For example, in Weber's system, where organic and raster road networks can be generated, a model is estimated in this research to estimate which of these two forms is likely to appear where. Given the chain of simulation events that link population forecasts back to road network type, such a model could potentially improve these forecasts.

Historical Survey of Planning Paradigms

As described above, City Generation Engines use a simple process for selecting the future road network type to generate. The current research aims to increase understanding of the factors influencing type of local road network that is constructed in future urban areas so that it might be applied in simulations. Sandalack et al. (2013) conducted the most comprehensive analysis of urban form in Canada. Urban form, a component of which is the local road network, structures the physical character of neighbourhoods. The authors found that neighbourhood types are the product of planning paradigms, societal values and economic influences. By understanding history, it might be possible to understand how historical planning paradigms are playing out in the present. More specifically, one can hope to understand why a given road network type might be selected for a particular neighbourhood in the present, by understanding the historical circumstances under which the road network design was conceived. With this in mind, the following section provides an overview of planning paradigms in Britain and North America over the 19th and 20th centuries that draws mostly from the book "Streets and the Shaping of Towns and Cities" (Southworth & Ben-Joseph, 2003). Understanding when and why different road network types were used can help in predicting future trends in their spatial distribution.

A historical analysis shows that the dominant planning trend in a given period cycled between practicality at one extreme and aesthetics or privacy at the other. In the early 18th century, the development of the railroad gave rise to the gridiron or grid street pattern. The main driving force of growth at the time was land speculation and so a design that could be built quickly and in a modular fashion was desirable. In reaction to the monotony of the gridiron, Olmstead and Vaux developed a suburban design in the late 1800s. Here streets followed natural topography and provided aesthetic views from homes (Southworth & Ben-Joseph, 2003, p. 32). The dominant trend subsequently wavered between these two extremes. The Garden suburb was the origin of many features of modern suburban street networks. For example, streets were designed to bend or terminate with views of buildings rather than open streets creating an enclosure within a block, a variation of the modern "loop" (Southworth & Ben-Joseph, 2003, p. 40). In 1904, Unwin & Parker introduced the cul-de-sac in their design of Hampstead Garden. The cul-de-sac is a residential street that excludes through traffic. Bends and cul-de-sacs also provided a spontaneous and unstructured aesthetic that could "meander about aimlessly, comfortably, following the natural contour and advantages of the land" (Southworth & Ben-Joseph, 2003, p. 44).

The City Practical movement in America in the early 1900s emphasized scientific management of cities through rational planning and utilitarian ethics. Here suburbs served to redistribute the middle class to outlying areas to relieve pressures on housing in the center. Providing fast and low cost transportation would encourage industry to locate on the fringe. With the opening of new automobile routes and streetcar lines to suburbs, a new fringe based on speculative development began to take shape. The prevalence of loop and lollipops here can be explained by the shifting of development from the hands of municipalities to private land developers. Loop and lollipops are a profit maximizing design: forming blocks with the largest amount of buildable area and requiring a shorter total road length than other types. At the same time, the rise of the automobile led to the decline of the gridiron street pattern for real estate development since it provided no safety or shelter from traffic.

With the "City Beautiful" movement, the concept of the neighbourhood unit emerged as a response to the fact that speculative growth diminishes the sense of community in residential neighbourhoods. Perry & Adams conceived of a neighbourhood unit with a population size that requires one elementary school, bounded by arterials, with hierarchical local streets that facilitate local traffic but discourage through traffic. Schools and neighbourhood institutions were often in the center, with shopping on the edges and streets running between them to provide ease of access. Staggered cross streets provided safety, attractiveness and variety (Southworth & Ben-Joseph, 2003, p. 68).

New Urbanism began in the 1980s in an effort to encourage higher densities, public transit use, and accommodate pedestrians and bicyclists. New Urbanist developments fall into two types: Pedestrian or Transit Oriented Development (TOD) and Neotraditional development. TOD reinforces use of public transportation by providing accessible transit stops and commercial areas. In contrast, the primary motivation of Neotraditional development in providing mixed use, mixed housing and walkable pedestrian networks is to reproduce the style of a classic small town. The interconnected pedestrian networks of cul-de-sacs here are often criticized as being explorable rather than walkable, where they don't provide access to important amenities. From a geometric point of view, a study of New Urbanist developments by Southworth & Ben-Joseph (2003) found that 10% were loop and lollipop, 40% were grids and 50% were loop/grid hybrids. Furthermore, 80% of the time alleys were present.

Sandalack et al. (2013) conducted an analysis of urban form in Calgary, the most comprehensive such analysis in Canada. The study classified all Calgary neighbourhoods into three main categories, each arising in a different period and having a characteristic block design. The study illustrates how the planning paradigms that developed in England and North America were appropriated over time in a particular context. The 'grid block' pattern characterizes the first phase of Calgary's urban development starting in 1883 until the Second World War. The block pattern usually extended the existing grid framework, subdividing it at right angles. This type is characterized by a grid block pattern with main-street mixed-use commercial streets. After the Second World War, economic development and visions of an idealized post-war suburban life fueled the development of auto-suburbs. This corresponded to Calgary's second ring of development, in which land uses were segregated through zoning regulations and single-family homes were organized in neighborhoods around schools. A new hierarchical street network emerged that incorporated crescents, cul-de-sacs and curved roads with the intention of breaking up the perceived monotony of the grid. This neighborhood type is characterized by a warped grid (crescents and drives) block pattern with a local shopping center. Calgary's third and current phase of urban development created the outer ring of suburbs. Although these neighborhoods are generally marketed on some aspect of their uniqueness, they share some essential qualities. High-volume collector roads with huge land buffers on either side separate neighborhoods. The neighborhoods themselves have highly impermeable curvilinear street systems so as to increase a household's perception of private space. Garages occupy much of the front lots, and sidewalks are often absent. This type is characterized by a curvilinear "loops and lollipops" street pattern with strip convenience stores and services. With a few exceptions, all Calgary neighborhoods have been described as one of these three types. The study illustrates how the design of neighborhoods in a given area was heavily influenced by the dominant planning trend in a given era. However, the historical conditions and purposes that gave rise to different neighborhood and road network types can help in understanding the conditions under which they might be built in the future on a more localized scale; conditions the road network models estimated in this research attempt to capture.

Conclusion

Several block subdivision algorithms have been developed and the parcels they generate have been compared with their observed counterparts both statistically and visually based on a number of different metrics. However these algorithms have never been compared with each other to see if certain ones work better overall or on certain block types. This research uses a common local road network classification combined with a parcel pattern classification to determine if any block subdivision algorithms perform better on given block designs than others. The aim is to define a relationship between optimal block subdivision algorithm and block design. A larger program that incorporates all of these block subdivision algorithms could make use of such a relationship to more accurately automate the block subdivision process within City Generation Engines and potentially within integrated modeling frameworks, once these are capable of representing geometric changes to the landscape. This relationship can also be used to more accurately generate missing parcel data used as a baseyear input to integrated land use-transportation models.

Block subdivision algorithms are often used to subdivide blocks in future residential areas, where local road networks haven't yet been built. Road generation algorithms that simulate them are currently capable of generating blocks of different types, and some of these within City Generation Engines. Currently, the local road network to be generated in an area is determined by the land use type or else the road network types in its vicinity. To help make best use of these algorithms in representing urban form, this research also attempts to anticipate the local road network type in future residential areas based on a wider variety of factors that characterize aspects of the terrain, the transportation network, land use and time of development. The estimated road network type categories match those used to test the block subdivision algorithms so that the findings of each study might be integrated into a program that estimates the local road network type in a future area, generates it and then executes the block subdivision algorithm best suited to subdividing it. It is hoped that these findings will allow for better representation of the spatial distribution of urban form in future residential areas within integrated modeling tools.

Rationale of Study and Articles in Brief

I chose to focus my research on block subdivision algorithms to complement available data, namely parcel data for the Montreal Metropolitan Community. This also seemed to be an opportunity to contribute to an interesting area of research concerned with integrating behavioural and geometric modelling.

My thesis is concerned with existing algorithms and methods for generating urban form elements within integrated modelling tools. It deals with the specific question of when to use which methods so as to more realistically represent different urban forms within future urban areas.

This question is addressed in order of which urban form elements are most directly relevant to population forecasts. First, several block subdivision algorithms are evaluated to compare how well each one functions for different block types. The next chapter addresses the question of which of these block types, or road networks, is likely to appear where. The two studies were designed so that their results could be combined with each other and within state of the art urban simulation tools.

Study Area

The study spans the Communauté métropolitaine de Montréal (CMM) (Figure 8), whose population was 3.7 million in 2013 (Affaires municipales, Regions et Occupation du territoire). This area was chosen because it complemented the available parcel data acquired from Foncier Quebec. The extent was chosen because it represents the political boundaries of the metropolitan area. It consists of a central island (the Island of Montreal) with surrounding regions connected by bridges. This physical separation manifests itself in a variation in culture and aesthetic preferences as well. Montreal has a decentralized planning structure organized around the Municipalité régionale de comté (MRC) with 14 MRCs spanning 82 municipalities. The MRCs are responsible for creating regional development plans that are flexible enough to allow local municipalities to identify land uses, urban limits and areas for urbanization, requirements for infrastructure, development densities as well as the location and type of major roads (Tomalty, 1997). This decentralized planning structure is likely to result in a large spatial variation in neighbourhood designs useful for this study.

Figure 8: CMM by AMT transit planning zones, Main: Montreal's urban core

Montreal’s long history provides a time slice over which a number of planning paradigms have been experienced (Dufresne et al., 2003). Founded in 1833 (CMM, 2005), the St-Laurence River heavily influenced settlement patterns there. Its banks were colonized first and agricultural development subsequently expanded to the north and south in a roughly monocentric growth pattern. North of the river is the Island of Montreal, representing about 10% of the land area. Old Montreal, located on the island’s southeast coast, was the first urban center (Dufresne et al., 2003). Now also an historic neighbourhood, its location forms the origin of the Central Business District.

Recently growth in the outskirts has exceeded that in the city center. Between 2001 and 2006, the average population growth rate was three times lower in the center (0.5%) than on the outskirts (1.5%) (CMM, 2010). Furthermore, exurban areas- where >50% of the population commutes to work- grew by 18,500 people from 2006-2011, the largest increase in any major Canadian city. The population of auto-suburbs grew by 163,000 (Gordon & Shirokoff, 2014). These demographic trends should provide a rich source of data for studying recent neighbourhood designs characteristic of fringe development.

Methods of Analysis

The methods used for this research are relatively simple by design and are covered for the most part in the Literature Review and elaborated here. The general method for determining whether a block subdivision algorithm performs well is to actually use it to subdivide a block and then compare the resulting parcels to their observed counterparts in a parcel dataset. The observed parcels come from the 2009 cadastral data of Quebec distributed by the Ministère des Ressources naturelles et de la Faune (Foncier Québec 2009). For the road network study, a simple logit modelling structure is used.

The first sections of the methodology describe methods used in the first Manuscript – Testing Block Subdivision Algorithms on Block Designs. The last sections describe methods used in the second Manuscript – Modeling Road Network Type. Among other things, this methodology will provide a detailed description of the process used to classify sample areas by road network type, which was used in both the road network models and in testing the block subdivision algorithms.

Metrics of similarity

Metrics to use as a basis of comparison between sets of observed and automatically generated parcels were selected from among those used in previous studies. Area and width were selected because they were used across all studies, are sometimes specified directly in policies and clearly relevant to population forecasts. Street access, or egress, was selected since it was also used in some previous studies and is a general requirement for residential parcels. In contrast, parcel shape is an ambiguous concept and was measured differently in each of the studies. Wickramasuriya et al. (2011) used Mean Shape Index, Khila and Dahal (2014) used Regularity Index and Shape Index, Vanegas et al. (2012) used width to length ratio. For this study, Shape Index (SI) was used because it can be calculated easily from the parcel perimeter and area and can be interpreted simply as how elongated a shape is. The study was limited to one shape metric to minimize the quantity of results for this meta-analysis while still capturing important parcel characteristics.

The SI is a measure of shape complexity found in FRAGSTATS, a computer program designed to compute a wide variety of landscape metrics for categorical map patterns (McGarigal & Marks, 1994). The SI corrects for the size problem of the perimeter-area ratio index, namely, that holding shape constant, an increase in area will cause a decrease in perimeter-area ratio since area increases at a higher rate than perimeter. The SI corrects for this problem by adjusting the ratio to a standard square having the same area as the parcel in question. The SI is calculated by dividing the parcel perimeter by the perimeter of a square parcel of the same area as follows:

Equation 1: Shape Index (SI)

$$SI = \frac{P_{ij}}{4 * \sqrt{a_{ij}}}$$

Where P_{ij} is the parcel perimeter and a_{ij} is the parcel area. Thus the SI would be 1 for a square and would increase as the parcel becomes more irregular. These three metrics were calculated for each parcel in each observed block as well as for those in each automatically subdivided block.

Tests of Similarity

In the study of block subdivision algorithms, three statistical tests were used to determine whether or not an algorithm generated parcels with similar metrics as those observed. Because of the nature of hypothesis testing, the tests could only definitively determine when the difference in metrics between observed and generated parcels wasn't zero, hence when the metrics' values were different. However, where the null hypothesis wasn't rejected this doesn't imply that the generated and observed parcels are the same. For the purposes of this study, this was still considered to be a good method for distinguishing cases where an algorithm performed well and cases where it didn't. However, in light of this the results should be interpreted as how well the different algorithms performed relative to each other, rather than whether or not they performed well in an absolute sense.

The three statistical tests used are the t-test, the Kolmogorov Smirnov (ks) test and the Fisher's exact test. Results of the t-test allow one to make inferences about the similarity between average parcel metrics of observed and generated parcels. The ks-test is a non-parametric test used to make inferences about the similarity between whole distributions of metrics. A bootstrap version of the ks-test was used to calculate accurate p-values on the test statistic, which was necessary since the distributions contained duplicate values (Jasjeet S. Sekhon, UC Berkeley). The Fisher's exact test allows one to test the egress metric in particular. Specifically, it allows for making inferences about the similarity of the proportions of observed and simulated parcels with egress. Each test was conducted on samples of at least 30 parcels but often more; a sample size generally regarded as sufficient where the distribution is not too abnormal and a high degree of confidence isn't required (Ruxton, 2006). This is because of the central limit theorem, which shows that as the sample size increases, the distribution of the averages of a random variable becomes increasingly normal.

The null hypothesis of the t-test is that the difference in mean area (or other metric) between the simulated parcels and observed parcels is 0. A p-value less than 0.05 implies that the two means are different and that the algorithm failed to reproduce parcels with statistically similar mean metric. In contrast, a p-value greater than 0.05 doesn't indicate the two means are the same, only that they cannot be concluded to be different. The null hypothesis of the ks-test is that the distributions of metrics for both

sets of parcels were sampled from populations with identical distributions. Again, a p-value less than 0.05 implies that the two parent distributions are different and the algorithm failed to produce parcels with a statistically similar distribution of a given metric. The null hypothesis of the Fisher's exact test is that whether or not a parcel has egress is independent of whether or not it is simulated or observed. That is, the proportions of simulated and observed parcels with egress are the same. Here a p-value less than 0.05 implies that the proportion of parcels with egress does depend on whether or not the parcels are observed or synthetic, and hence that these proportions are different. Details about how each of the tests works is included in the manuscript Testing Block Subdivision Algorithms on Block Designs.

Site Sample Size

In the past, tests of block subdivision algorithms have been done on one or two samples of each block type (Dahal & Chow, 2014; Vanegas et al., 2012; Wickramasuriya et al., 2011). Results of these studies generally allow one to say whether or not a block subdivision algorithm performed well on a given block type, but nothing about *how* well they performed. This makes it difficult to compare between algorithms with the aim of finding an optimal one for a specific block type. Doing so would require a more quantitative characterization of the degree to which each algorithm performs well on a given block type.

To remedy this, around 30 sites of each block type were tested. Each site from the sample was compared individually with the observed mean. The results for a given algorithm and site type were aggregated to give a proportion of times the statistical test was passed. This provides a quantitative basis for comparing the performance of the different algorithms for a given site type and metric.

Sample areas

The rest of the methodology deals with methods used in the second manuscript – Modeling Road Network Type. Before the sample road network areas could be classified, they needed to be created from the continuous road network surface. A number of different studies investigate the relationship between road network type or built environment variables and other factors and these take a variety of approaches to defining spatial units. Studies that investigate how built form variables affect mode choice might define sample units as the route for a trip being studied (Rodriguez & Joo, 2004; Winters et al., 2010). Studies in which road network type is the variable being studied often use the grid cell approach (Lovegrove & Sayed, 2006; Sun & Lovegrove, 2013). This involves finding the average block size within the area in question and then superimposing a grid with cells of that size onto the area. Cells can be aggregated to repeat the experiment at multiple scales to address modifiable aerial unit problems. The most frequent spatial unit used in studies of road network or built environment type is the TAZ (Traffic Analysis Zone). This is a unit that corresponds roughly to the Census tract and is often used as a proxy for a neighborhood (Rifaat & Tay, 2008). It is considered a spatial unit with roughly stable population. This approach is often used in combination with the grid cell approach, by aggregating grid cells within a TAZ (Sun & Lovegrove, 2013). Another approach to grouping neighborhoods is cluster analysis. To study the effects of neighborhood type on housing value, Song & Quercia (2008) clustered zones into traditional, neo-traditional and conventional suburbs, based on the similarity within a set of predefined design features of neighborhoods. The original study by Southworth and Owens (1993) that conceived of the road type classification system studied urban form at the lot level, the neighborhood level and community level. The neighbourhood unit used measured about 100 acres and encompassed an area that would take less than ten minutes to walk across.

Some of these studies used a qualitative characterization of built environment (Rifaat & Tay, 2008; Southworth & Owens, 1993). While others used a quantitative characterization based on the similarity within a number of predefined neighbourhood design features (Song & Quercia, 2008). The study by Rifaat and Tay (Rifaat & Tay, 2008) used the Southworth & Owens classification system, but merged the two grid categories and the two lollipop categories and additionally incorporated a mixed category.

The difference in the research conducted in this thesis and all of these studies is that here, road network type is the dependent variable. It wouldn't be practical to have a non-uniform road network type -category, given that there are so many possible combinations of them, and since many road networks at the TAZ level are mixed, this would greatly reduce the sample size. Furthermore, the grid cells wouldn't align perfectly with the borders of neighborhoods, probably introducing more mixed categories. One solution to this would be to reduce the size of the grid cells, however, with decreasing size, the cells would get more difficult to classify since key links in the design might be missing.

For this study, it was important to have sample areas with uniform and recognizable road network type and a sufficiently large sample size for the model estimation. As a result, neither the grid cell nor the TAZ spatial units were used. A novel way of defining neighborhood units was developed here to ensure as many as possible were of a uniform and recognizable type. Considering this, a clustering approach was taken in this research to isolate neighborhoods based on observed characteristics rather than on a predefined unit or size. The method produced an average neighborhood size of 481,844 m^2 similar to that in the Southworth and Owens (1993) study of 405,000 m^2. By clustering residential housing units, discrete areas were created. The final 120 m cluster tolerance was found by incrementally increasing the distance until clusters were formed without obvious gaps. This value can be thought of as a metric of the maximum distance between houses within the same development. In certain cases, the clusters still had gaps and so these were identified and the clusters were closed using an ArcGIS Geodatabase topology. The clusters were defined using residential building points, and so were often jagged where housing is dispersed. To create more regular sample areas, these were smoothed by applying a 200m buffer followed by a 200m negative buffer. This way the new sample areas were roughly the same size and shape, but the edges were smooth and didn't omit important terrain features or road links. Finally, the sample areas were clipped by a buffered layer of highways, arterials and collectors (at 25m, 15m, and 15m respectively), in keeping with the definition of a neighborhood as an area surrounded by major roads (Sun & Lovegrove, 2013). Road links were then spatially joined to the clusters for classification. The independent model variables were also calculated based on these clusters.

Road Network Type Classification

The following section describes the road network classification system used in both manuscripts included in this thesis. It also describes the process by which sample areas were classified into road network categories to assign dependent variables for the model estimation.

Many studies of classification systems have been conducted. These highlight that categories we assign to objects are socially constructed rather than essential properties of the objects themselves. For example, there are around a dozen different names for the word snow and ten for the word ice in Inuktitut (Shneider, 1985). The book “Sorting Things Out: Classification and Its Consequences” (Bowker & Star, 1999), demonstrates how classification systems have been devised to render certain groups invisible to the benefit of other groups. Urban form typologies are also socially constructed. This is apparent from the large number of road network type classification systems, each one assigning another label to essentially the same real object. In the book “Streets and Patterns”, Stephen Marshall shows that there is little consensus on the meanings we attach to terms for labeling road network types. This stems from the fact that different types are distinguished by spatial or geometric properties and these are difficult to fully describe using language. Images can provide examples of different types, but they are far from being complete definitions. As a result, the practice of classifying road networks patches into types is highly ambiguous.

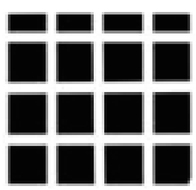
Marshall offers a way around this ambiguity; to accept that “the act of classification effectively carves a series of discrete types from the ‘morphological continuum’ of all patterns” (Marshall, 2005, p.81) and that where the divisions are made on this continuum is somewhat arbitrary. In the case of road networks, the continuum is described as being a perfect grid on one end of the spectrum, and a perfectly branching tree on the other end. Rather than ask what the road network categories are and how to assign real objects to ideal categories, we should ask why we are classifying and how we can best devise categories and ways of mapping objects to categories for this purpose.

With this in mind, a number of road network classification systems were explored (Cherry & Nagel, 2009; Sandalack et al., 2013; Southworth & Owens, 1993). Of these, the one by Southworth & Owens (1993) was selected. It is both a geometric and functional classification system, informed by an extensive historical analysis of planning paradigms in both urban and suburban areas and a morphological analysis of actual sites. The authors studied several US neighborhoods in the San Francisco Bay region that would have represented the urban fringe at different periods in the 19th and 20th centuries. They defined a neighborhood as a one hundred acre area that takes less than ten minutes to walk across (~two thousand feet). From the study, they defined five different types of local road network patterns- gridiron, fragmented parallels, warped parallels, loops & lollipops and lollipops on a stick- as well as several mixed categories.

The following is an overview of each road network type included in the classification system taken from Southworth & Owens (1993). It includes a short geometric description of each type, its functional form, and the overall character of the neighborhood. All the information was relevant to constructing a classification key for assigning road network patches to categories, described later in the Methodology.

- <u>Gridiron</u>

Figure 9: Gridiron



The gridiron is formed by two series of parallel lines crossing at right angles. The resulting blocks are equally sized square or rectangular shapes. Large landscape features such as institutions, parks and water bodies sometimes break the pattern. It has more land devoted to streets, more blocks and intersections than the other patterns. The gridiron is easily expandable and requires little planning. It offers the largest number of route choices and shortest trip lengths resulting in the most walkable neighborhoods. With non-hierarchical, perfectly connected streets, there is nothing to discourage through traffic.

- <u>Fragmented parallels</u>

Figure 10: Fragmented parallels

Like the gridiron, this type is orthogonal but the blocks are long and narrow and sometimes L-shaped. It favors T-intersections. This type has similar total road length as gridiron, but with fewer blocks and access points in and out. The long blocks provide more street frontage for residential building lots. However, this pattern lengthens pedestrian trips compared to gridiron and also limits through traffic.

- <u>Warped parallels</u>

Figure 11: Warped parallels

In this type, the blocks of the fragmented parallels pattern are warped or curved. The long narrow blocks provide optimal street frontage for residential building lots. Occasionally cul-de-sacs fill in areas formed from the curves. Curves provide a more rural character and shorten the visual length of streets, though they aren't a response to topography. There are a similar number of intersections, access points and route choices as fragmented parallels, but curves make the streets more confusing to navigate.

- <u>Loops & cul-de-sacs</u>

Figure 12: Loop & lollipop

The loop and lollipop pattern lacks the interconnectedness of the other types. It is highly hierarchical, with both loop and cul-de-sacs being entirely residential and limiting through traffic. It has a high ratio of 3-way to 4-way intersections (i.e.: 14:1). There are a few variations of this type, with more uneven topography lending itself to more free form patterns. Lazy loops and lollipops weave around topography following the most convenient path (i.e.: along lines of constant elevation). Clustered loops and lollipops have denser development surrounded by open space unsuitable for development due to slopes, soils, seismic factors, endangered species, streams or power line easements. These give the impression of self-contained enclaves rather than of a community with a common identity where faceless connector roads seem to weave aimlessly. This type provides low route choice, is usually only residential, has poor access to facilities, high traffic and low walkability.

- Lollipops on a Stick

Figure 13: Lollipops on a stick

The final category, omitted from this study, is the antithesis of the gridiron. It is a tree-like network, where dead ends branch off of a few through streets. There is a minimal amount of road connectivity, with few and large blocks. Nested branches provide access to block interiors. It has a very low number of intersections, route choices and access points. Since this type doesn't form blocks in the conventional sense, namely ones that are good candidates for automatic subdivision, it was excluded from the study.

This classification system is the most widely cited local road network classification system in both qualitative and quantitative studies of urban form (Burton et al., 2011; Rifaat & Tay, 2008). From the above descriptions, it is apparent that each of these categories is defined both by its geometric properties as well as the purpose it was conceived for. Though the road network models estimated in this research didn't capture the entire decision making processes of planners, or their reasons for choosing certain functional forms over others, it was possible to model which forms would function better in which areas based on pre-existing characteristics of the areas themselves. This relationship might be lost on strictly geometric classifications. It was also helpful to have a functional classification system when classifying the sample areas that would become the model units. It proved difficult to differentiate patches of road network based on pattern alone, given that these patterns lie somewhere along a continuum and the differences between them are often subtle. With functional categories, it was possible to interpret what the predominant functions of the streets in a given patch were (i.e.: limiting access or visual interest), rather than classify seemingly haphazard collections of lines.

Given the inherent subjectivity in classification and the importance of accurate classification of modeling units, a classification key was developed inspired by taxonomic keys that help biologists identify the species of an unknown organism (Borges, 1996; Timme, 1991). Here, keys are constructed so that the user is presented with a series of choices about the characteristics of the unknown organisms; by making the correct choice at each step of the key, the user is ultimately led to the identity of a specimen. The two main types of taxonomic keys are the dichotomous key (JFP, 2016) and polyclave or synoptic key (SCFS, 2016). The dichotomous keys consist of a series of paired statements, or termed couplets, that describe some feature of the organism. The statements, or leads, are mutually exclusive. To use the key, the user begins with the first couplet and selects the statement that best fits their organism. This will direct them to another couplet and ultimately provide the identity of the specimen. With the polyclave key, there is a series of characteristics or features, and a list of species associated with each characteristic. The user can enter the key at any point, by selecting a characteristic of their species and then copying down the list of species that possess the feature. Then the user selects another characteristic and eliminates any species not common to both lists. This process continues until the specimen is identified. Polyclave keys have gained popularity because they lend themselves to being computerized.

For the sake of simplicity, a series of dichotomous keys were used to classify the road network sample areas or patches (Appendix A). Most of the ambiguity was between adjacent road network types on the grid-dendritic spectrum, and so a separate key was developed each with only two outcomes representing a pair of these adjacent types. For example, it is often difficult to distinguish between a gridiron patch and a fragmented parallels patch because fragmented parallels are essentially incomplete gridirons. The key for this set of types served to specify how incomplete a grid must be before it is considered fragmented parallels. In contrast, it is unlikely that a gridiron and a loop and lollipop patch would be confused, and so no key was created for discerning these two types. There was, however, a key that served to distinguish a sample area of a uniform type, ultimately included in the sample, from a mixed type, automatically excluded from the sample. In addition, the key that served to distinguish between warped parallels and loop and lollipop also identified New Urbanism types since these tend to contain elements of both. A brief description of each road network type is given to direct the user toward a specific key.

The keys elaborated here are perhaps less complete than taxonomic keys and require the user to exercise more judgment in the process of identifying a specimen. However they serve to distinguish elements that are more ambiguous to begin with, namely road networks that are artificial and likely to be mixed vs. species that are by definition and at least in principle mutually exclusive. The heuristics used in the paired statements of the dichotomous keys were informed by the literature whenever possible (Southworth & Ben-Joseph, 1993, 2003; Southworth & Owens, 1993), however many of them were admittedly subjective and somewhat arbitrary. Given the absence of a complete definition of each road type category from the literature, and the inherent ambiguity in the boundaries of the categories themselves (Marshall, 2005), this was thought to be inevitable. It is hoped that as long as the heuristics were applied consistently for all sample areas, I was doing the best that I could with the expertise that I had. There were five keys in total, included in Appendix A, and they were used to classify all road network patches.

Spatial Autocorrelation

Spatial autocorrelation refers to the tendency for objects closer in space to be more similar. More formally, it is defined as the dependency found in a set of cross-sectional observations over space and occurs when individuals in a population are related through their spatial location (Anselin, 2001). The mechanisms by which this occurs vary depending on the phenomenon being observed. Since no other road network type models have been documented, the following draws from models that predict the timing of conversion of non-urban parcels to residential use, so called land-use change models.

A number of land-use change models have been estimated at the parcel level, and some of these incorporating spatial autocorrelation. Spatial autocorrelation here signifies that a parcel is more likely to convert to residential use if parcels in its vicinity have already converted. This is partly because clustered development can benefit from nearby infrastructure and services. Such studies often employ a variation of the multinomial logit model with a spatial autoregressive lag model that has a spatially weighted dependent variable. Accounting for spatial autocorrelation in multinomial logit models is important since these models assume that the random components of the utilities of the different alternatives are independent (Mohammadian and Kanaroglou 2003). If the modelling units influence each other, then this assumption is violated. In this case, the central limit theorem, that states that the distribution of the average of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution(Griffiths et al., 1993), may not hold. This could lead to false inferences about the significance of parameter estimates as well as to insufficient estimates of the dependent variable. However, calculation of the spatially weighted variable is complex for multinomial logit models, due to the fact that maximum likelihood estimation of the spatially lagged variable coefficients requires multiple inversions of the very large spatial weight matrix. Mohammadian and Kanaroglou (2003) used this technique to model the housing types developers choose to construct. They used a distance decay function to capture potential for interaction between sample units. A Generalized Method of Moments (GMM) estimator can alternatively be used but requires similar inversion of the spatial weight matrix (Pinske & Slade, 1998). Klier and McMillen (2008) propose a spatial logit estimator based on Pinske and Slade's GMM estimator for a binary discrete choice model. Their method was found to accurately identify the coefficient of the spatially lagged dependent variable. Carrión-Flores (2009) extended this method to the multinomial logit model case in a way that doesn't require inversion of a large spatial weight matrix.

Spatial Autocorrelation in Context

For a number of reasons, the above approaches weren't used in this study. Firstly, it was important that the road network type model be easy to estimate so that it might be easily incorporated into an integrated land use-transportation modelling tool. Given my expertise, the above approaches would have required more time than was available, and perhaps this is a good enough reason not to use it in this context. Rather than incorporating a spatially lagged variable into the model, an effort was made to capture spatial dependence through a series of standard binary variables with alternative specific coefficients. These were derived from a spatial weight matrix, calculated according to the method used in Ding (2001) and described further below.

Similar to the procedure for estimating spatial autoregression models, a spatial weight matrix was estimated to define the neighbourhood of influence of a given sample unit. In other words, the area within which other units can be expected to affect the outcome of the unit in question. However, unlike in spatial autoregression models, in this study, the spatial weight matrix was used to calculate an independent binary variable indicating whether a sample unit has at least one other sample in its neighbourhood with the same outcome as itself. This was a simplified approach to accounting for spatial autocorrelation in the model, to help offset the effects of not accounting for it at all. The section below describes the spatial analysis method used for defining the neighbourhood of influence, a necessary input to the spatial weight matrix calculation.

The way the spatial weight matrix is calculated can influence whether or not spatial autocorrelation is detected in the model. The spatial weight matrix is meant to define the sample units that are in a given sample unit's neighbourhood and the degree of the dependence between them. There are two main approaches for defining a neighbourhood, through the adjacency concept or through the distance concept (Ding, 2001). It is difficult to apply the adjacency concept in this study, since the sample areas weren't contiguous but were rather defined by clustering residential building points and then clipping these using buffered major roads. The distance approach is also problematic since the sample areas are of different sizes and have variable dispersion. A fixed distance might overestimate neighbourhoors within dense residential areas where sample areas are closer together and smaller. Conversely, the same distance might underestimate neighbours within suburban areas with sprawling development and larger, more spread out communities. A technique was applied to build the spatial weight matrix that uses a variable search distance that depends on the size of the sample area. The relationship between sample area and search distance is defined in the equation below.

Equation 2: Spatial weight matrix calculation

$$\psi_{ij} = 1 \; if \; d_{ij} \leq b\sqrt{\frac{area_i}{\pi}}, and \; \psi_{ij} = 0 \; if \; d_{ij} > b\sqrt{\frac{area_i}{\pi}}$$

Where ψ_{ij} is a binary variable equal to 1 if a sample area is a neighbor and 0 otherwise, d_{ij} is the threshold distance between neighboring sample areas, b is a constant and $area_i$ is the area of the sample unit in question. The case where b=2 represents an estimate of rooks case contiguity, where sample areas are considered to be neighbours if they share an edge with a given unit. The case where b=3 represents queen's case contiguity, where a unit is also a neighbour if it only shares a corner with a given unit. It was thought that the rook's case would be more applicable here, since road networks that share an edge are more likely to be physically connected and therefore more likely to interact. This spatial weight matrix was computed in Python from the centroids of sample areas.

From this spatial weight matrix, four binary variables were calculated. These captured whether or not a sample unit has at least one neighbour of each of the four road network types. In estimating the model, all four binary variables were included in each utility function. However, only binary variables of the same type were found to be significant within their respective utility functions. This suggests a positive spatial autocorrelation and captures an aspect of it, albeit in a somewhat rudimentary way.

In general, spatial dependence within modelling is captured along two dimensions: 1) the potential for interaction between sample units of similar value and 2) the degree to which this potential interaction translates into actual interaction for the particular phenomena in question. For example, two sets of neighborhoods might be the same distance apart (i.e.: have the same potential for interaction), but their stray cat populations might influence each other to a larger degree than their frequency of ice cream shops (i.e.: phenomena have different actual interactions). The first dimension is often captured in a spatial weight matrix that measures the extent of proximity between sample units of the same type. The second dimension is often captured in coefficients to the weight matrix, called spatial lag coefficients (Carrión-Flores et al., 2009; Mohammadian & Kangarolou, 2003). In the simplified models estimated here, an attempt was made to capture the second dimension within the alternative specific coefficients of the binary variables. However these binary variables only captured whether or not a sample area had at least one neighbour with road network of the same type. As a result, the first dimension of spatial dependence, or potential interaction between sample units of the same type, was an oversimplification. In reality, the odds of a road network being a given type should increase with the number or area of neighbours of that type. However this dimension wasn't incorporated due to time constraints. The method used in this research was thought to be adequate for capturing spatial dependence simply, without having to go through the complex process of estimating spatially lagged variables. A main goal for this model was to make it easy to integrate into urban simulation tools, so it was important that it not be too difficult to estimate.

It is worth noting that for Manuscript 1 (Testing Block Subdivision Algorithms on Block Designs), sites of the same type were included as replicates of the same experiment and so each of the sites was tested in isolation from the others. As a result, there was no need to account for spatial autocorrelation between them. Within a site, spatial autocorrelation between parcel metrics was considered implicitly in the block type classification, as outlined in the Manuscript. The category encompassing more uniform parcel sizes could be expected to have higher spatial autocorrelation than the more variable parcel size category. Incorporating a metric of spatial autocorrelation, such as Moran's-I, in a comparison of observed and generated parcels was beyond the study scope.

Manuscript 1 Testing Block Subdivision Algorithms on Block Designs

Context

This section includes a co-authored paper that builds on recent progress in the area of automatic block subdivision. The paper extends the methods previously used in testing automatic block subdivision algorithms so as to provide a basis of comparison between them. The methods are also designed to produce statistically reliable conclusions, so that they might serve as actionable guidelines on when to use which algorithms.

In the following manuscript, published in the Journal of Geographical Systems (Wiseman & Patterson, 2016b), I had the role of lead author. Zachary Patterson, the paper's second author, formulated the research question and analysis methods. Initial results of this research were presented at the Transportation Research Board's 94th annual meeting in Washington, DC, in January 2015. The methodology was then updated to include repetition of the experiment on a larger number of sites within Montreal and to incorporate a more common local road network classification system.

The findings indicate that different types of blocks have different better algorithms suited to subdividing them. They also indicate which of these algorithms is better for each block type in a common classification of residential blocks. Since these finding are based on a sufficiently large sample size, they can be considered statistically reliable insofar as the idea of a "better algorithm" is defined in the paper.

The results, included in the Results and Conclusions sections of the paper, could be applied either to generating parcels in future residential areas or to generating missing current parcel data for integrated modelling tools. They are as follows: The Oriented Bounding Box algorithm is better for Gridiron and Fragmented grid sites with more uniform parcel sizes as well as Warped non-uniform sites. The Generalized Parcel Divider 1 is better for Gridiron non-uniform sites. The Straight Skeleton algorithm is better for Loop and lollipop sites with uniform parcels, Fragmented non-uniform and Warped uniform sites. It also produces more statistically similar parcel shapes and street access, whereas the Oriented Bounding Box algorithm produces more statistically similar average parcel areas and widths overall.

The paper appears below as it was accepted in the Journal of Geographical Systems in November 2015.

Abstract

Integrated land use-transportation models predict future transportation demand taking into account how households and firms arrange themselves partly as a function of the transportation system. Recent integrated models require parcels as inputs and produce household and employment predictions at the parcel scale. Block subdivision algorithms automatically generate parcel patterns within blocks. Evaluating block subdivision algorithms is done by way of generating parcels and comparing them to those in a parcel database. Three block subdivision algorithms are evaluated on how closely they reproduce parcels of different block types found in a parcel database from Montreal, Canada. While the authors who developed each of the algorithms have evaluated them, they have used their own metrics and block types to evaluate their own algorithms. This makes it difficult to compare their strengths and weaknesses. The contribution of this paper is in resolving this difficulty with the aim of finding a better algorithm suited to subdividing each block type. The proposed hypothesis is that given the different approaches that block subdivision algorithms take, it’s likely that different algorithms are better adapted to subdividing different block types. To test this, a standardized block type classification is used that consists of mutually exclusive and comprehensive categories. A statistical method is used for finding a better algorithm and the probability it will perform well for a given block type. Results suggest the Oriented Bounding Box algorithm

performs better for warped non-uniform sites, as well as gridiron and fragmented uniform sites. It also produces more similar parcel areas and widths. The Generalized Parcel Divider 1 algorithm performs better for gridiron non-uniform sites. The Straight Skeleton algorithm performs better for loop and lollipop networks as well as fragmented non-uniform and warped uniform sites. It also produces more similar parcel shapes and patterns.

Introduction

Increasingly, planning for transportation infrastructure investments is being integrated with land use planning. Successful integrated planning, it is hoped, can provide competitive alternatives to automobile use and thereby reduce congestion on roadways and greenhouse gas emissions. Integrated land use-transportation modeling is increasingly a method chosen to help forecast transportation demand in this context. It is used because it explicitly accounts for the effect of transportation network performance on population and employment distribution, and thereby overall transportation demand (Borning et al., 2008; Waddell et al., 2007).

Within the field of integrated modeling, there has been a trend toward representing phenomena at increasingly finer spatial units. The finest unit used in integrated modeling to date is the parcel. This is also the spatial unit to which all other data (households, jobs, buildings, etc.) are linked (Waddell, 2009). Increasingly fine spatial representation has been motivated by understandings in complex systems, behavioral theory and statistical aggregation bias, and made possible by advances in Geographic Information Systems (Xie & Batty, 2003). This increasingly disaggregated spatial representation presents many opportunities but also many challenges within the field of integrated modeling.

For one, parcel-level models require a large amount of highly detailed data (e.g. parcel data), which can be unavailable or incomplete (Schirmer, 2010) and whose absence can delay running simulations with integrated models (Patterson & Bierlaire, 2010). Moreover, future parcel data is by definition not available, but clearly relevant to forecasting future locations of households and employment with parcel-based models. Recently, a number of algorithms have been developed that can help with both of these problems – so-called *parcel generation or block subdivision* algorithms. Some of these algorithms have been implemented in City Generation Engines, such as Esri's CityEngine and Synthicity's UrbanCanvas, as visualization tools integrated with simplified behavioral models (Vanegas et al., 2012). Others have been developed as stand-alone applications and could potentially be linked to integrated models (Dahal & Chow, 2014; Wickramasuriya et al., 2011). Both sets of tools can be used to generate missing data from parcel datasets needed as inputs to parcel-level integrated models.

Given the different approaches that block subdivision algorithms take it's likely that different algorithms are better adapted to subdividing different block types. This study aims to find the better algorithm suited to subdividing each one. It serves as a comparative analysis of the performance of previously developed algorithms. In evaluating these algorithms, their developers tested them on a few sites each from their own informal site type classifications, some of which encompassed only a subset of possible blocks. The developers also used their own metrics and tests for the evaluation. While each of the developers provided evidence that their block subdivision algorithms perform well on a given block type, the inconsistent block types, comparison methods and metrics used in their evaluation make it difficult to compare the algorithms with each other. As such, in this paper a consistent methodology is used that provides a quantitative basis for comparing between them so that an optimal block subdivision algorithm for each block type can emerge. This study also introduces a comprehensive and mutually exclusive block type classification where previous studies each used their own, posing another difficulty in comparing their results.

To this end, parcels are generated using three algorithms on the same set of blocks from a parcel database for Montreal, Canada. The resulting synthetic parcels are compared statistically to the real-world parcels associated with the blocks from the database. The comparison is based on indicators used in the past by algorithm developers to evaluate their own algorithms. The test is repeated on ~30 sites per block type and the proportion of passed statistical tests for each algorithm is recorded. Ideally, the use of statistics and a comprehensive block type classification can be applied to define a set of rules for best matching block subdivision algorithms to block types. As such, this paper provides guidelines to users of block subdivision algorithms for preparing missing current or future parcel data for integrated models. The methodology can also be reused for testing newly developed block subdivision algorithms to make it clearer which block type(s) the algorithm ought to be used on. In the longer run, the goal is to use these findings to define a set of rules to be incorporated into an automatic block subdivision program: one that automatically executes the appropriate algorithm for the block type to be subdivided. This could then be incorporated into a future integrated model that simulates land subdivision processes.

The literature review describes previous developments in block subdivision algorithms and in methods used to test their performance. The methodology section outlines the block type classification used, as well as the statistical tests used to evaluate the algorithms' performance. The results show how each algorithm performed on each block type, and a discussion section attempts to outline a set of relationships that define which algorithm is better for which block type. The paper is finished with a few additional remarks and directions for future research in the conclusion.

Literature Review

This literature review has two parts. The first describes the literature relating to the history and development of block subdivision algorithms. The second describes how the success of block subdivision algorithms has been evaluated in the past.

Block Subdivision Algorithms

The idea to use computer programming to model patterns in the urban environment originates with Parish and Müller (2001). They adapted the use of L-systems, which had been successfully used to generate realistic trees in computer graphics programs, to the generation of road and highway networks. Parish and Müller (2001) also demonstrated how such systems could be extended to subdivide land into lots, and create the appropriate geometry for buildings on the lots. Their land subdivision algorithm recursively subdivides blocks along the longest pair of approximately parallel edges, until parcel sizes are under a user-specified threshold area. One of the disadvantages of this algorithm is that parcels with no street access are deleted, leaving holes in the middle of blocks. This method also produces some irregularly shaped parcels. It is worth noting that this appears to fit the description of the block subdivision algorithm used in the integrated modeling system, PECAS (Hunt and Abraham 2009), although details on the exact algorithm used aren't available.

To Parish and Müller's recursive algorithm, Weber et al. (2009) incorporate a varying maximum area threshold depending on a parcel's land use type. More recently, Vanegas et al. (2012) implemented an Oriented Bounding Box (OBB) algorithm as a method for recursively splitting street blocks into parcels. Use of the OBB produces more regularly shaped parcels, and ensures a maximum number are oriented parallel to an adjacent street. Their algorithm also tries to ensure street access by splitting the bounding box along either the longest or widest edge.

Vanegas et al. (2012) also introduced the Straight Skeleton (SS) algorithm. It is based on the straight skeleton shape, formed by collapsing the edges of a polygon inward and tracing the intersection points of each set of edges according to Aicholzer's motorcycle algorithm (Aicholzer & Aurenhammer, 1995). This shape is then split at regular intervals determined by a user-specified parcel width. Diagonal edges are then shifted to be perpendicular to roads. This algorithm ensures all parcels have street access. Similar to the Straight Skeleton algorithm, the Offset Subdivision algorithm contains an additional parameter specifying the distance to set the far edge of the parcels from the street. This produces a perimeter-block design; one whose parcels surround the outer edge of the block and contain a large central parcel typically occupied by a school or park. All algorithms developed by Vanegas et al. (2012) use block-level descriptive parcel parameters specified by the user, such as minimum parcel area and width.

The above algorithms were developed for City Generation Engines, such as CityEngine and UrbanCanvas. City Generation Engines can structure the forecasts of integrated models, namely future data on population, jobs and buildings, into plausible 3D cities. They also support manual editing of an urban system for more localized planning (Vanegas et al., 2012). In this context, development in an area is simulated according to a predefined process. First a road network is grown according to a road generation algorithm, then a block subdivision algorithm is executed to generate parcels, and finally a rule file determines the type of building to place on each parcel. All these steps can optionally be executed on top of a digital elevation model, so that the generated shapes adjust to the topography. In contrast, Wickramasuriya et al. (2011) developed a single algorithm to generate both roads and parcels within a block. Their program overlays four different orthogonal grids onto an area of land, selects the one that maximizes the number of parcels or minimizes the length of roads, and then clips the grid to the land's boundaries. Such a system serves as a stand-alone land subdivision tool for use by land developers or urban planners. Being modular and open-source, it could also be incorporated into integrated land use transportation modelling systems.

Building on the work of Wickramasurya et al. (2011), Dahal & Chow (2014) developed the ArcGIS Parcel Divider python toolset, which creates new roads and parcels on previously undeveloped tracts of land. The toolset contains six different algorithms, two of which can be used on any block type and the others on blocks with specific shapes such as T, L, or cul-de-sacs. Of the two more general algorithms, Generalized Parcel Divider 1 is designed for any block type while Generalized Parcel Divider 2 "...tends to generate block pattern with Manhattan-style street network." (p. 6)

Generalized Parcel Divider 1 (GPD1) uses a combination of recursive binary subdivision and grid drawing. First the algorithm recursively subdivides a land tract's oriented bounding box until the width of the bounding box is <= 2.5 times the user specified average parcel length. In a later step, the contours of this final bounding box are turned into roads, ensuring all parcels it contains have street access. A series of grid lines are drawn within the bounding box, perpendicular to its longest edge and at intervals determined by the user-specified average parcel width. Grid lines perpendicular to the shortest edge of the bounding box are then drawn at intervals determined by the user-specified average parcel length. Undersized parcels are merged and the resulting grid is then clipped to the boundaries of the input land area. Generalized Parcel Divider 2 is similar, but skips straight to the grid drawing steps. In so doing, it is nearly identical to Wickramasuriya et al.'s (2011) algorithm.

Testing Block Subdivision Algorithms

In the presentation of their algorithms, developers typically test their algorithm's ability to generate parcel patterns similar to observed, real-world counterparts. The method used across all studies to determine whether an algorithm performs well, is to subdivide a block with the algorithm and then to compare the resulting parcels to their observed counterparts in a parcel dataset. In the past, the algorithm developers have, however, used their own informal block type categories some encompassing only a subset of possible block types (Wickramasuriya 2011, Vanegas et al. 2012, Dahal & Chow 2014). Each study has also used its own methods and ad-hoc indicators for comparing the two sets of parcels. This has made it difficult to compare their performance, prompting the question raised in the current research of which algorithm is best suited to subdividing a given block type.

With respect to site selection, each study tested their algorithms on a few site types, chosen to represent distinct types of areas with block characteristics that are expected to affect the algorithms' performance. Each study has also selected one or two examples of each site type on which to test the algorithms. Vanegas et al. (2012) selected three site types based on land use type and density, block shape and parcel variability and tested their algorithms on one large site of each type. One site type was described as "...a mixed-use suburban area composed of rectangular blocks with both straight and curved edges...The set of blocks exhibits significant variability in both the area, the aspect ratio, and the minimum width of the parcels." (p. 9). Wickramasuriya et al. (2011) classified sites into two types: one with parallel road and parcel orientations and uniformly shaped parcels, and the other with varying road and parcel orientations and shapes. Three examples of the former and one example of the latter site type were selected for testing. Since the toolset of Dahal & Chow (2014) has some algorithms designed for particular block shapes (i.e. L-, and T-shapes) they test these algorithms on examples of blocks with these specific characteristics. With respect to their two more general algorithms (GPD1 and GPD2), GPD1 was tested on one site of an irregular and another of a regular shape, while GPD2 was tested visually on a site with an irregular shape.

With respect to the comparison of simulated to real-world parcels, Wikramasuriya et al. (2011) used t-tests and correlation coefficients to statistically compare the number of lots and mean lot sizes of the two sets of parcel distributions. They also compared the standard deviations of the sets of distributions and their mean shape index (MSI), a variation on the shape index metric used in FRAGSTATS (McGarigal & Marks, 1994), using a standard error calculation. Vanegas et al. (2012) pooled the parcels generated by the OBB, SS and Offset Subdivision algorithms into a single sample, before comparing them to the observed parcel distributions. The statistical component of their method involved overlapping the frequency distributions of metrics of the simulated and real-world parcels to conduct a visual comparison of their similarities. Dahal & Chow (2014) compared the total number of lots and mean lot size of each site using a calculation of standard error ($[Z_{modeled} - Z_{reference}]/Z_{reference} \times 100$). They also compared counts of parcels without egress and ones with access to more than one street.

Given the different methods to test the algorithms, each study also came to different types of conclusions about its own algorithms. Vanegas et al. (2012) found that their algorithms produced parcels with similar frequency distributions of metrics as those observed, as well as similar spatial distributions of metrics depicted in color coded maps. Furthermore, all generated parcels were found to have dimensions and aspect ratios that were adequate for containing buildings. The SS algorithm always produced parcels with an egress while the OBB algorithm sometimes didn't. Wickramasuriya et al. (2011) found that their algorithm generated parcels with statistically similar area distributions and counts for the first, more regular block type, but failed to do so for the second, more irregular block type. On the first block type, it also produced parcels of highly irregular shapes and sizes adjacent to curved or irregular block boundaries. Dahal & Chow (2014) found their algorithms to produce unrealistically uniform parcel shapes and sizes, with similar widths but longer lengths than those observed. The total number of lots was similar for all sets of parcels. All generated parcels had access to at least one street.

The above research focused on developing algorithms for generating parcels and also tested their performance on different block types. Each study, however, used a different method for classifying blocks, some encompassing only a subset of possible blocks, as well as a different method of testing its algorithms' performance. They also only tested their respective algorithms on one or two sites of each block type. As a result, it is difficult to perform a meta-analysis of the results, to determine whether or not there is an algorithm that performs better than others, on any given block type. To fill in this gap, this study selects three of the most recently developed general algorithms described in the literature, and uses them to generate parcels on the same set of real-world blocks. The algorithms are applied to different blocks based on a comprehensive block type classification, and the simulated blocks are compared to real-world blocks using a comprehensive set of indicators and statistical tests based on those used in the literature before. Approximately 30 examples of each block type from this classification system were selected as test cases on which to apply the three algorithms, thus allowing both a common basis of comparison and more statistically robust results than in previous literature.

Methodology

The general methodology adopted was to apply three block subdivision algorithms to the same sites, selected as examples of the block type categories, and to perform statistical tests comparing the characteristics of the simulated parcels with their real-world counterparts. The rest of this section describes: how the algorithms themselves and required input parameters were chosen, the block type classification used, and how the sites were selected; what parcel characteristics were compared and what statistical tests used to compare them.

Selection of Algorithms

In order to conduct a comparison of block subdivision algorithms, a number of candidate algorithms were available. As described in the literature review, there were: the original Parish and Müller algorithm (2001) and its more recent incarnation (Weber et al., 2009); the algorithm used in the PECAS model (Hunt & Abraham, 2007); Vanegas et al.'s oriented bounding box (OBB) and straight skeleton (SS) algorithms (2012); Wickramasuriya et al.'s algorithm (2011); and Dahal and Chow's (2014) algorithms. When selecting the candidate algorithms, a number of criteria were used. First, an explicit and detailed description of the algorithm needed to be documented in the academic literature. Because we could not find an explicit and detailed description of the algorithm used in PECAS, we did not include this algorithm. Second, only "general algorithms" were considered in the analysis. The term "general algorithm" is used to distinguish them from algorithms specifically designed to subdivide a particular type of block. For example, the Offset Subdivision algorithm (Vanegas et al., 2012) reproduces a perimeter block design and one could safely assume that it is the best algorithm for this type of block. Similarly, the Cul-de-sac Creator algorithm (Dahal & Chow, 2014), could be expected to best subdivide blocks at cul-de-sacs. The more general algorithms were chosen for this study because there is a degree of ambiguity about their performance in relation to each other and to different block types. As a result, a few of Dahal and Chow's algorithms, namely, Cul-De-Sac Creator, L-Shaped Parcel Divider, T-Shaped Parcel Divider, Multi Family Parcel Divider (for multi-family housing lots), Divider with Inner Roads (for blocks with inner looped roads) were not considered in the analysis. Finally, algorithms with documented weaknesses were also removed from consideration. This eliminated Parish and Müller's (holes within blocks and irregularly shaped parcels), Wickramasuriya et al.'s and Dahal and Chow's GPD2 (parcels with irregular shapes and orientations within irregular blocks) algorithms. This left the OBB, SS and GPD1 algorithms as candidates to test.

The application of the OBB and SS algorithms was done through ESRI's implementation in CityEngine. Application of the GPD1 algorithm was done in ArcGIS with the Parcel Divider Toolset after modifying the code to prevent it from generating roads within input blocks and enabling it to read subdivision parameters that vary from block to block. The road generation functionality was disabled to make it easier to evaluate in relation to the other algorithms (that subdivide blocks without creating roads, and which indeed require a road network as an input) and to specific block types. For example, since this algorithm can only generate straight roads, enabling this function would have changed the two curvilinear block types.

Description of Input Parameters

Each algorithm has a different set of input parameters whose values must be specified by the user. For the purpose of this study, input parameters were classified into two types: 1) those which determine the characteristics of the resulting parcels, i.e.: deterministic parameters, and 2) those which constrain the characteristics of the resulting parcels, i.e.: constraint parameters. For example, the width parameter functions as a constraint in the OBB algorithm and as a deterministic parameter in the GPD1 and SS algorithms. In the case of OBB, the width parameter determines the width below which no more subdivision occurs. This leads to many possible width values of the resulting parcels, and so this parameter can be said to be non-deterministic. On the other hand, for SS and GPD1, the input width determines the width of the resulting parcels (before slivers are merged with non-sliver parcels), by drawing a split line at intervals equal to it. To account for this difference, the input width for the OBB algorithm was set as:

Equation 3: Minimum width input parameter for OBB

$$W_{min} = W_{avg} - 1 \times sd_w$$

Where W_{min} is the minimum width a parcel can take, W_{avg} is the average width of parcels for the block, and sd_w is the standard deviation of parcel widths for the block. For the SS and GPD1 algorithms, the average parcel width by block is the width input value.

This convention was varied slightly for loop & lollipop road networks, with curved blocks and highly irregular parcel shapes. Since the parcel width parameter is difficult to compute, let alone conceptualize, the width of the Minimum Bounding Rectangle of the parcel is used instead. For curved or irregular parcels, the Minimum Bounding Rectangle width appears to be an overestimation of the actual street frontage of the parcel. To compensate for this, for blocks with extreme curves or irregular angles (loop & lollipop) the minimum width parameters were set at 1 standard deviation below the others. That is, for SS and GPD1:

Equation 4: Loop and lollipop width input parameter for SS and GPD1

$$W_{input} = W_{avg} - 1 \times sd_w$$

Equation 5: Loop and lollipop width input parameter for OBB

$$W_{input} = W_{avg} - 2 \times sd_w$$

Where W_{input} is the width input value, W_{avg} is the average width of the bounding box for all the parcels in the block, and sd_w, is the standard deviation of the average width of the bounding box for all the parcels in the block.

For the Straight Skeleton and Oriented Bounding Box algorithms, the width input value is enhanced by a split irregularity parameter, which displaces the parcel's split line from its default position to create less uniform parcels. This displacement distance is sampled proportionally from a distribution defined by a mean equal to the algorithm's input width and a variance equal to 3 times the irregularity parameter. To populate this parameter for the SS algorithm, the blocks with smallest and largest width standard deviations were determined for the entire sample (0 and 12 respectively). The blocks' width standard deviations were then converted to a scale between 0 and 1; the range of this input parameter. The maximum irregularity parameter value (ie: 1) was divided by the maximum width standard deviation (ie: 12) to give a conversion factor of 0.083. This factor was then multiplied by each of the blocks' standard deviations to give their input irregularity values, as follows:

Equation 6: Irregularity input parameter

$$\omega_{input} = \left(\omega_{max}/sd_{\omega_{max}}\right) \times sd_{\omega}$$

Where ω_{input} is the input irregularity, ω_{max} is the maximum input irregularity (1), $sd_{\omega_{max}}$ is the maximum width standard deviation (12), and sd_{ω} is the block's width standard deviation. An exception was made for block types with uniform parcel sizes, where the irregularity was set at 0. By definition, these parcels are highly regular and most of the variability in parcel width appears to be a result of irregularities in block shape, rather than an intentional design feature. Further varying these widths in the synthetic parcels through an irregularity parameter was found to have overestimated the width variance.

Since the OBB algorithm's width parameter is a constraint, there is already a degree of variability in the parcel widths it produces. As a result, the split irregularity parameter was generally set at 0, unless the standard deviation of the block's parcel widths was >= 10, in which case it was set at 0.03.

The other input parameters, namely, minimum and maximum area, are constraint parameters and the following maximum and minimums are used:

Equation 7: Minimum and maximum area input parameters

$$A_{max/min} = A_{avg} \pm 2 \times sd_A$$

Where $A_{max/min}$ are the maximum and minimum values (e.g. of parcel area), A_{avg} is the average value for the parcels in a block, and sd_A is the standard deviation of the average value for the parcels in a block. One exception was for the GPD1 minimum area parameter that was set at:

Equation 8: Minimum area input parameter for GPD1

$$A_{min} = A_{avg}/2$$

Where A_{min} is the minimum parcel area and A_{avg} is the average area of parcels in the block. The minimum area was left open in the paper and it was thought that this option would result in a majority of parcel areas around the mean. It is worth noting that these methods for arriving at the deterministic and constraint parameters were chosen from several trials, based on visual and sometimes statistical analysis. Finally, there were some required inputs (e.g. length in the GPD1 algorithm) for which there was no ambiguity in the value of the parameter required. For example, the GPD1 algorithm required an average value for the length of the parcel, but this is consistent with the length of the parcel's bounding box. As a result, average parcel length could be used. The *Force street access* parameter required by the OBB algorithm was set to always ensure street access, a parameter value of 1, in light of the general requirement that all residential parcels have access to a road. These input parameters are summarized for each algorithm in the tables below under the actual parameter names used in their respective programs.

Table 2: Input parameters used in the SS algorithm

Input parameter	Type	Road network type				Rationale
		Gridiron	Fragmented Grid	Warped Grid	Loops & Lollipops	
lotAreaMin	Constraint	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	Used in paper[a]
lotWidthMin	Deterministic	W_{avg}	W_{avg}	W_{avg}	$W_{avg} - 1 \times sd_W$	Used in paper[a], Experimentation
irregularity	Constraint	$0\ or\ 0.038 \times sd_W$	$0\ or\ 0.038 \times sd_W$	$0\ or\ 0.038 \times sd_W$	0	Trial and Error

[a]Vanegas et al. (2012)

Table 3: Input parameters used in the GPD1 algorithm

Input parameter	Type	Road network type				Rationale
		Gridiron	Fragmented Grid	Warped Grid	Loops & Lollipops	
Width	Deterministic	W_{avg}	W_{avg}	W_{avg}	$W_{avg} - 1 \times sd_W$	Used in paper[a], Experimentation
Length	Constraint	L_{avg}	L_{avg}	L_{avg}	L_{avg}	Used in paper[a]
AvLotSize	Constraint	A_{avg}	A_{avg}	A_{avg}	A_{avg}	Used in paper[a]
sizeTo Merge	Constraint	$A_{avg}/2$	$A_{avg}/2$	$A_{avg}/2$	$A_{avg}/2$	Experimentation

[a]Dahal & Chow (2014)

Table 4: Input parameters used in the OBB algorithm

Input parameter	Type	Road network type: Gridiron	Fragmented Grid	Warped Grid	Loops & Lollipops	Rationale
lotAreaMin	Constraint	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	$A_{avg} - 2 \times sd_A$	Used in paper[a]
lotAreaMax	Constraint	$A_{avg} + 2 \times sd_A$	$A_{avg} + 2 \times sd_A$	$A_{avg} + 2 \times sd_A$	$A_{avg} + 2 \times sd_A$	Used in paper[a]
lotWidthMin	Constraint	$W_{avg} - 1 \times sd_W$	$W_{avg} - 1 \times sd_W$	$W_{avg} - 1 \times sd_W$	$W_{avg} - 2 \times sd_W$	Used in paper[a], Experimentation
forceStreet Access	Deterministic	1	1	1	1	Experimentation
Irregularity	Constraint	[0,0.3]	[0,0.3]	[0,0.3]	0	Experimentation

[a]Vanegas et al. (2012)

Block Type Categorization and Test Site Selection

The block type classification used in this study is based on a classification of local road networks developed by Southworth & Owens (1993). This classification has since been used in a number of studies where local road network type categories are required, both quantitative (Burton et al., 2011; Rifaat & Tay, 2008) and qualitative (Garde, 2008; Sandalack et al., 2013; Tasker-Brown & Pogharian, 2000). Southworth & Owens classified residential road networks into five categories at the highest level. These include: gridiron, fragmented parallels, warped parallels, loops and lollipops and lollipops on a stick. The categories are based on the evolution of planning paradigms for local road networks from the 1900s – 1980s and represent a progression from grid, highly connected types to dendritic, highly disconnected types. Schirmer (2010) developed a classification of parcel patterns including two categories for residential blocks: Residential 1, containing parcels of more uniform size and Residential 2, containing parcels of variable sizes. Taking into account these different classifications, the block type classification adopted in this study consisted of two dimensions. The first identified the road network type that defined a group of blocks and determined the range of possible block shapes within it. The second dimension identified the residential block type, namely, Residential 1 (with uniform parcel sizes) or Residential 2 (with variable parcel sizes). Figure 14 shows the seven block types and example sites selected from Montreal's parcel database to represent them. To select the final sites used in this study, all residential road network patches within the CMM were classified into one of the four road network types or else got assigned to a mixed category. Next, 30 test sites were selected for each road network type to encompass blocks most characteristic of that category. Parcel pattern types were identified from a parcel database. The fifth local road network type, namely, lollipops on a stick, was omitted from the study since it doesn't consist of enclosed blocks surrounded by roads; the kind of shape that existing algorithms are capable of subdividing.

Description of Comparative Tests

The simulated parcels were compared with their observed counterparts on the basis of four different metrics that are the most commonly used in the literature (see literature review): area, shape index (SI), width and egress (egrs). Area was chosen as an indicator of parcel size, SI as an indicator of parcel shape (FRAGSTATS: McGarigal and Marks 1994), width as an indicator of the amount of street frontage of the parcels and egress as an indicator of whether or not a parcel has street access. The t-test was used to determine whether or not the mean area, SI and width metrics of the simulated parcels were statistically similar to the mean metrics of their observed counterparts. The t-test seemed like the most appropriate test and was also consistent with the literature. The Kolmogorov-Smirnov test was used to test whether or not the distributions of metrics of the simulated parcels were statistically similar to the distributions of their observed counterparts. The Fisher's exact test was used to determine whether the proportion of parcels with egress was the same for both observed and simulated parcels. The Fisher's exact test permits testing proportions that are close to or equal to 1, which is often the case with street access.

The distributions of observed and simulated parcel metrics weren't generally normally distributed, yet the t-test is technically designed to compare two t-distributions. However, for any sufficiently large sample ($n > 30$), the central limit theorem states that the mean of a randomly sampled variable drawn from any distribution will itself be normally distributed (Griffiths et al., 1993). Furthermore, this distribution of means will have the same mean as the original sample itself. Therefore, the t-test is used to compare the means of these two distributions.

The Kolmogorov-Smirnov test is a non-parametric test, useful in determining whether two random variables could have been drawn from the same distribution. It computes the empirical cumulative distribution function (ecdf) of a sorted set of sample values, to demonstrate the percentage of the data below each of the values (Crawley, 2013). The test statistic, or D value, gives the greatest deviation in percentage below a given value between the two ecdfs being compared. Where a distribution of metrics has duplicate values or ties, the KS-test generates a warning in R that the p-value is approximate. As a result, a bootstrap version of the KS-test (ks.boot), developed by Jasjeet S. Sekhon at UC Berkeley, was used that returns exact p-values in all cases.

The Fisher's exact test is used to determine if there are nonrandom associations between two categorical variables (Conover, 1999). Observed data is categorized by variable level and counts associated with each category are arranged in an n by n matrix. The test then determines if the relative proportions of one variable are independent of the second variable. It does this by calculating the conditional probability of getting the actual matrix given the particular row and column sums. Unlike other tests of its kind, it gives an actual p-value and can be used with very small datasets or where the cell proportions are close to or equal to zero and one.

This study takes a sample of 30 sites, each one containing at least 30 parcels, and compares each site individually to its observed counterpart. Aggregating the results over the 30 sites yields a proportion of sites where the null hypothesis wasn't rejected. That is, where the algorithm didn't produce a mean metric that's outside the 95% confidence interval of the observed mean metric. In the context of this study, this is considered to be an indicator of the probability with which an algorithm will perform well for a given site type. By comparing the probabilities for each of the algorithms for a given site type, a "better" algorithm for the set of sites of that type emerges. This improves upon previous methods for testing block subdivision algorithms, in which only one or two sites of each type were used as test cases. It is thought that testing on 30 sites per block type gives a better picture of how well the algorithms perform on sites of that type.

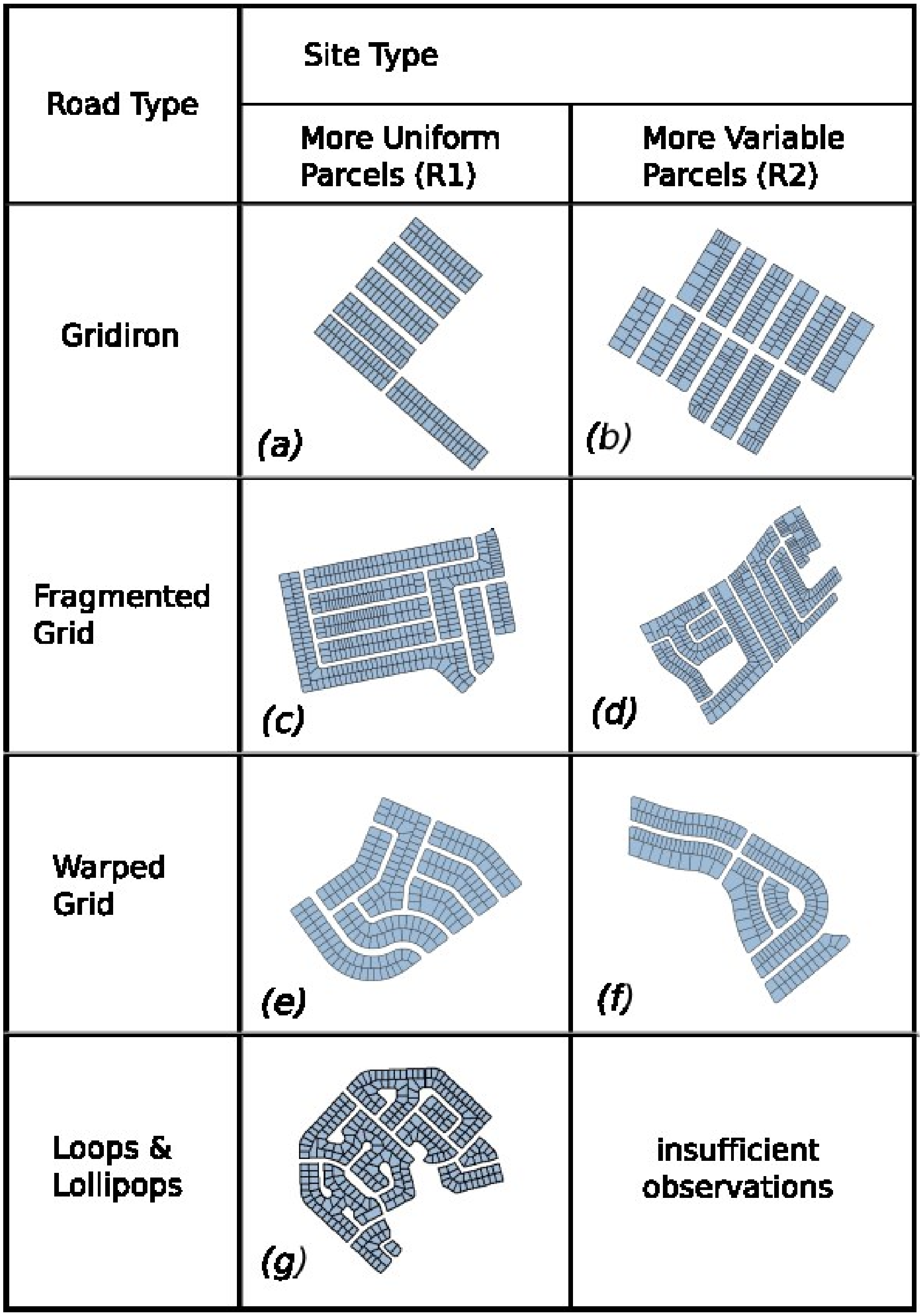


Figure 14: Block types and sites from Montreal's parcel database, Wiseman & Patterson (2016)

Data

The observed parcels, to which the simulated parcels are compared, come from the 2009 cadastral data of Québec purchased from the Ministère des Ressources naturelles et de la Faune (2009) with the funding from the Canadian Foundation for Innovation. The database contains parcel shapefiles that cover the entire territory of the Montreal Metropolitan Community (Communauté métropolitaine de Montréal - CMM), excluding some rural areas where data were missing. A road network by DMTI (2013) was overlaid onto the parcel fabric and used to identify sites from the four road type categories.

The blocks were derived from the most accurate input data available under usual circumstances, namely, the presence of a road network. The road widths were estimated from the number of lanes multiplied by 3.5 m (3.83 yd), a common lane width. The blocks were created in CityEngine by filling in the negative space of the road network. In total 222 sites were selected consisting of between 30 and 35 sites per type, each site containing 3-9 blocks.

Results

The null hypothesis of the t-test is that the difference in mean area (or other metric) between the simulated parcels and observed parcels is 0. A p-value less than 0.05 implies that the two means are different and that the algorithm failed to reproduce parcels with statistically similar mean metric. In contrast, a p-value greater than 0.05 doesn't indicate the two means are the same, only that they cannot be concluded to be different. The null hypothesis of the ks-test is that the distributions of metrics for both sets of parcels were sampled from populations with identical distributions. Again, a p-value less than 0.05 implies that the two parent distributions are different and the algorithm failed to produce parcels with a statistically similar distribution of a given metric. The null hypothesis of the Fisher's exact test is that whether or not a parcel has egress is independent of whether or not it is simulated or observed. That is, the proportions of simulated and observed parcels with egress are the same. Here a p-value less than 0.05 implies that the proportion of parcels with egress does depend on whether or not the parcels are observed or synthetic, and hence that these proportions are different.

The results of these statistical tests, aggregated by site type, can be found in Table 5 and Table 6. Here, each row represents a different site type and each column another algorithm/metric combination. The numbers represent the proportion of non-rejected null hypotheses out of total number of t-tests and Fisher's exact tests for Table 5 (or ks-tests for Table 6). Since the numbers are aggregated by site type, the total number of tests is equal to the number of selected sites per site type and is between 30 and 35. The shaded values indicate the algorithm with highest proportion of non-rejected null hypotheses for a given metric and site type (grey). Ties are shaded in light grey. For example, in the first entry in Table 5- parcel areas for the gridiron, uniform site types produced by the Straight Skeleton algorithm- 3% of the sites had synthetic parcel areas whose average couldn't be shown to be statistically different from their observed counterparts. Here, 100% of sites had proportions of parcels with egress that couldn't be shown to be statistically different from the observed proportions. Similarly, for the same algorithm and metric, but for the gridiron, non-uniform site types, 7% of the sites had average synthetic parcel areas that couldn't be shown to be statistically different from their observed counterparts in the parcel database. For the gridiron non-uniform site types, 80% of the sites had average mean shape indices (si) that couldn't be shown to be different from their observed counterparts. Furthermore, this was the highest proportion for this site type and metric so it was highlighted in grey.

The results of the ks-tests are included in Table 6 and they indicate that the algorithms don't generally reproduce statistically similar distributions of metrics to their observed counterparts. For example, all sites of type gridiron, uniform parcels subdivided by the Straight Skeleton algorithm were found to have statistically different area distributions to their corresponding observed sites. The GPD1 algorithm also produced parcels whose area distributions were always statistically different from their realistic counterparts in all gridiron uniform sites. Again, the highest proportions of non-rejected null hypotheses for each site type and metric are shaded in grey.

Table 5: Proportion of non-rejected null hypothesis for t-Tests and Fisher's exact tests by Metric, Algorithm and Site Type

Site type	SS				GPD1				OBB			
	area	width	Si	egrs	area	Width	Si	egrs[b]	area	width	Si	Egrs
Gridiron - R1	0.03	0.13	0.17	1.00	0.03	0.17	0.13	0.33	0.20	0.30	0.30	1.00
Gridiron - R2	0.07	0.40	0.80	1.00	0.13	0.53	0.77	0.61	0.40	0.50	0.47	0.93
Fragmented - R1	0.03	0.09	0.27	1.00	0.00	0.09	0.30	0.17	0.27	0.21	0.15	0.82
Fragmented - R2	0.12	0.33	0.67	0.97	0.12	0.61	0.52	0.43	0.30	0.24	0.27	0.73
Warped - R1	0.00	0.15	0.24	1.00	0.00	0.09	0.18	0.11	0.18	0.29	0.12	0.74
Warped - R2	0.03	0.00	0.60	1.00	0.00	0.17	0.43	0.19	0.20	0.53	0.17	0.80
Loops, lollipops-R1	0.09	0.84	0.03	1.00	0.22	0.06	0.00	0.09	0.72	0.00	0.19	0.78

[a] Shaded values indicate highest proportions per metric (grey) and ties (light grey).
[b] Results for this metric only valid for the block subdivision portion of the algorithm. Enabling the road generation portion would yield much higher proportions of passed tests, but these would no longer be directly comparable with the other algorithms (Dahal and Chow 2014).

Table 6: Proportions of non-rejected null hypotheses for ks-Tests for Different Metrics by Algorithm and Site Type[a]

Site type	SS			GPD1			OBB		
	area	width	Si	area	width	Si	area	width	si
Gridiron - R1	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00
Gridiron - R2	0.00	0.07	0.10	0.00	0.00	0.03	0.03	0.07	0.07
Fragmented - R1	0.00	0.00	0.03	0.00	0.00	0.09	0.00	0.00	0.00
Fragmented - R2	0.03	0.03	0.09	0.00	0.00	0.03	0.03	0.03	0.06
Warped - R1	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00
Warped - R2	0.00	0.00	0.13	0.00	0.00	0.03	0.03	0.03	0.00
Loops, lollipops-R1	0.16	0.13	0.03	0.09	0.00	0.00	0.00	0.00	0.00

[a] Shaded values indicate highest proportions per metric (grey) and ties (light grey).

Discussion

One goal of this study is to help users of block subdivision algorithms come to a decision about which algorithm to use for which site type, in order to create the most accurate parcel data possible within integrated models. With that in mind, we can define a "better" algorithm as follows: A better algorithm for a particular site type is one that performed better than the other two. Better performance being measured by the proportion of non-rejected t-test and Fisher's exact test null hypotheses of total number of sites, on average for all metrics for that site type. Because the null hypotheses of the ks-tests were mostly rejected, these results weren't incorporated into the conclusions. It's important to note that this doesn't necessarily mean the better algorithm produces non-statistically different average parcel metrics the majority of the time for a given site type; just that it does so more often than the others.

The average proportion of non-rejected null hypotheses over all metrics for a given site type was calculated and included in Table 7. The largest of these (shaded in dark grey) is used to derive a rule that states the better algorithm to use for that site type.

For example, the Straight Skeleton algorithm produced average parcel metrics that can't be shown to be different from their observed counterparts in 33% of gridiron, uniform sites. For gridiron, non-uniform sites, the same algorithm produced average parcel metrics that couldn't be shown to be statistically different from their observed counterparts 57% of the time. The average proportion of the GPD1 algorithm for this site type was highlighted in dark grey, because it has the highest average proportion of non-failed statistical tests (ie: 61%).

Table 7: Average proportions of non-rejected null hypotheses for t-Tests and Fisher's exact tests for all Metrics by Algorithm and Site Type and Rules Defining a Better Algorithm for each[a]

Site type	Average of metrics			Derived rule
	SS	GPD1[b]	OBB	
Gridiron - R1	0.33	0.33	0.45	OBB is the better algorithm for Gridiron uniform types
Gridiron - R2	0.57	0.61	0.58	GPD1 is the better algorithm for Gridiron non-uniform types
Fragmented - R1	0.35	0.17	0.36	OBB is the better algorithm for Fragmented uniform types
Fragmented - R2	0.52	0.43	0.39	SS is the better algorithm for Fragmented non-uniform types
Warped - R1	0.35	0.11	0.33	SS is the better algorithm for Warped uniform types
Warped - R2	0.41	0.19	0.43	OBB is the better algorithm for Warped non-uniform types
Loops, lollipops-R1	0.49	0.09	0.42	SS is the better algorithm for Loop & lollipop types

[a] Shaded values indicate highest proportions per metric (dark grey) that determine the rule.
[b] Results apply only to the block subdivision portion of this algorithm.

While this study has focused on easily measurable parcel characteristics, it is also important to note that the algorithms vary in other important respects such as their ability to generate road networks and to adjust roads or parcels to fit the underlying topography of an area. The GPD1 algorithm can generate roads around blocks created from the recursive binary subdivision process. These roads are mostly straight and at right or slightly smaller angles. Newly generated roads ensure parcels have egress but since the road generation function was disabled in this study, a number of the synthetic parcels generated by this algorithm have none (Figure 20b). A road can also be generated to connect the new area to the existing road network. However this is a straight-line road, that doesn't break for terrain features or adhere to any network structure. This algorithm is also not sensitive to topography. While the GPD1 algorithm is part of a self-contained parcel generation toolset that generates roads and parcels, the OBB and SS rely on linked tools to generate roads within a larger simulation. Several linked road generation algorithms can create either rectilinear or organic roads with sinuosity as well as local roads or access structures. These algorithms have an input setting that controls how parcels adjust to fit a Digital Elevation Model (DEM). The roads can also adjust to fit a DEM resulting in three-dimensional blocks with different sizes and shapes than their two-dimensional counterparts. Such changes in input blocks are likely to have an effect on parcel characteristics, particularly for the SS algorithm which is sensitive to input block shape. While it would have been ideal to incorporate information from DEMs in the block subdivision of the SS and OBB algorithms, the available DEM was not accurate enough to be used (Smith & Sandwell, 2003). At the same time, the topography of the block type examples used in this study is not generally dramatic, so it is expected that the results would not change very much with the incorporation of a DEM. It is worth mentioning that a LIDAR-based DEM is expected in the next few years for the region, and as a result, it may be possible to revisit this issue in future research.

The SS algorithm always produces parcels with egress, while the OBB algorithm doesn't (Vanegas et al 2012). The OBB algorithm can be forced to produce parcels with egress, however at times it doesn't succeed and at others results in parcels that are unrealistically narrow (Figure 17c). Occasionally, however, residential parcels may be interior parcels connected to the road network through alleys or footpaths. The OBB algorithm is also appropriate for industrial and commercial parcels that sometimes have no egress. While the latter two algorithms provide more realism within their simulation tools, the former algorithm is modular, generates more features and can be integrated into a wider variety of simulation systems. A more quantitative comparison of these features, and how they affect parcel characteristics, was beyond the scope of this study.

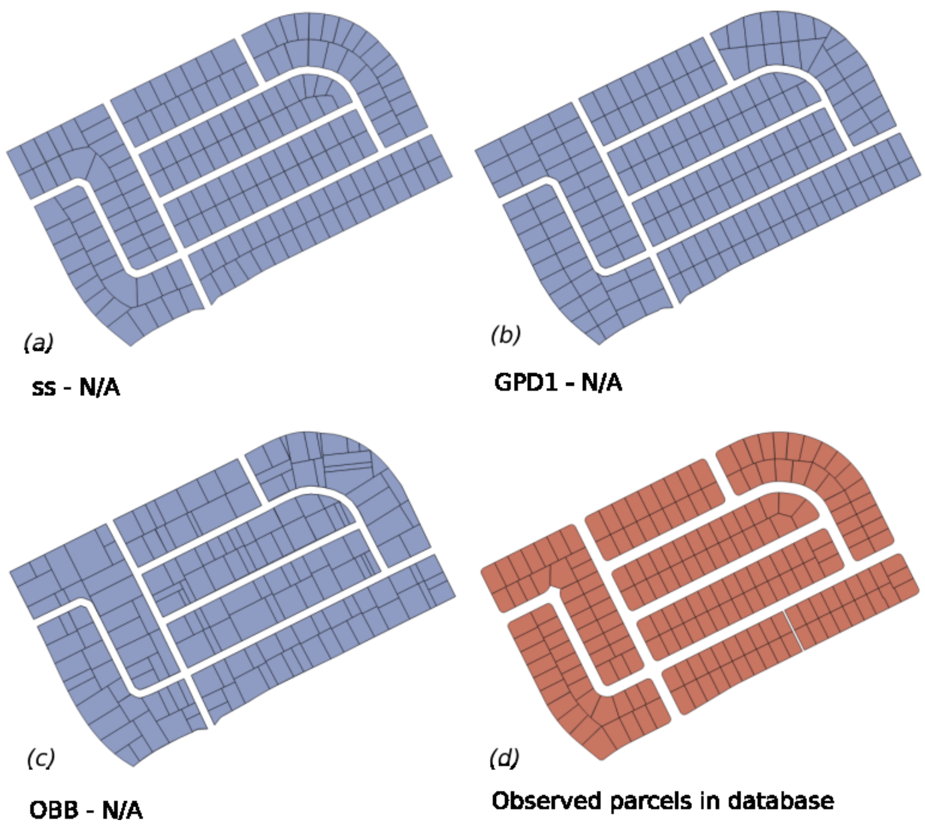
The results of the study suggest that in all cases, there is an algorithm that performs better for a site type relative to the others (Table 7). For loop and lollipop sites, the Straight Skeleton algorithm produced, on average, a non-statistically different average parcel metric in 49% of sites. This is a higher percentage than the GPD1 algorithm for the same site type (9% of sites) as well as the OBB algorithm (42% of sites). The SS algorithm also performed better than the other two algorithms for the fragmented non-uniform and warped uniform site types (52% and 35% probability of performing well respectively). Similarly, the OBB algorithm performed better than the other two algorithms for the warped non-uniform site type (43% probability of performing well) and the gridiron and fragmented uniform site types (45% and 36% respectively). The

GPD1 algorithm performed better than the other two for the gridiron non-uniform site types (61% probability of performing well). All of these "better" algorithms produced non-statistically different parcel metrics between ~20 and 60% of the time. This suggests the algorithms can be further developed to reproduce parcels with similar characteristics to ones that are observed, assuming this is a desirable goal.

The ks-tests indicate that in general, the distributions of area, width and SI of synthetic parcels aren't statistically similar to those of the observed parcels (Table 6). With few exceptions, the Straight Skeleton algorithm produced the highest proportion of sites with non-statistically different distributions for all site types. This indicates that the SS algorithm more accurately reproduces distributions of parcel metrics than the other two, though it still does so infrequently.

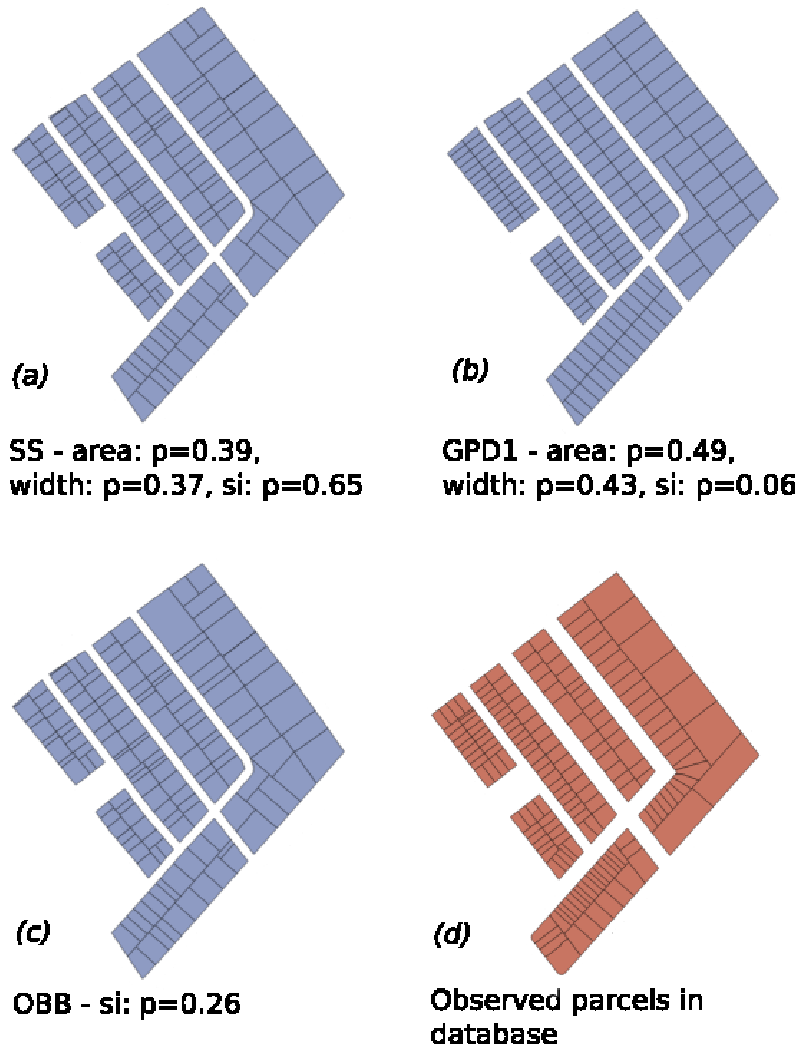
Differences between synthetic and observed parcels may also be caused by inaccuracies in block shapes derived from the road network when compared with the observed ones from the parcel database. Input blocks with the spatial accuracy of those in the parcel database are not generally available when synthetic parcels are needed. However, as the input data approaches these ideal conditions, the algorithms should perform better and better. Such conditions can be brought about by higher accuracy in road network data. They can also potentially be brought about through the study of municipal guidelines, which determine the range of possible lane widths and sidewalk widths for a given area such as those by the Pioneer Institute for Public Policy Research (2005). Knowing these values can improve the representation of road widths, resulting in more accurate block shapes and sizes used in subdivision.

It seems that all the algorithms performed better for site types with non-uniform parcels than for ones with uniform parcels, often by orders of magnitude of 2 or higher. The difference in results between non-uniform and uniform site types was least extreme for the OBB algorithm. An example of a fragmented grid, uniform site and a fragmented grid, non-uniform site subdivided by each algorithm are shown in figures 15 and 16. The test results demonstrate that in the uniform site, no non-statistically different parcel metrics are produced other than egress, while in the non-uniform site each algorithm produced between 2 and 4 non-statistically different average parcel metrics. These sites are representative of the general trend in the results. The primary reason for this trend is surely that the variance in the indicators for the regular sites is lower than in the irregular sites, and as a result, it is easier to have a significant t-test than in the case of irregular sites. In this sense we can say the t-test is more forgiving in the case of irregular than regular block types. Another possible explanation is that the algorithms produced less uniform parcels where the blocks bend. These parcel irregularities would be more comparable to the non-uniform site types than the uniform ones. Furthermore, since the OBB algorithm recursively subdivides parcels, the parcels in an area are a function of the shapes produced in the previous recursion. In this case, the overall shape of the block would have less of an impact on the parcels than the other algorithms. It seems that the OBB produced more uniform parcels at bends and angles than the other algorithms, which would explain why the differences between results of non-uniform and uniform site types are less extreme here.



Only p-values of significant metrics reported.

Figure 15: Fragmented grid - uniform parcels generated by algorithms, Wiseman & Patterson (2016)



Only p-values of significant metrics reported.

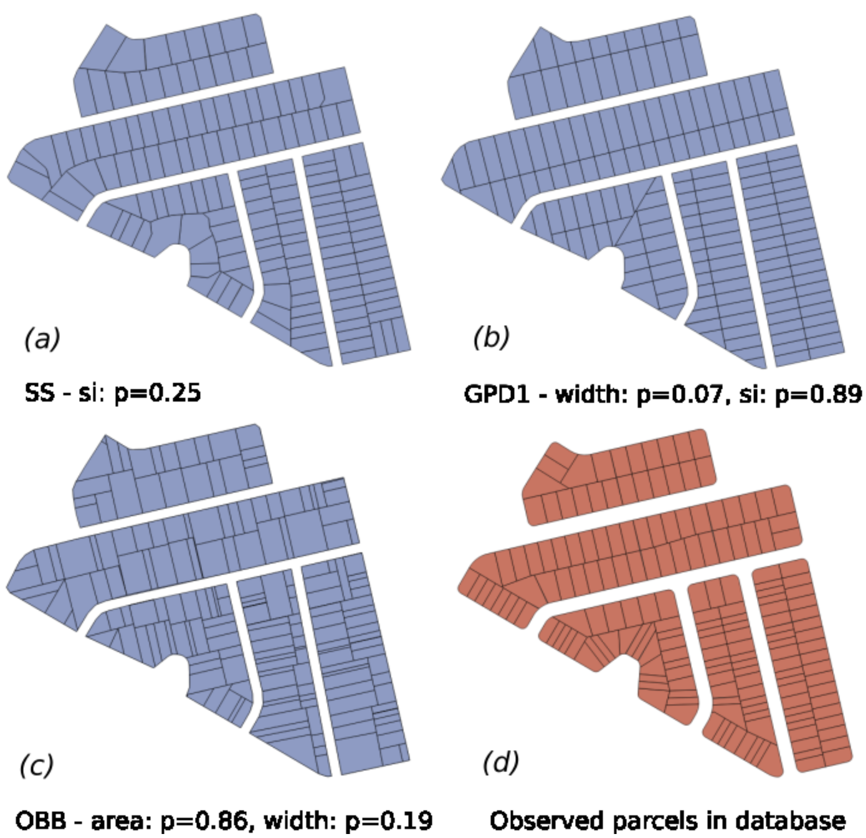
Figure 16: Fragmented grid - variable parcels generated by algorithm, Wiseman & Patterson (2016)

One consistent trend is that certain algorithms tended to perform better for certain metrics. This is most apparent from the t-test and Fisher's exact test results (Table 5). For example, in 4 out of 7 site types, the SS algorithm produced the highest proportion of non-rejected null hypotheses for the SI metric and in all site types for the egress metric. The OBB algorithm performed better for the area metric (7 out of 7 site types) and width metric (4 out of 7 site types). Moreover, for site types where it was the better algorithm for the SI metric, the SS algorithm usually reproduced non-statistically different SI metrics >50% of the time. These results make sense given the algorithm designs. For example, the SS algorithm was designed by looking at how planners approach the geometric problem of subdividing street blocks using hand drawn sketches. It's not surprising that the resulting parcel shapes, street accessibilities and patterns are most similar to those observed. Unlike the other two, the OBB algorithm has two area input parameters that are constraints as opposed to deterministic, as well as a constraint width parameter, possibly an indication of why it most often produced more accurate parcel areas and widths.

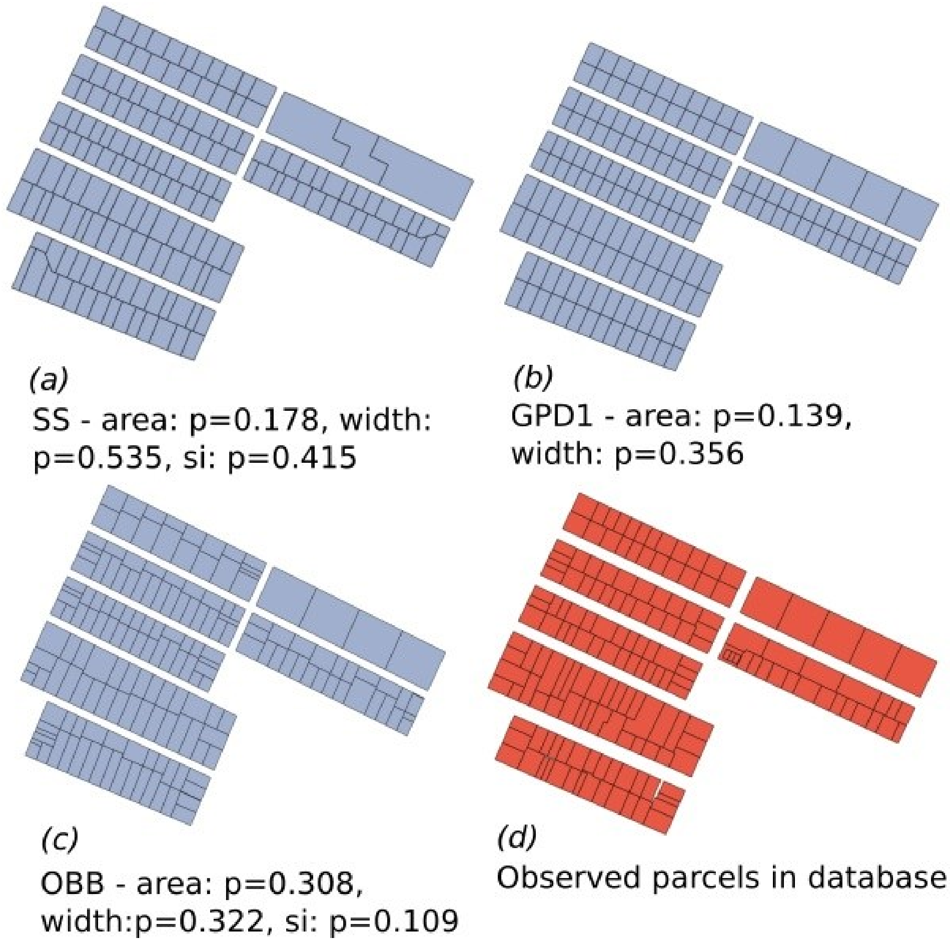
This finding may be especially useful in light of the fact that planning policy constraints often correspond directly to parcels. A maximum parcel area is often set to regulate development density and a street frontage (ie: parcel width) is used to regulate density and aesthetic characteristics of a neighbourhood. The results of this study can be used to select a better algorithm for simulating development under such policies. For a development scenario that specifies a certain minimum parcel width, the OBB algorithm would most closely reproduce those widths in the synthetic parcels. For a scenario that specifies a certain minimum or maximum parcel area, the OBB algorithm would also be a better choice. On the other hand, the SS algorithm most accurately reproduces parcel shapes, patterns and street accessibility, which is most helpful with visualizing aesthetic characteristics of a future city or for calculating walking scale accessibility metrics.

These results not only suggest a better algorithm to use in each of these scenarios, but they also allow us to predict the probabilities that these algorithms will perform well. Testing on 30 sites allows us to form predictions about how likely the algorithms will perform well on members of the population of sites of that type. This is particularly helpful if the geometries they produce contribute to things that are ultimately measured, for example in estimates of the spatial distribution of future populations, which are partly a function of parcel layout. These results could help make best use of the available algorithms within integrated models for most accurately predicting population, as well as contribute to estimating uncertainties of these predictions.

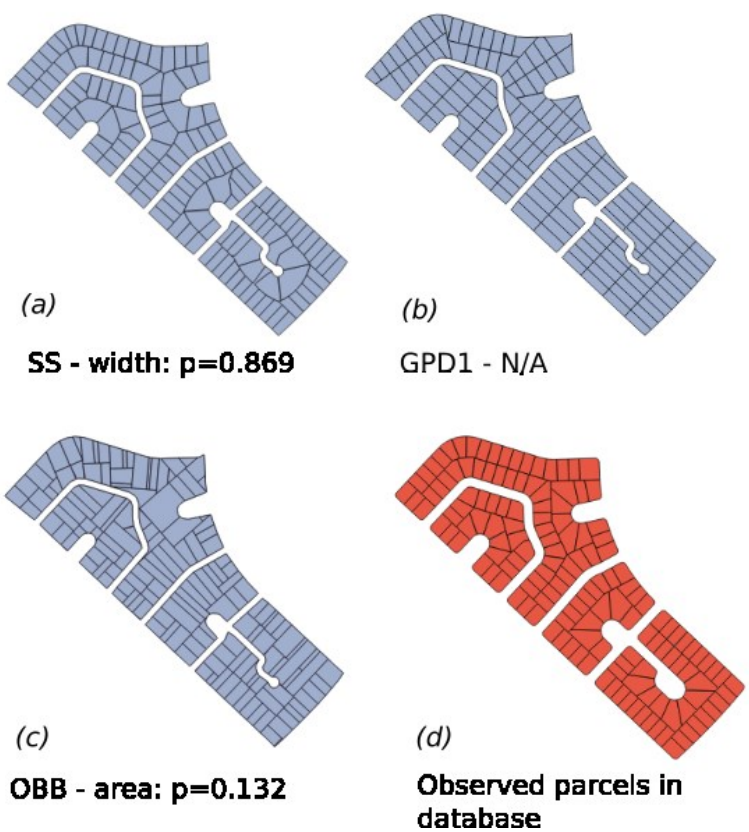
Furthermore, qualitative studies can suggest which algorithm produces parcels that look most similar to their observed counterparts. However, that algorithm doesn't necessarily produce parcels with most similar counts, areas or widths. In fact, visual observation of the synthetic parcels produced in this study suggests they often don't. In the example below (Figure 17), a visual inspection might lead one to conclude the SS algorithm reproduces parcels most similar to their observed counterparts. A statistical test would however lead us to the conclusion that the OBB algorithm reproduces parcels that are not statistically different from their observed counterparts with respect to average width (p=0.19), area (p=0.86) and egress (p=0.13) metrics, while the GPD1 reproduced non-statistically different width (p=0.07), SI (p=0.89) and egress (p=0.48) metrics. The SS reproduces non-statistically different average SI (p=0.25) and egress (p=1.00) metrics. Furthermore, a statistical comparison over multiple sites of the warped non-uniform site type would lead us to the conclusion that the OBB algorithm most often reproduces non-statistically different average parcel metrics for this site type (OBB – 43%, GPD1 - 19%, and SS - 41%). Several sites subdivided by the three algorithms are included for comparison in figures 17-20 below.



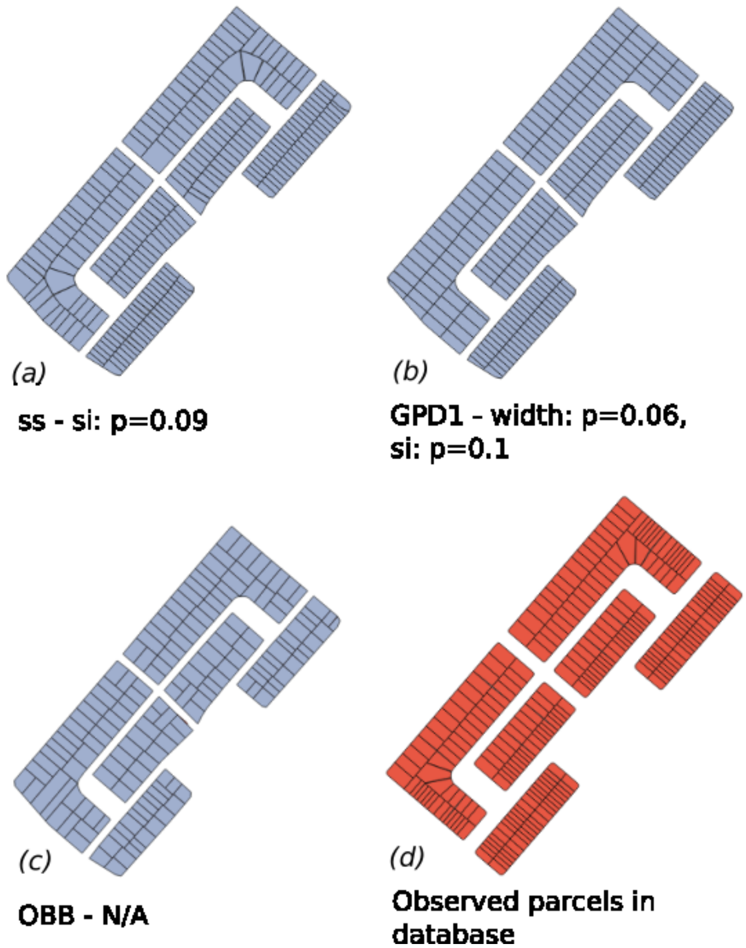
Only p-values of significant metrics are reported.
Figure 17: Warped grid - variable parcels generated by algorithms, Wiseman & Patterson (2016)



Only p-values of significant metrics are reported.
Figure 18: Gridiron - variable parcels generated by algorithms, Wiseman & Patterson (2016).



Only p-values of significant metrics are reported.
Figure 19: Loops & Lollipops - uniform parcels generated by algorithms, Wiseman & Patterson (2016)



Only p-values of significant metrics are reported.
Figure 20: Fragmented grid - variable parcels generated by algorithms, Wiseman & Patterson (2016)

Conclusions

This study tested how well parcels simulated with three block subdivision algorithms compared with their observed counterparts from a parcel database from Montreal, Canada. The authors of each of these algorithms (Vanegas et al. 2012, Dahal & Chow 2014) have already evaluated their own algorithms on a few sites from their own set of informal block type categories, some of which only encompassed a subset of all possible blocks. Furthermore, each developer based their evaluation on their own statistical or non-statistical comparison methods and on their own set of parcel indicators. This has made comparing the different algorithms problematic, presenting ambiguity as to which one to use on which block type. The current study, by comparing the different block subdivision algorithms according to the same set of indicators and across the same 30 sites from the same comprehensive block type classification, has helped to work toward defining a relationship between block type and an algorithm better suited to subdividing it. Finally, by testing each algorithm on 30 sites per type, it presented a method for finding the likelihood this algorithm will perform well either on the whole or for a given metric. This has the potential to increase the accuracy of synthetic parcel data produced using existing algorithms.

The results suggest that the OBB algorithm more often produces non-statistically different parcels on warped non-uniform sites as well as on gridiron and fragmented uniform sites. Meanwhile the SS algorithm more often produces non-statistically different parcels on loop and lollipop networks as well as fragmented non-uniform and warped uniform site types. The GPD1 algorithm more often produces non-statistically different parcels for gridiron non-uniform sites. It should be emphasized that the results for the GPD1 algorithm aren't a commentary on the performance of the algorithm itself, but rather on the part tested here, that is, the block subdivision feature. It was designed for a use case in which both new roads would be generated and the resulting blocks subdivided. However the road generation feature was disabled to create conditions under which the three algorithms could be more easily compared. These better algorithms performed well between ~20 and 60% of the time. In addition, the SS algorithm tends to produce the most non-statistically different average parcel shapes and street access and often does so the majority of the time, while the OBB algorithm does so for average parcel areas and widths. Assuming that it's a desirable goal to reproduce statistically similar parcel characteristics rather than parcels with realistic characteristics more generally, there is room for further development of these algorithms.

Currently within city generation engines, a user manually selects the subdivision algorithm to use on a set of blocks or else if the city is automatically generated, a single subdivision algorithm is used for all blocks. This study envisions a system where different subdivision algorithms are automatically executed in different locations, depending on the type of block there. This is to take advantage of the differing capabilities of each algorithm to subdivide different types of blocks. The results from this study map block type to a better algorithm suited to subdividing it. This mapping can be encoded into a larger function whose purpose is to take the block type as input and output a better algorithm suited to subdividing it. The pseudo-code of this function is included below:

```
CASE algorithm OF
  Gridiron AND Uniform : OBB
  Gridiron AND Non-uniform : GPD1
  Fragmented AND Uniform : OBB
  Fragmented AND Non-uniform : SS
  Warped AND Uniform : SS
  Warped AND Non-uniform : OBB
  Loops & Lollipops AND Uniform : SS
 ENDCASE
```

Figure 21: Pseudo-code of algorithm selection function

The different cases represent the set of rules for associating a block type with its better algorithm. Within a larger program that detected the block type in a given area, this function could automatically execute the better algorithm suited to subdividing it.

In addition to the rules outlined above, such a meta block subdivision program would require a linked model to either detect the block type that exists in a given area, for example, using a machine learning algorithm or multi category logit model with graph network metrics as predictor variables. Or else it would need a linked model to predict which type of block or road network is likely to exist and then generate it with the appropriate road generation algorithm. The variation in performance of the different algorithms suggests that such a meta program that incorporates them all and executes the best one for a given block type or metric could stand to improve the spatial accuracy of parcels in input data and in simulations of urban expansion, potentially improving the population forecasts on which they're based.

Manuscript 2 Modeling Road Network Type

Context

The following co-authored paper is designed to complement the above chapter (Testing Block Subdivision Algorithms on Block Designs) that recommends an optimal block subdivision algorithm to use for different block types. Since the block type of a future residential area is by definition unknown, this second manuscript attempts to estimate how likely each of these block types is to appear where.

The following manuscript was presented at the Transportation Research Board for its Annual Meeting in January 2016 (Wiseman & Patterson, 2016a). I had the role of lead author and Zachary Patterson, the second author, formulated the research question and advised the model estimation process. This study was designed to support the first paper by making its results more applicable in a practical context. It also represents a first attempt at estimating the spatial distribution of local road network types in future residential areas. Together with findings of which block subdivision algorithms works better for which road network types, it attempted to form a process for generating a plausible picture of the spatial distribution of urban form in future residential areas. Furthermore, the two studies are designed so that they could potentially be integrated into the simulation chain of integrated modelling tools and City Generation Engines.

The research uses the simple method of a logit model to estimate road network type. It takes into account a small subset of factors that might influence the local road network type in a future residential area, and doesn't explicitly take into account any decision-making processes of planning agents. As a result, the models are limited in their ability to predict road network type but since they are the first attempt at doing so the methodology employed represents progress. The research also contributes to an understanding of the factors associated with different local road network types in residential areas.

Abstract

Integrated land use-transportation modeling is used to predict future transportation demand given the way households and firms arrange themselves in urban areas as a function of among other things, the transportation system. More recent models operate at

the very fine spatial scale of the land parcel. At this scale, physical structures of the urban system change and these changes can take on different forms. Modeling these changes in general and their specific forms in particular can improve population forecasts and the travel demand forecasts on which they're based. Road geometry structures finer scale geometries like parcels and buildings, which have direct consequences on population predictions. This study tries to predict the type of local road network likely to exist in a future residential area. To this end, two multinomial logit models of road network type are estimated. Another two binary logit models are estimated based on more aggregate road network type categories; currently capable of being generated within GIS tools coupled to integrated models. Based on these models, road network type is found to be a function of a number of variables including: slope of terrain, period of development, proximity to a river and adjacency to a road network of the same type, among others. Anticipating the type of road network in a future urban area could lead to more accurate modeling of housing type, population density and demographic structure.

Introduction and Literature Review

The complex processes giving rise to changes in urban activities and landscapes lend themselves to being modeled through urban simulation. Urban simulations are being used more and more to assess alternative transportation investments, land use regulations and environmental policies (Vanegas et al., 2009a). Integrated land use-transportation modeling is a type of urban simulation that predicts travel demand in a way that is sensitive to how population and employment arrange themselves as a function of the transportation system.

There is a trend in integrated modeling toward representing phenomena at increasingly finer spatial scales, down to the scale at which they occur in reality - the parcel (Waddell, 2009). The advantages this has on realism of simulations are partly offset by increased requirements for highly detailed input data (Patterson & Bierlaire, 2010). In addition, geometric changes to the landscape can occur, the simulation of which requires extremely memory intensive computations (Waddell, 2009). Currently, such computations take place in external tools called City Generation Engines, for visualizing outputs of integrated models, among other things (Vanegas et al., 2009a; Weber et al., 2009).

Such fine-grained dynamic spatial representation presents the opportunity to represent urban form within simulations, such as different road network, parcel and building patterns. There are many automatic methods for generating road networks, including L-systems (Parish & Muller, 2001), Rapidly-exploring random trees (Lavalle, 1998) and agent-based approaches (Smelik et al., 2009). Many of these approaches are capable of generating local road networks of specific types (Chen et al., 2008; Glass et al., 2006; Kelly & McCabe, 2007; Sun et al., 2002), some of them integrated into City Generation Engines (Aliaga et al., 2008; Vanegas et al., 2009a; Weber et al., 2009). A number of block subdivision algorithms have also been developed, for subdividing the negative space of the road network into land parcels (Dahal & Chow, 2014; Parish & Muller, 2001; Vanegas et al., 2012; Weber et al., 2009; Wickramasuriya et al., 2011). Wiseman and Patterson (2016b) demonstrated that there is a different better algorithm for creating realistic parcels within each road network type. Buildings are generated on parcels using CGA (Computer Generated Architecture), which can include rules that define the types of buildings allowed on different parcel geometries. Schirmer (2010) proposed a logit model for predicting types of new buildings within integrated models based partly on a parcel's geometry. Clearly there is a chain of simulation events that link population forecasts back to road network geometry such that the latter is likely to affect the former.

While algorithms have been developed to create different kinds of road networks, little has been done to evaluate what types of road networks are likely to appear where. The aim of this research is to work towards models that will help in making this decision. In this paper, a series of logit models are estimated to meet different research goals. One goal is to understand, as much as possible, the factors influencing road network type and the other, to provide a tool for estimating road network type in future neighborhoods within urban simulations. Incorporating such a model into urban simulation tools could potentially result in more realistic parcel shapes and patterns, which could lead to more accurate overall population forecasts and demographic structures.

The rest of this paper is composed of the following topics: a description of the road network model, a description of the study area, the data and methods used to prepare the sample units and variables, the model results, and finally, a discussion and conclusion and description of potential future work.

Local Road Type Model

This study develops local road network models, that is, models to predict the type of road network in a residential area as a function of attributes of the area itself. In this case, the dependent variable has *J* levels, each level representing one of the local road network types. The Binary logit model is used for statistical classification problems where the variable of interest has two levels. The Multinomial logit model is an extension of this and is used where the variable of interest is categorical with greater than two levels. Both model types are used here and have essentially the same structure as described in Long (1997) and summarized briefly below.

A residential area i is characterized by road network type $j=1$ to J according to a linear function of a vector of predictor variables X. The predictor variables have an associated vector β_m, including an intercept β_{0m} and coefficients β_{km} for the effect of X_k on outcome m.

To ensure the probabilities are non-negative, the exponential of $x\beta_m$ is taken $\exp(x\beta_m)$. This value can be converted into a probability of observation i belonging to category m by ensuring the outcomes sum to 1. This is done in the equation below:

Equation 9: Probability function for logit models

$$\Pr(y = m|x_i) = \exp(x_i\beta_m) \,/\, \Sigma_{j=1,J} \exp(x_i\beta_j)$$

The probability is given by the base e to the power of $x\beta_m$ divided by the sums of the base e to the power of $x\beta_j$ for all the categories $j=1$ to J.

In this study, four models are estimated, two are complete models and the other two are reduced models. Of these, two are multinomial with four levels and the other two are binary with only two levels. The four level models estimate road network types based on the most common local road classification system. Since City Generation Engines don't simulate each of these types in all their detail, the other two models combine them into two aggregate types with associated automatic generation methods. For the complete models, as many variables as were available and that were believed to potentially influence road network type were tested. The aim of these models was to understand as generally as possible the factors affecting road network type. The reduced models were developed so that they could be used in a more realistic predictive environment; in the context of integrated land-use transportation simulation where it would be useful to predict the nature of road networks in expanding urban areas. As such, they use only variables that would be known before the neighborhoods were developed and that could be easily incorporated within the base year database of an integrated model.

Study Area

The study spans the Communauté métropolitaine de Montréal (CMM), composed of a central island and surrounding regions. The island is linked to the outer communities, many of them suburban, by bridges (Tomalty, 1997). As the main transportation corridor for colonists, the St-Lawrence River heavily influenced settlement patterns in Montreal. Its banks were colonized first and eventually urbanization followed in a monocentric growth pattern (Dufresne et al., 2003). Old Montreal, located on the island's southeast coast, is the region's oldest urban center and is the origin of this growth pattern.

Recently population growth in the outskirts has been three times higher than in the central city (CMM, 2010). Furthermore, exurban areas grew by 18,500 people from 2006-2011, the largest increase in any major Canadian city (Gordon & Shirokoff, 2014). Montreal has a decentralized planning structure organized around the MRC (Municipalité régionale de comté) with 14 MRCs spanning 82 municipalities. The MRCs are responsible for creating regional development plans that are flexible enough to allow municipalities to identify how and where to develop as well as the location and type of major roads (Tomalty, 1997). This decentralized planning structure is likely to result in a large spatial variation in local road network types.

Data and Methods

The study concentrated on residential road network types. To estimate the different models, it was necessary to split the study area into sample areas, and classify them according to road network type. To be useful in constructing the models, the sample areas had to be residential and small enough to be mostly of a single type. To do so, residential building points from Ville de Montreal (2011) were clustered at a distance of 120 m using the ArcGIS “Aggregate Points” tool (Esri, Redlands CA). This distance was chosen from among several tested, and seemed to produce the most coherent areas most representative of neighborhoods. The clusters sometimes contained holes, which were identified and filled in using an ArcGIS geodatabase topology. To smooth jagged edges, a 120 m buffer was applied to the clusters followed by a 120 m inverse buffer. To separate clusters that spanned several neighborhoods they were clipped by a buffered layer of highways, arterials and collectors (25m, 15m and 15m respectively) from GeoBase (2010). The local road network from DMTI (2013) was then spatially joined to these polygons, to assign sets of links to sample units.

Once the sample units were created, they were classified into four of the road network types described in Southworth & Owens (Southworth & Owens, 1993) as well as a mixed category for all the hybrid types, which were excluded from the sample. The classification system of Southworth & Owens was selected from among several (Cherry & Nagle, 2009; Sandalack et al., 2013), as it is most often cited in the literature and composed of the most basic types. The four types used represent categories along a spectrum of more grid-like to more tree-like and are as follows: 1) gridiron, 2) fragmented parallels, 3) warped parallels, and 4) loops and lollipops. The lollipops-on-a-stick type was excluded since its branching structure doesn’t form enclosed blocks that urban simulations require in creating subsequent geometries. In total, there were 869 sample areas in the reduced model (gridiron: 146, fragmented: 329, warped: 236, loops and lollipops: 158) and 842 sample areas in the complete model (Figure 22). Some of the sample areas needed to be dropped from the complete model due missing data in calculating the extra predictor variables.

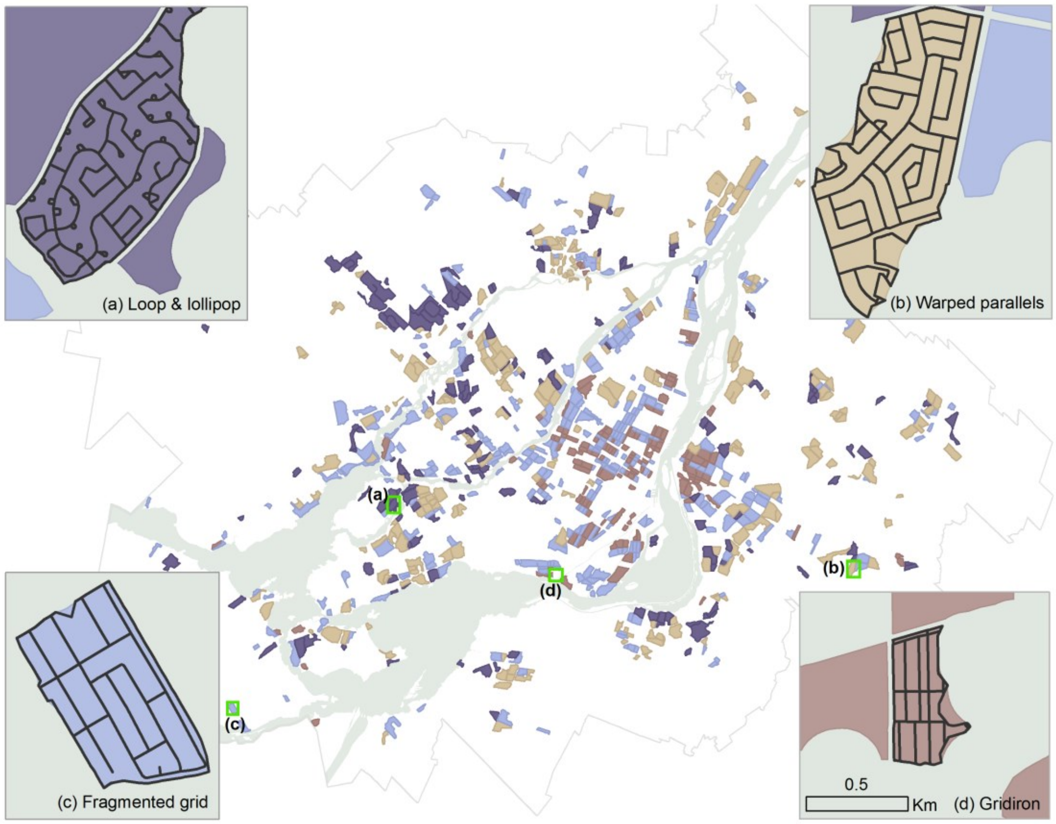


Figure 22: Sample areas color coded by road network type, (Wiseman & Patterson, 2016b)

Description of Explanatory Variables

Planners and architects have a palate of road network types to draw from when designing a new residential area. There is little information in the academic literature that describes how planners come to the decision of what type of local road network to use where. Many studies have sought to characterize neighborhoods of different local road network types with respect to traffic safety (Rifaat & Tay, 2008; Sun & Lovegrove, 2013), mode choice (Rodriguez & Joo, 2004; Snellen et al., 2002; Winters et al., 2010), walkability (Sandalack et al., 2013), vehicle miles travelled (Cervero & Murakami, 2010), population density, housing type (Peponis et al., 2007; Wang, 1998), and housing value (Song & Quercia, 2008). A historical survey of residential development can give insight into the functions of different road network designs and how these emerged from different planning paradigms, emphasizing practicality or aesthetics at the extremes (Southworth

& Ben-Joseph, 2003; Southworth & Owens, 1993). For example, the gridiron emerged during the industrial revolution as an efficient way to quickly expand the inhabitable area while allowing for speculation. Fragmented parallel patterns were developed to create more privacy by limiting through traffic. The long curved blocks of warped parallels aimed to provide residents with interesting sight lines. Loops and lollipops were designed to create enclaves where houses could have privacy from one another and to limit through traffic. While these features indicate why planners might have used a given road network design, they provide little insight into why one might be used over another based on pre-existing characteristics of the sites themselves. As a result, the variables tested here were informed by the limited academic literature available on the subject, other loosely related models, personal communication with planners and common sense.

Independent variables for the road network model were calculated in ArcGIS and QGIS. Their data sources include: heavy rail stops provided by the Agence métropolitaine de transport (2008), a digital elevation model at 20m resolution from GeoBase (2015), a road network from GeoBase (2010), housing type and greenspace data from Ville de Montréal (2012), and predominant age of housing construction for the latest available data at time of writing from Statistics Canada (2006).

For the reduced model aimed at predicting road network type, a number of spatially explicit explanatory variables were calculated, all of which would be known for a potential residential area before any local road network is constructed. These fit into 3 categories: spatial features of the terrain, transit variables and time period of construction. The first of the spatial features was road distance in meters from the central business district (Newburn & Berck, 2006; Pocewicz et al., 2007). Leveraging Montreal's monocentric growth pattern (Dufresne et al., 2003), this variable was meant to represent the relative time of construction of the different sample areas. The coefficient was expected to be negative for road network patterns popular in the distant past when the city center was developed, namely gridiron and fragmented parallel patterns. It was expected to be positive for road network patterns popular in the more recent past, namely, warped parallels and loops and lollipops. A binary variable representing whether or not the average slope of the terrain is greater than 2% was calculated. Design codes specify maximum slope tolerances allowed for building roads and slope of terrain is a major factor in decisions of which new areas to develop. Constructing grids in general, and on sloped terrain in particular, requires longer road lengths and is therefore more costly than constructing loop and lollipop networks, which easily follow contour lines (Southworth & Ben-Joseph, 2003; Southworth & Owens, 1993). It was expected that a >2% slope would increase the odds of constructing a loop and lollipop network and decrease the odds of constructing a grid network. However, since the urban core, as defined in Patterson et al. (2014), was developed when gridirons were in fashion, and a large portion of the core is occupied by a mountain (Mount-Royal), sloped terrains are more likely to be gridirons within the core. To capture this effect, the slope variable was interacted with a binary variable indicating whether or not the sample was inside or outside this core. A variable representing straight-line distance from the sample's centroid to nearest highway and another to the nearest major road were also calculated (Newburn & Berck, 2006; Pocewicz et al., 2007). This was meant to capture the scale differences of neighborhoods of each road network type, with loops and lollipops being automobile scale and grid types walking scale. Sample area was also included as a variable for this reason. Whether or not the sample is within a 100m buffer of a major river (or any river) was calculated. Being near a river was expected to increase the likelihood of building a warped parallels or loop and lollipop road network since these were seen as compatible with the geometric and aesthetic qualities of rivers. Two transit variables were calculated, namely, whether or not the area is within an 800m buffer of a heavy rail stop (metro and commuter train) and whether or not the area is within an 800m buffer of a metro only (El-Geneidy et al., 2014). It was expected that for the two grid types (gridiron and fragmented parallels) this coefficient would have a positive sign since they would provide a walkable network to heavy rail. While for the warped parallels and loop and lollipop types, the coefficient was expected to be negative since walkability to heavy rail is a desirable neighborhood characteristic and these would decrease it. A number of variables were calculated to capture temporal trends in road network design. The number of years since 1951 when most housing was constructed was used as a continuous variable. A series of binary variables were also calculated indicating whether or not the most housing was constructed within a given time interval. The continuous time variable was expected to have a positive sign for warped parallels and loop and lollipop road network types and a negative sign for gridiron and fragmented parallels types. Binary variables were expected to have positive signs consistent with historical knowledge about when each road network type was most popular (Southworth & Ben-Joseph, 2003; Southworth & Owens, 1993). For example, the period after 1970 was expected to be positive for loop and lollipop road networks, in keeping with its popularity beginning in this era.

For the complete models, aimed at a quantitative understanding of relationships between road network type and neighborhood characteristics, a fourth variable category was included, namely, current land use characteristics of the area. Percent residential area occupied by multifamily housing was expected to have a positive coefficient for gridiron and fragmented parallel patterns, while percent residential area occupied by single family housing was expected to have a positive coefficient for warped parallels and loop and lollipop networks (Jochen Eckart, personal communication). Multifamily housing is typically more compatible with the practicality and efficiency of grid-type networks and single family housing with the privacy and aesthetics of loop and lollipop and warped type networks. A binary variable representing whether or not the area contains greenspace was expected to have a positive coefficient for warped parallel patterns and loops and lollipops, since these favor aesthetic features.

Spatial Dependence

Designing a neighborhood that fits into the surrounding aesthetic is often desirable and so the road network types of surrounding neighborhoods are also important factors (Jochen Eckart, personal communication). Failing to account for this type of spatial autocorrelation in multinomial logit models violates the assumption that the error components of the utilities of the different alternatives are independent. Normally this is addressed by introducing a spatially lagged variable and coefficient into the observed components of the utility functions. However, due to constraints on time and model complexity within simulations, a different approach was taken here to try to account for spatial autocorrelation within the modelling process.

To incorporate this spatial dependence into the model, a set of binary variables were calculated, each one capturing whether or not a given sample unit had at least one road network of a given type adjacent to it. Since strictly speaking, the sample units weren't adjacent in a topological sense (they had buffered roads in between), rather than calculating these variables using an adjacency matrix, a distance-based matrix was used according to the method described in Ding (2001). Here, the author used a variable search distance based on the area of the sample unit in question to define its neighborhood. This distance is calculated by the following:

Equation 10: Spatial Weight Matrix calculation

$$\psi_{ij} = 1 \; if \; d_{ij} \leq b\sqrt{\frac{area_i}{\pi}}, and \; \psi_{ij} = 0 \; if \; d_{ij} > b\sqrt{\frac{area_i}{\pi}}$$

Where $\psi_{ij} = 1$ means sample j is the neighbhor of i, d_{ij} is the cutoff search radius, b is a constant set at either 2 (for the rook's case) or 3 (for the queen's case), and $area_i$ is the area of the sample unit in question. For this study b=2 was used, meaning that being within twice the radius of a circle of the same area as a given sample area constitutes a neighbor.

Model Results

Many models were tested in Biogeme (Bierlaire, 2003), beginning by including as many variables as described in the previous section and then paring them down to those with a statistically significant effect. The results of the models with four outcome variables, representing the most common road network type categories, are presented first. Next, the models with two outcome variables, representing aggregate categories of road network types from the first model, are shown. Both the four- and two- outcome model types each have a complete and a reduced version. The former containing the maximum number of predictive variables and the latter only with variables that would be known at simulation time.

Four-Outcome Models

The model's predictor functions are those for the four road network types. The ASC (Alternative Specific Constant) indicates the preexisting odds that a road network is of the type of the predictor function it is in. Gridiron and loop and lollipop types both had insignificant ASCs, fixed at zero for the model, and as a result were both used as omitted categories. The results of the complete and reduced models are found in Table 8 and

Table 9 In the four-outcome complete model (Table 8) there are nine distinct variables in the entire model and the model's predictor functions have between 3-5 variables each. One variable is continuous while the others are binary. The likelihood ratio test indicates that the probability of the model coefficients actually being 0 as opposed to those estimated is extremely unlikely with a p-value of 3.9×10^{-4}. The adjusted rho square is 23%.

Table 8: Variables Included in Four-Outcome Complete Model and What they Represent

Variable	Coefficient(β)	p-value	exp(β)
Gridiron			
Gridiron ASC	0.00 (FIXED)		
Gridiron adjacent	1.0	0.00	2.801
Percent residential area occupied by multifamily housing	0.05	0.00	1.051
Warped Parallels			
Warped ASC	2.34	0.00	10.381
Warped adjacent	1.13	0.00	3.096
Within 800 m of heavy rail	-0.461	0.02	0.631
Fragmented Parallels			
Fragmented ASC	1.89	0.00	6.619
Fragmented adjacent	0.505	0.00	1.657
Percent residential area occupied by multifamily housing	0.0304	0.00	1.031
Loops and lollipops			
Loops and lollipops ASC	0.00 (FIXED)		
Loops and Lollipops adjacent	1.18	0.00	3.254
Average slope > 2% x outside urban core	0.807	0.01	2.241
Majority of housing constructed after 1970 Dummy	1.46	0.00	4.306
Contains greenspace dummy	0.781	0.00	2.184
Init log-likelihood	-1129.830		
Final log-likelihood	-855.721		
Likelihood ratio test	548.218	3.9x10-4	
Adjusted rho-square	0.232		

In general, the variables in the model are consistent with *a priori* hypotheses about their effects on road network type. Consistent with what was learned from architects, having an adjacent road network of a given type increases the odds of being the same type compared to all others combined. Higher percentages of multifamily housing are associated with gridiron and fragmented parallels types. Being within 800m of a heavy rail stop decreases the odds of being warped parallels. Having a slope >2 % increases the odds of being loops and lollipops, if outside the urban core. Having housing built mostly after 1970 increases the odds of being loops and lollipops. Having greenspace also increases the odds of being loops and lollipops, in keeping with the aesthetic nature of this type.

The degree to which these variables influence the road network type of a given area can be interpreted through the values of their respective coefficients. The expected values of these coefficients, that is, their influence on the odds of being a given road network type in relation to all of the others without this variable, is given in the exp(β) column. The first step to interpreting the exp(β) values is to subtract 1 from the number. The resulting value represents how much an independent variable increases (if positive) or decreases (if negative) the odds of having the outcome of that predictor function. For binary variables, the resulting value can be thought of as an increase or decrease in the odds of a neighborhood being a given road network type, if the variable is true, relative to all the other types combined that don't have this variable in their functions. For continuous variables, the resulting value can be thought of as a relative increase or decrease in odds of being a given road network type, for every one unit increase in an independent variable, relative to all other road network types combined that don't have this variable in their functions.

For example, increasing multifamily housing by 1% increases the odds of being gridiron by 5% and fragmented grid by 3%. While this variable has a large effect on the model results, in general it wouldn't be known when trying to predict road network type, unless it was specified in a policy or plan. Being within 800m of a heavy rail stop decreases the odds of being warped parallels by 37%. Having a slope greater than 2% increases the odds of being loops and lollipops by 1.24 times if in an area outside the urban core. Both having housing mostly built after 1970 and greenspace increases the odds of being loops and lollipops by 3.3 times and 1.2 times respectively.

Table 9 shows the reduced, 4-outcome road network model. The likelihood ratio test of the four outcome reduced model indicates that the probability of the model coefficients actually being 0 as opposed to those estimated is extremely unlikely with a p-value of $4.1x10^{-3}$. The adjusted rho-square is lower than the previous model since it includes fewer variables.

Table 9: Variables Included in Four-Outcome Reduced Model and What they Represent

Variable	Coefficient(β)	p-value	exp(β)
Gridiron			
Gridiron ASC	0.00 (FIXED)		
Gridiron adjacent	2.03	0.00	7.614
Average slope > 2% x within urban core	1.46	0.00	4.306
Years since 1951 when majority of housing was constructed	-0.0531	0.00	0.948
Within 100 m of a water body	0.647	0.03	1.910
Warped Parallels			
Warped ASC	0.3	0.02	1.35
Warped adjacent	1.44	0.00	4.221
Within 800 m of heavy rail	-0.451	0.01	0.637
Fragmented Parallels			
Fragmented ASC	1.14	0.00	3.127
Fragmented adjacent	0.565	0.00	1.759
Years since 1951 when majority of housing was Constructed	-0.0193	0.00	0.981
Loops and lollipops			
Loops and lollipops ASC	0 (FIXED)		
Loops and Lollipops adjacent	1.37	0.00	3.935
Average slope > 2% x outside urban core	0.589	0.02	1.802
Init log-likelihood	-1203.304		
Final log-likelihood	-972.562		
Likelihood ratio test	461.484	4.1x10-3	
Adjusted rho-square	0.182		

For the variables common to both full and reduced models, namely, adjacent type, slope and proximity to heavy rail, the degrees to which they affect the odds of different road network types are comparable. In general, the other variables in the model also fit with *a priori* hypotheses about their effects on road network type. The odds of being gridiron or fragmented parallels decrease over time, consistent with the evolution of planning paradigms for residential areas. Also, this model confirms that having a slope >2% and being within the urban core increases the odds of being a gridiron, due to the presence of Mount-Royal. Another interesting result is that contrary to expectation being near a river increases the odds of being a gridiron. Perhaps areas near rivers are furnished with pedestrian networks to allow the larger public to access them. The fact that being near a *major river only*, one that drains directly into an ocean, was insignificant, suggests this relationship isn't due to the fact that the banks of the Saint-Lawrence River were settled when grids were prevalent.

Again, a more quantitative interpretation of the influence of each variable on the odds of being a given road network type, in relation to all other types without this variable, can be made using the expected values of their coefficients (exp(β)). For the adjacent type variables, having a neighbor of a given type increases the odds of being that type by 6.6 times if the neighbor is a gridiron, by 3.2 times if it's warped parallels, by 76% if its fragmented parallels and by 2.9 times if loops and lollipops relative to the other categories. For the slope variables, an average slope >2% increases the odds of being a grid by 3.3 times if inside the urban core and loops and lollipops by 2.9 times if outside the urban core. Similar to the complete model, being within 800m of a heavy rail stop decreases the odds of being warped parallels by 36%, suggesting that more walkable road networks tend to be used in areas with nearby heavy rail. Being within 100m of a river increases the odds of being a gridiron by 91%. The odds of being gridiron and fragmented parallels decrease by 5% and 2% respectively for every year past 1951 that most housing is constructed.

Two-Outcome Models

The two-outcome models used most of the same variables as the previous models. Here the outcome variable had 2 levels: 1) gridiron OR fragmented parallels, representing the grid types, and 2) warped parallels OR loops and lollipops, representing the free form, aesthetic types referred to as organic. The results of the complete and reduced model can be found in Table 10 and Table 11.

Table 10: Variables Included in Two-Outcome Complete Model and What they Represent

Variable	Coefficient(β)	p-value	exp(β)
Grid types*			
Grid type ASC	0.00 (FIXED)		
Grid type adjacent	0.770	0.00	2.16
Percentage of residential land area occupied by multifamily housing	0.0287	0.00	1.029
In urban core	2.39	0.02	10.913
Organic types**			
Organic type ASC	0.00 (FIXED)		
Organic type adjacent	0.950	0.00	2.586
Within 800 m of heavy rail	-0.579	0.02	0.560
Contains greenspace dummy	0.716	0.00	2.046
Majority of housing constructed after 1970 dummy	0.620	0.00	1.859
Init log-likelihood	-564.915		
Final log-likelihood	-362.498		
Likelihood ratio test	404.834		
Adjusted rho-square	0.346	7.3x10-4	

*Includes both Gridiron and Fragmented types.
**Includes both Warped and Loop and lollipop types.

The complete two-outcome model Table 10 has an adjusted rho-square of 0.346. The likelihood ratio test (p = 7.3x10^{-}4) indicates that it is unlikely that the model coefficients are actually 0 rather than those estimated. The greater quality of this model compared to the four outcome models suggests that these variables are more strongly associated with a particular planning paradigm (practicality vs. aesthetics) rather than a particular road network type.

The signs of the variable coefficients are consistent with hypotheses and with the other models. Being adjacent to a given type increases the odds that an area is of same type by 1.2 times for grid types and by 1.6 times for organic types. A 1% increase in multifamily housing increases the odds of being a grid type by 3%. Being within the urban core increases the odds of being a grid type by 10 times. Being within 800m of heavy rail decreases the odds of being an organic type by 46%. While having greenspace and being developed after 1970 increases the odds of being an organic type by a factor of 1 and 1.9 respectively. The two-outcome reduced model (Table 11) has a likelihood ratio p-value of $2.4x10^{-4}$ and an adjusted rho-square of 0.298. It shares four variables in common with the complete two-outcome model, the only other variable being surface area. The common coefficients have similar signs and values between models and so only notable differences are presented here. Being in the urban core increases the odds of being a grid type by 26 times. Having the majority of housing constructed after 1970 increases the odds of being an organic type by 2.3 times. Finally, a 1 hectare increase in surface area decreases the odds of being a grid type by 1% in keeping with the hypothesis that these are more walking scale neighborhoods.

Table 11: Variables Included in the Two-Outcome Reduced Model and What they Represent

Variable	Coefficient(β)	p-value	exp(β)
Grid types*			
Grid type ASC	0.00 (FIXED)		
Grid type adjacent	0.955	0.00	2.60
In urban core	3.31	0.00	27.385
Surface area (ha)	-0.00876	0.00	0.912
Organic types**			
Organic type ASC	-1.01	0.00	0.364
Organic type adjacent	1.05	0.00	2.858
Within 800 m of heavy rail	-0.463	0.02	0.629
Majority of housing constructed after 1970 dummy	1.19	0.00	3.287
Init log-likelihood	-601.652		
Final log-likelihood	-415.584		
Likelihood ratio test	372.136	2.4x10-4	
Adjusted rho-square	0.298		

*Includes both Gridiron and Fragmented types.
**Includes both Warped and Loop and lollipop types.

Model Validation

In response to a reviewer’s comments on this thesis, the road network type models estimated in this study were validated through a simulation. First, the models were used to simulate outcome road network types for neighborhoods in the sample. The frequencies of simulated outcomes were then compared to the frequencies of observed road network types. A strong model would accurately assign outcomes to a high proportion of the sample. Normally an out-of-sample validation of the models would be performed, wherein a portion of the data would be used to estimate the models and a smaller proportion used to validate them (Anguita et al., 2012). However due to insufficient sample size this was not possible. As a result, an in-sample validation was performed, wherein all of the data was used to both estimate and validate the models (Anguita et al., 2012). The results of this validation are included in tables 23-27, in the form of frequency histograms of simulated outcomes for each observed set of neighborhood types.

The tables show that the models didn't succeed in accurately simulating outcomes for a large proportion of the sample, and this is true for all observed road network types. This indicates the models have low predictive capabilities. The four-outcome reduced model simulated the correct outcomes most frequently, but not by a large amount. For example, 55 of 146 observed gridirons were simulated as being gridirons, but 49 were also simulated as being fragmented parallels (Figure 23). For the four-outcome complete model, however, observed gridirons are more frequently simulated as being fragmented parallels and observed loop & lollipops are more frequently simulated as being warped parallels (Figure 24). These inconsistencies aren't present in the validation of the two-outcome models, where the more similar road network types are grouped together into aggregate categories. Here, the correct road network type was simulated the majority of the time, but there was still a substantial proportion of the sample that was simulated as the other type. For example, for the two-outcome reduced model, 73% of the grid sample units were simulated correctly (346 simulated grids out of 474 observed grids) (Figure 25). Similarly 66% of observed organic types were simulated as being organic (261 out of 394). The validation results are similar for the two-outcome complete model. The two-outcome models are more reliable than the four-outcome models. However, considering the high degree of abstraction and aggregation in the data, one would expect them to simulate a larger frequency of observed road network types correctly.

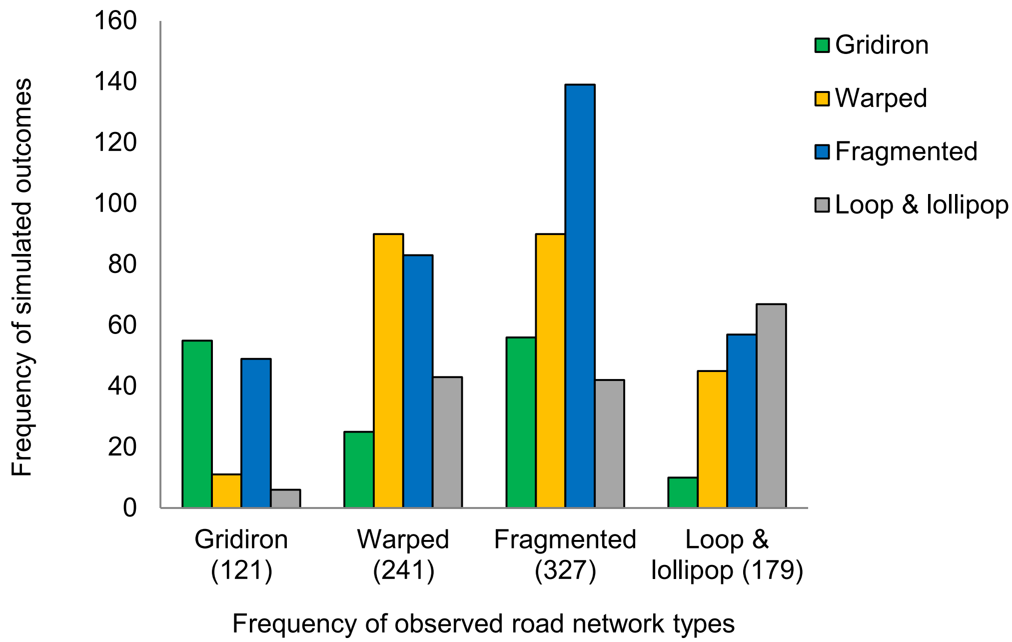


Figure 23: Four-outcome reduced model validation results

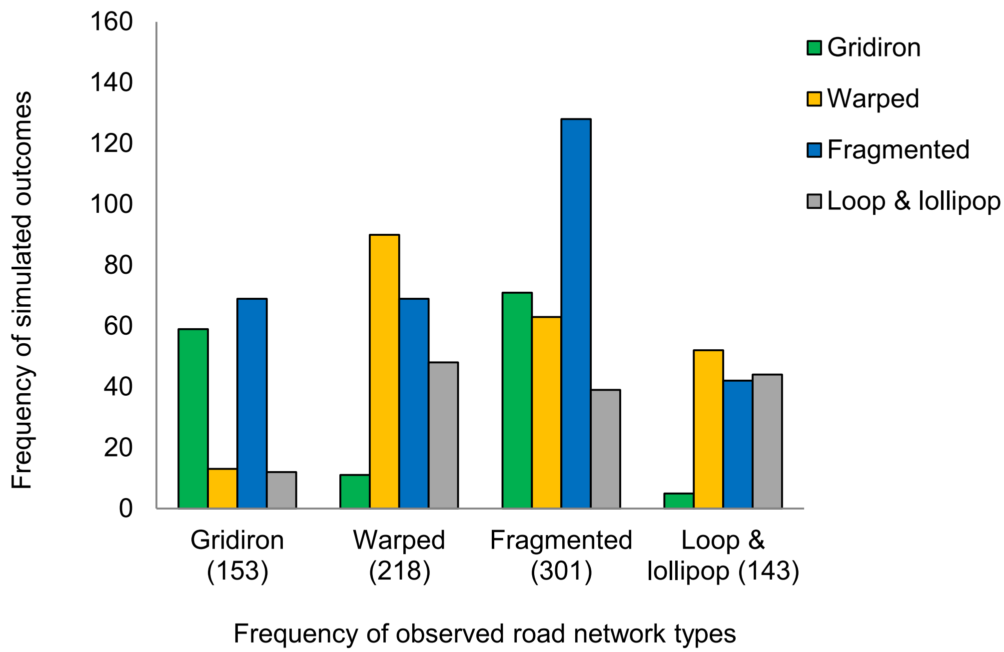


Figure 24: Four-outcome complete model validation results

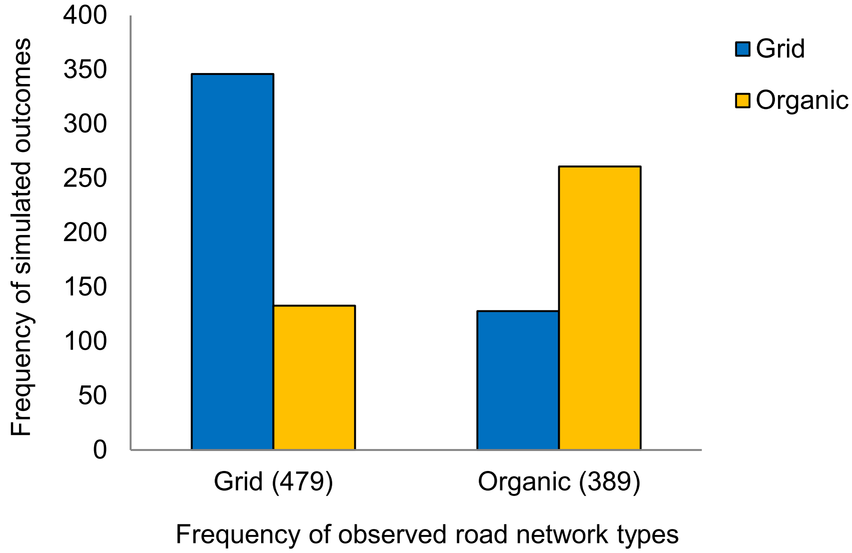


Figure 25: Two-outcome reduced model validation results

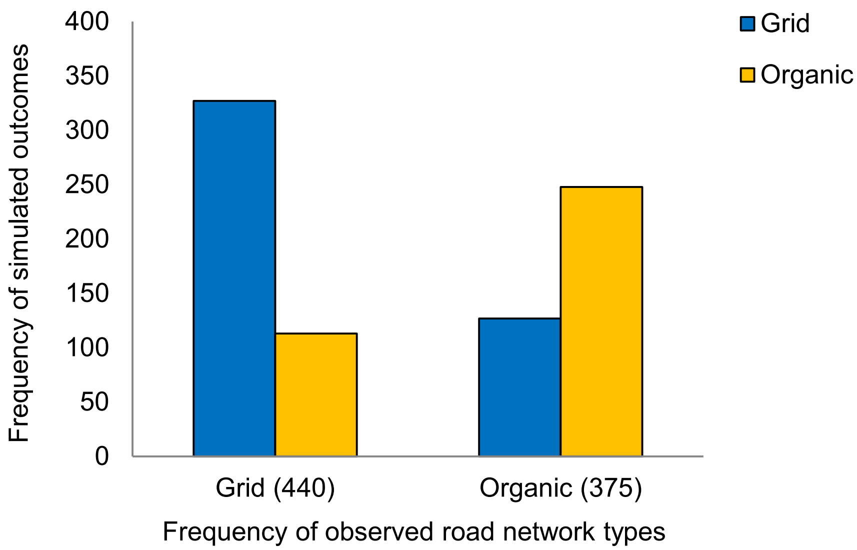


Figure 26: Two-outcome complete model validation results

Discussion & Conclusion

A number of automatic road generation algorithms have been developed, many of them aimed at recreating road networks of specific types. Within visualization and simulation tools, these form the basis of finer geometries, such as parcels and buildings both of which have a number of automatic methods for being generated as well. This research has attempted to create a model for automatically deciding which road generation algorithm to use in a future residential area, to best leverage the available algorithms in urban simulations.

More specifically, the purpose of this study was to develop statistical models to understand the factors influencing local road network type with the idea of being able to use such models to predict road network type in integrated modeling systems. The model results demonstrate that for the study area used, predicting road network type of future residential areas using site-specific factors that are known in advance is limited. However, these models do not include all the variables with potential to influence road network type. It is possible that applying a logit model structure to this problem could potentially yield a useful model if more factors were incorporated. These factors would likely be ones calculated endogenously to the integrated modeling system, such as market value of land and housing or population and employment forecasts in the region. Other factors could also include those specified in policy, such as population density, parcel size and housing type constraints at a local level or the extent of the urban growth boundary on a more regional level.

In that case, the models estimated here might provide a helpful framework for future modelling efforts. The four-outcome models both contained 9 distinct variables while the two-outcome models contained 7 or 6. Many of the variables were used in more than one model. The variable coefficients were usually similar between models but sometimes varied widely (ie: being within the core). In practice, the two-outcome reduced model has corresponding algorithms for being generated within urban simulations and so it could be directly applied in this context, although given the validation results, this isn't necessarily worthwhile. In all models, there is strong spatial dependence between an area's road network type and that of its neighbors. In at least some of the models, road network type was found to be influenced by the following factors: whether or not the terrain has slopes, the percentage of multifamily housing, proximity to a river or heavy rail stop, time of building construction and whether or not the area is within the urban core.

Furthermore, all the variables used in the reduced models are easy to calculate, and would be simple to incorporate into an integrated model's base year dataset. For the complete models, their variables may be specified in development plans or policies, in which case they could be used for simulation purposes.

Future Work

By incorporating more knowledge of road network design practices, the predictive capabilities of the models could be improved. Future work could involve testing these improved models within integrated land use-transportation models or City Generation Engines. In combination with rules relating better block subdivision algorithms to road network types, more reliable parcel geometry may result. Finally, an exploration of the data used to create these models and the model parameters could help define quantitative urban typologies that relate road network type to neighborhood features.

Conclusion

This thesis set out to build on the extensive set of algorithms for generating geometries within integrated modelling and city simulation tools. It did this not through creating additional algorithms, but rather by providing guidance on when to use which one so as to more realistically capture the spatial distribution of urban form in simulations.

The research succeeded as far as providing guidelines for users of block subdivision algorithms on which one to use for subdividing a specific block type. The block type classification incorporates dimensions of both road network and parcel pattern type, and comprises forms most commonly found in North America. It could also be used for testing new algorithms, thereby providing a basis for comparing them and informing future guidelines. These guidelines, by contributing to more realistic parcel attributes, may even help improve population and demographic forecasts of modelling tools. However, if nothing else they at least provide guidance to users on when to implement which algorithms.

The idea of testing an algorithm on 30 sites of a given type may prove to be a useful methodological contribution to testing algorithms more generally. Particularly where such algorithms aim to generate features that are ultimately counted or measured, or where such features depend on those being generated (and where a single better algorithm may not exist for all cases). The review article of procedural methods for terrain modeling by Smelik et al. (2009) illustrates the diversity of methods for generating trees, height-maps, and road networks among other objects. Such methods have focused on generating plausible rather than quantitatively accurate scenes, however as methods continue to improve, the latter may become desirable. In this case it might be appropriate to apply a more statistically reliable method for determining which algorithms perform better overall or in specific cases. For example, algorithms that generate different types of local road networks might be evaluated based on how similar their total road length or intersection counts are compared to observed road networks (Wickramasuriya et al., 2011). This may lead to more accurate integrated modeling simulations of walking scale metrics.

This thesis also created a methodology for modelling road network type, an as-yet unexplored research area. Though the quality of the modelling results could be improved, doing so will become increasingly relevant as walking scale indicators of travel behaviour are incorporated into integrated modelling tools (Waddell, 2009). That way the effects of policy on walkability and dispersion of future neighorhoods might be assessed. Furthermore, having identified factors associated with different local road network types, and the strength of their associations, contributes to a quantitative understanding of urban form. Specifically, several factors were identified, from a large set of hypothesized ones, which were actually found to have an influence on local road network type within a logit modeling framework. While the variable coefficient estimates are unreliable, it is possible to use the number of models a given variable appears in as a rough indicator of how strongly it influences road network type compared to another variable.

Limitations

The studies conducted here have several weaknesses. The testing of block subdivision algorithms was limited to residential blocks and didn't make full use of the algorithms' capabilities. For example, functions designed to take topography into account or generate new roads weren't used due to data constraints or else to facilitate comparison between the algorithms.

The block subdivision rules are based on hypothesis tests that give a simple binary indicator of whether or not an algorithm performed well, or rather didn't perform poorly, for a given site. In aggregate, these test results give an indication of how well the algorithm performed overall. The way the results were reported, there is little indication of the degree of similarity between generated and observed parcel sets for a given site and metric.

When considering which areas the block subdivision algorithm rules can be applied to, it is important to note that the block types were derived from a study of urban form in North America (Southworth & Owens, 1993). As a result, the rules would be most safely applied to North American regions. While planning trends in Europe influenced the development of North American cities, this influence is only as old as these cities themselves. Therefore, the study results wouldn't be applicable to European areas constructed before the 19th century or to other areas with entirely different planning influences (Southworth & Ben-Joseph, 2003).

The study area also represents a local, somewhat atypical North American context, which may further limit how widely the results can be applied. Unlike the majority of North America, the territory of Montreal was subject to the seigneurial system, an institutional form of land distribution established in New France in 1627 (Harris, 1984). Unlike in New England, where farm plots were square, the plots in New France were thin and rectangular. The legacy of this system remains in the access roads that shaped these rectangular plots. A given local road network type could have taken slightly different forms within each of these access structures. In that case, applying the better block subdivision algorithm to its corresponding block type within North America may not yield the same results as the study.

Furthermore, the road network models estimated here are limited in their capability to predict actual road network types. As the validation results suggest, the models are mostly capable of distinguishing between the more aggregate grid-type and organic road network types. However, they can't distinguish between finer road network categories than these. The modeling results could at best be said to estimate the type of road network in future areas, but this falls short of actual prediction. For this reason, they aren't ready to be implemented in simulations.

To explain this, it is possible to point to some large simplifying assumptions relating to spatial autocorrelation and to factors influencing road network type. With respect to spatial autocorrelation, here, the potential interaction between sample areas based on spatial location was reduced to a simple binary variable indicating whether or not the sample area has at least one neighbour of the same road network type. In reality, the odds of a road network being a given type should increase with the number or area of neighbours of that type.

With respect to factors influencing road network type, these models only include preexisting characteristics of the areas themselves, as well as one housing policy constraint. They mostly fail to incorporate an understanding of how and the extent to which policy and land use regulations influence local road network type. On the whole, these latter characteristics might play a larger role in influencing road network type, in which case efforts to model road network type based on preexisting characteristics of the areas themselves are inherently limited.

Another weakness of this study is perhaps that its value is based on the premise that modeling the spatial distribution of block types within integrated modelling frameworks would increase accuracy of population and employment forecasts enough to influence travel demand. This premise is just a hypothesis. While it is reasonable to assume that generating more accurate parcels would improve accuracy of forecasts at a local level, it is unclear whether this improvement would be large enough to measurably affect travel demand at the more aggregate scale at which it is usually assessed. It might be helpful to verify this hypothesis before continuing any future research in this area.

Future Work

The guidelines defined in Testing Block Subdivision Algorithms on Block Designs can be implemented immediately to generate missing parcel data or to manually edit urban simulations. For them to be useful in automating the block subdivision process, however, there would need to be some way to automatically capture the block type or local road network type of residential areas. If the road network were generated within the tool, as in city simulations, there should be a way of documenting the algorithm that was used and hence which block type was created. Otherwise, for generating missing parcel data, some way of classifying road network patches by type could be elaborated if a fully automated approach is desired. This could be through a machine learning approach or using a logit model approach similar to ones estimated here, but instead for classifying road networks that already exist. Such a model could be estimated using quantitative variables calculated from graph metrics.

Ideally, the road generation algorithm executed, and block types created, would be a good estimate of what these would be in the future. That way population and demographic forecasts could be more accurate and the way policies influence these urban forms might be evaluated. The methods for modelling local road network type developed here could help achieve this. First, they would need to be integrated into integrated modelling tools or city generation engines, in addition to the datasets on which they depend. This effort is probably not justified by the predictive capability of the road network type models estimated here. Still the methods used here provide a good basis for improvement. Incorporating more knowledge into the models about how neighborhood design decisions are made, like the influence of road construction costs and value of different housing types, could stand to improve them. In addition, incorporating additional policy constraints as variables into the model, such as any requirements for multifamily housing, would surely enhance them as well.

There is also potential to improve the way spatial dependence was incorporated into the modelling process used in this study. Here, it was incorporated through a binary variable indicating whether or not the sample area has at least one neighbor of the same road network type. Using a similar approach as in this research, it would have been possible to calculate the number of neighbors or area of neighbors of the same type as a given sample area. Then these could have been incorporated into the model as continuous rather than binary variables. Alternatively, a spatial multinomial logit model according to one of the methods described in the methodology section could also be used. Considering the potential for strong spatial dependence between features of the urban system that are based on geometry or form, such as local road network type, this could prove worthwhile.

Though the quality of the modelling results could be improved, doing so will become increasingly relevant as walking scale indicators of travel behaviour are incorporated into integrated modelling tools (Waddell, 2009). That way the effects of policy on walkability and dispersion of future neighourhoods might be assessed. Similarly, as integrated modeling tools become capable of assessing parcel-level policies related to urban sprawl, a more accurate representation of parcel characteristics and patterns should facilitate this.

References

Adler, R. (2006). In Defense of NEPA: The Case of the Legacy Parkway. *Journal of Land, Resources, & Environmental Law, 26*(2).

Aicholzer, O., & Aurenhammer, F. (1995). A Novel Type of Skeleton. *Journal of Universal Computer Science, 1*(12), 752-761.

Alexandridis, K., & Pijanowski, B. (2007). Assessing Multiagent Parcelization Performance in MABEL Simulation Model Using Monte Carlo Replication Experiments. *Environment and Planning B: Planning and Design, 34*(2), 223-244.

Aliaga, D., Vanegas, C., & Benes, B. (2008). Interactive Example-Based Urban Layout Synthesis. *ACM Trans. Graph. , 27*(5).

AMT. (2008) Montreal Transit Stops. Agence métropolitaine de transport. Montréal, QC.

Anas, A., & Arnott, R. J. (1993). Development and testing of the Chicago Prototype Housing Market Model. *Journal of Housing Research, 4*(1), 73-129.

Anguita, D., Ghio, A., Oneto, L., & Ridella, S. (2012). In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Trans Neural Netw Learn Syst, 23*(9), 1390-1406.

Anselin, L. (2001). Spatial Econometrics. In B. H. Baltagi (Ed.), A Companion in Theoretical Econometrics. Malden, MA: Blackwell Publishing Ltd. .

Ben-Akiva, M., & Lerman, S. (1987). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, Massachusetts: MIT Press.

Bierlaire, M. (2003). *BIOGEME: a free package for estimation of discrete choice models*. Paper presented at the 3rd Swiss Transport Research Conference Monte Verita, Ascona.

Borges, S. (1996). Taxonomic key and annotated bibliography to the earthworms of Puerto Rico. *Annals of the New York Academy of Sciences, 776*, 249-256. doi: 10.1111/j.1749-6632.1996.tb17424.x

Borning, A., & Davis, J. (2005). *Three Contexts of Appropriation for an Urban Simulation System*. Paper presented at the 9th European Conference on Computer Supported Cooperative Work, Paris, France.

Borning, A., Waddell, P., & Förster, R. (2008). Digital Government: E-Government Research, Case Studies, and implementations. In H. Chen, L. Brandt & V. Gregg (Series Eds.),

Bowker, G., & Star, S. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, Massachusetts: The MIT Press.

Burton, E. J., Mitchell, L., & Stride, C. B. (2011). Good places for ageing in place: development of objective built environment measures for investigating links with older people's wellbeing. *BMC Public Health, 11*, 839-851. doi: 10.1186/1471-2458-11-839

Carrión-Flores, C., Flores-Lagunes, A., & Guci, L. (2009). *Land Use Change: A Spatial Multinomial Choice Analysis* Paper presented at the World Conference on Spatial Econometrics, Barcelona, Spain.

Cervero, R., & Murakami, J. (2010). Effects of built environments on vehicle miles traveled: evidence from 370 US urbanized areas. *Environment and Planning A, 42*, 400-418.

Chen, G., Esch, G., Wonka, P., Muller, P., & Zhang, E. (2008). *Interactive Procedural Street Modeling*. Paper presented at the Proceedings of the 35th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY. USA.

Cherry, N., & Nagel, K. (2009). *Grid / Street / Place: Essential Elements of Sustainable Urban Districts*. Chicago, IL: American Planning Association Press.

Cherry, N., & Nagle, K. (2009). *Grid / street / place : essential elements of sustainable urban districts* (1st ed.). Chicago, IL: APA Planners Press.

Clay, M. J. (2010). Developing an integrated land-use/transportation model for small to medium-sized cities: case study of Montgomery, Alabama. *Transportation Planning and Technology, 33*(8), 679-693. doi: 10.1080/03081060.2010.527178

CMM. (2005). Heritage Policy: City of Montreal.

CMM. (2010). Portrait of Greater Montréal *Metropolitan Reports* (Vol. 1): City of Montreal.

CMM. (2012) Données numériques de la cartographie d'utilisation du sol. Communauté Métropolitaine de Montréal. Retrieved from http://cmm.qc.ca/

Conover, C. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.

Crawley, M. J. (2013). *The R Book* (2nd ed.). West Sussex, United Kingdom: John Wiley & Sons.

Dahal, K. R., & Chow, T. E. (2014). A GIS toolset for automated partitioning of urban lands. *Environmental Modelling & Software, 55*, 222-234. doi: 10.1016/j.envsoft.2014.01.024

De Palma, A., Picard, N., & Motamedi, K. (2014). Application of UrbanSim in Paris (Ile-De-France) Case Study *cahier de recherche*.

Ding, C. (2001). An Empirical Model of Urban Spatial Development. *Review of Urban & Regional Development Studies, 13*(3).

DMTI. (2013) CanMap Route Logistics. DMTI Spatial Inc. Markham, Ontario.

Dufresne, M., Dufaux, F., Stott, M., Coulu, E., & Perdriel, E. (2003). Étude de caractérisation de la forme urbaine montréalaise, secteur Centre: Groupe Cardinal Hardy

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tetreault, P., & Surprenant-Legault, J. (2014). New evidence on walking distance to transit stops: Identifying redundancies and gaps using variable service areas. *Transportation 41*(1), 193-200.

ESRI. CityEngine. from http://www.esri.com/software/cityengine/index.html

FQ. (2009) Les données du cadastre du Québec en couverture ArcInfo. Foncier Québec Québec, Québec. Retrieved from www.mrnf.gouv.qc.ca/foncier/

Garde, A. (2008). City Sense and Suburban Design: Planners' Perceptions of the Emerging Suburban Form. *Journal of the American Planning Association, 74*(3), 325-342.

GeoBase. (2010) National Road Network (NRN). Geogratis. Sherbrooke, Québec. Retrieved from http://geogratis.gc.ca/

GeoBase. (2015) CanVec+. Geogratis. Sherbrooke, Québec. Retrieved from http://geogratis.gc.ca/

Glass, R. K., Morkel, C., & Bangay, S. D. (2006). *Duplicating road patterns in South African informal settlements using procedural techniques*. Paper presented at the Proceedings of the 4th international conferenece on Computer graphics, virutal reality, visualization and interaction in Africa AFRIGRAPH, Cape Town, South Africa http://doi.acm.org/10.1145/1108590.1108616

Gordon, D., & Shirokoff, I. (2014). Suburban Nation? Population Growth in Canadian Suburbs, 2006-2011: School of Urban and Regional Planning, Queen's University

Griffith, E. C. (1907). *The rise and development of the gerrymander.* PhD., The University of Chicago, Chicago. Retrieved from https://archive.org/details/risedevelopmento00grif

Griffiths, W. E., Hill, R. C., & Judge, G. C. (1993). *Learning and Practicing Econometrics*. Hoboken, New Jersey: John Wiley & Sons.

GVRD. (1975). Livable Region 1976/1986 Proposals to manage the growth of Greater Vancouver Regional District

Harris, R. C. (1984). *The Seigneurial System in Early Canada: A Geographical Study*: McGill-Queens University Press.

Hassan, M., Joo, Y., & Jun, C. (2010). A Review of the Development and Application of UrbanSim Integrated Land-Use and Transportation Model. *International Journal of Urban Studies, 14*(3), 295-306.

Hillier, B. (1996). Cities as movement economies. *Urban Design International, 1*(1), 41-60.

Hunt, J. D. (1994). Calibrating the Naples land use and transport model. *Environment and Planning B: Planning and Design, 21*, 569-590.

Hunt, J. D., & Abraham, J. E. (2007). PECAS - for Spatial Economic Modelling: Theoretical Formulation *System Documentation Technical Memorandum* (pp. 1-129). Calgary: HBA Specto.

Hunt, J. D., Kriger, D. S., & Miller, E. J. (2005). Current operational urban land‐use‐transport modelling frameworks: A review. *Transport Reviews, 25*(3), 329-376. doi: 10.1080/0144164052000336470

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature, 22*(4), 323-340. doi: 10.1177/0885412207314010

Irwin, E. G., Bell, K. P., & Geoghegan, J. (2003). Modeling and Managing Urban Growth at the Rural-Urban Fringe: A Parcel-Level Model of Residential Land Use Change *Agricultural and Resource Economics Review 32*(1), 19.

JFP. (2016). Jepson Flora Project. *Jepson eFlora*. Retrieved April 13, 2016, from http://ucjeps.berkeley.edu/IJM_fam_key.html

Kelly, G., & McCabe, H. (2007). *Citygen: An Interactive System for Procedural City Generation*. Paper presented at the The Fifth Annual International Conference in Computer Game Design and Technology, Liverpool, UK.

Klier, T., & McMillen, D. P. (2008). Clustering of Auto Supplier Plants in the United States: Generalize Method of Moments Spatial Logit for Large Samples. *Journal of Business Economics and Statistics, 26*(4), 460-471.

Kryvobokov, M., Mercier, A., Bonnafous, A., & Bouf, D. (2013). Simulating housing prices with UrbanSim: predictive capacity and sensitivity analysis. *Letters in Spatial Resource Sciences, 6*(1), 31-44.

Lavalle, S. M. (1998). Rapidly-exploring random trees: A new tool for path planning *TR* (Vol. 98). Iowa: Iowa State University.

Lechner, T., Watson, B., Brozefski, C., & Wilensky, U. (2006). *Procedural Modeling of Urban Land Use*. Paper presented at the SIGGRAPH '06: ACM SIGGRAPH 2006, New York, NY.

Lee, S., Tremble, M., Vaivai, J., Rowangould, G., Tayarani, M., & Poorfakhraei, M. (2015). Central New Mexico Climate Change Scenario Planning Project: Final Report.

Lewis, S. (1998). Land use and transportation: Envisioning regional sustainability. *Transport Policy, 5*, 147-161.

Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Lovegrove, G., & Sayed, T. (2006). Macro-level collision prediction models for evaluating neighbourhood traffic safety. *Can. J. Civ. Eng., 33*, 609-621.

MAMROT. (2011) Localisation des Immeubles. Direction du Bureau municipal, de la géomatique et de la statistique. Montréal, Québec. Retrieved from http://www.mamrot.gouv.qc.ca/

Marshall, S. (2005). *Streets and Patterns*. New York, NY: Spon Press.

Martínez, F. J. (1996). MUSSA: a land use model for Santiago City. *Transportation Research Record*(1552), 126-134.

McFadden, D. (1980). Economic Models for Probabilistic Choice Among Products. *The Journal of Business, 53*(3).

McGarigal, K., & Marks, B. J. (1994). Fragstats: Spatial Pattern Analysis Program for Quantifying Landscape Structure

Miller, E. J. (2003). Land Use: Transportation Modeling. In K. G. Goulias (Ed.), *Transportation Systems Planning: Methods and Applications* (pp. 24): CRC Press.

Modelistica. (1995). TRANUS Integrated Land Use and Transport Modeling System (Version 5.0). Retrieved from http://www.odot.state.or.us

Mohammadian, A., & Kangarolou, P. (2003). *Applications of Spatial Multinomial Logit Model to Transportation Planning*. Paper presented at the 10th International Conference on Travel Behaviour Research Lucerne.

Newburn, D., & Berck, P. (2006). Modeling Suburban and Rural-Residential Development Beyond the Urban Fringe. *Land Economics 82*(4), 481-499.

Newman, P., & Kenworthy, W. (1989). Gasoline Consumption and Cities. *Journal of the American Planning Association, 55*(1), 24-37. doi: 10.1080/01944368908975398

Parish, Y., & Muller, P. (2001). *Procedural Modeling of Cities*. Paper presented at the Proceedings of the 28th annual conference on Computer graphics and interactive techniques, Los Angeles, CA.

Patterson, Z., & Bierlaire, M. (2010). Development of prototype UrbanSim models. *Environment and Planning B: Planning and Design, 37*(2), 344-368.

Patterson, Z., Saddier, S., Rezaei, A., & Manaugh, K. (2014). Use of the Urban Core Index to analyze residential mobility: the case of seniors in Canadian metropolitan regions. *Journal of Transport Geography, 41*, 116-125.

Peponis, J., Allen, D., French, S., Scoppa, M., & Brown, J. (2007). *Street Connectivity and Urban Density: spatial measures and their correlation*. Paper presented at the Proceedings of the 6th Interational Space Syntax Symposium Istanbul.

Pinske, J., & Slade, M. E. (1998). Contracting in Space: An Application of Spatial Statistics to Discrete-Choice Models. *Journal of Econometrics, 85*, 125-154.

PIPPR. Pioneer Institute Public Policy Research. (2005). Housing Regulation Database Massachusetts Municipalities from http://www.masshousingregulations.com/

Pocewicz, A., Nielsen-Pincus, M., Goldberg, C. S., Johnson, M. H., Morgan, P., Force, J. E., Waits, L. P., & Vierling, L. (2007). Predicting land use change: comparison of models based on landowner surveys and historical land cover trends. *Landscape Ecology, 23*(2), 195-210. doi: 10.1007/s10980-007-9159-6

Putnam, S. H. (1996). Extending DRAM Model: theory-practice nexus. Transportation Research Record. *Transportation Research Record*(1552), 112-119.

Rifaat, S. M., & Tay, R. (2008). *Effect of Street Pattern on the Severity of Crashes Involving Vulnerable Road Users*. Paper presented at the Proceedings of the 18th Canadian Multidisciplinary Road Safety Conference Whistler, British Columbia.

Rodriguez, D. A., & Joo, J. (2004). The relationship between non-motorized mode choice and the local physical environment. *Transportation Research Part D, 9*, 151-173.

Ruxton, G. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioural Ecology, 17*(4), 688-690.

Sandalack, B. A., Alaniz Uribe, F. G., Eshghzadeh Zanjani, A., Shiell, A., McCormack, G. R., & Doyle-Baker, P. K. (2013). Neighbourhood type and walkshed size. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 6*(3), 236-255. doi: 10.1080/17549175.2013.771694

SCFS. (2016). Plants of Sagehen Creek Basin, CA - Wildflowers. Retrieved April 13, 2016, from http://www.inaturalist.org/guides/1104

Schirmer, P. (2010). *Options and constraints of a parcel based approach to 'UrbanSimE'*. Paper presented at the Swiss Transportation Research Conference, Monte Verita/Ascona.

Shneider, L. (Ed.) (1985). Québec: Les Presses de l'Université Laval.

Smelik, R. M., de Kraker, K. J., Tutenel, T., Bidarra, R., & Groenewegen, S. A. (2009). *A survey of procedural methods for terrain modelling*. Paper presented at the Proceedings of the CASA Workshop on 3D Advanced Media in Gaming and Simulation, Amsterdam, The Netherlands.

Smith, B., & Sandwell, D. (2003). Accuracy and resolution of shuttle radar topography mission data. *Geography Research Letters, 30*(9). doi: 10.1029/2002GL016643

Snellen, D., Borgers, A., & Timmermans, H. (2002). Urban form, road network type, and mode choice for frequently conducted activities: a multilevel analysis using quasi-experimental design data. *Environment and Planning A, 34*, 1207-1220.

Song, Y., & Quercia, R. (2008). How are neighbourhood design features valued across different neighbourhood types? *Journal of Housing and the Built Environment 23*(4), 297-316.

Southworth, M., & Ben-Joseph, E. (1993). *Regulated streets: the evolution of standards for suburban residential streets*. Berkeley, California: University of California Berkeley.

Southworth, M., & Ben-Joseph, E. (2003). *Streets and the Shaping of Towns and Cities*. Washington D.C.: Island Press.

Southworth, M., & Owens, P. (1993). The evolving metropolis: Studies of community, neighborhood, and street form at the urban edge. *Journal of the American Planning Association, 59*(3), 271-287.

StatsCan. (2006) Total number of occupied private dwellings by period of construction - 20% sample data for census tracts, 2006 census. CHASS. Retrieved from http://dc.chass.utoronto.ca/census/2006/index.html

Sun, J., & Lovegrove, G. (2013). Comparing the road safety of neighbourhood development patterns: traditional versus sustainable communities. *Can. J. Civ. Eng., 40*, 35-45.

Sun, J., Yu, X., Baciu, G., & Green, M. (2002). *Template-based Generation of Road Networks for Virtual City Modeling*. Paper presented at the Proceedings of the ACM symposium on Virtual Reality Software and Technology New York, NY. USA.

Synthicity. UrbanCanvas. from http://www.synthicity.com/urbancanvas/

Tasker-Brown, J., & Pogharian, S. (2000). Learning from Suburbia: Residential Street Pattern Design *Socio-economic Series* (pp. 1-6). Ottawa, Canada: Canada Mortgage and Housing Corporation

Timme, S. L. (1991). *How to construct and use a dichotomous key* Paper presented at the Proceedings of the 12th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).

Tomalty, R. (1997). *The compact metropolis: Growth management and intensification in Vancouver, Toronto and Montreal*: ICURR Publications.

Vanegas, C. A., Aliaga, D. G., Beneš, B., & Waddell, P. (2009a). Visualization of Simulated Urban Spaces: Inferring Parameterized Generation of Streets, Parcels, and Aerial Imagery. *IEEE Transactions of Visualization and Computer Graphics 15*(3), 424-435.

Vanegas, C. A., Aliaga, D. G., Beneš, B., & Waddell, P. A. (2009b). Interactive design of urban spaces using geometrical and behavioral modeling. *ACM Transactions on Graphics, 28*(5). doi: 10.1145/1618452.1618457

Vanegas, C. A., Aliaga, D. G., Wonka, P., Müller, P., Waddell, P., & Watson, B. (2010). Modelling the Appearance and Behaviour of Urban Spaces. *Computer Graphics Forum, 29*(1), 25-42. doi: 10.1111/j.1467-8659.2009.01535.x

Vanegas, C. A., Kelly, T., Weber, B., Halatsch, J., Aliaga, D. G., & Muller, P. (2012). Procedural Generation of Parcels in Urban Modeling. *Computer Graphics Forum, 31*(2), 681-690. doi: 10.1111/j.1467-8659.2012.03047.x

Waddell, P. (2009). *Parcel-Level Microsimulation of Land Use and Transportation: The Walking Scale of Urban Sustainability*. Paper presented at the Workshop on Computational Algorithms and Procedures for Integrated Microsimulation Models, Jaipur, India.

Waddell, P. (2010). Modeling Residential Location Choice in UrbanSim. In F. Pagliara, J. Preston & D. Simmonds (Eds.), *Residential Location Choice: Models and Applications* (pp. 165-180): Springer Berlin Heidelberg.

Waddell, P., Moore, T., & Edwards, S. (1998). *Exploiting Parcel-Level GIS for Land Use Modeling*. Paper presented at the ACSE Conference on Transportation, Land Use and Air Quality, Portland, Oregon

Waddell, P., Ulfarsson, G. F., Franklin, J. P., & Lobb, J. (2007). Incorporating land use in metropolitan transportation planning. *Transportation Research Part A: Policy and Practice, 41*(5), 382-410. doi: 10.1016/j.tra.2006.09.008

Waddell, P., & Ulfrasson, F. (2004). *Introduction to Urban Simulation: Design and Development of Operational Models* (Vol. 5): Pergamon Pres.

Waller, L., & Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*: Wiley-Interscience.

Wang, F. (1998). Urban population distribution with various road networks: a simulation approach *Environment and Planning B: Planning and Design, 25*, 265-278.

Weaver, W. (1948). Science and Complexity. *American Scientist, 36*(536).

Weber, B., Muller, P., Wonka, P., & Gross, M. (2009). Interactive Geometric Simulation of 4D Cities. *EUROGRAPHICS, 28*(2).

Weeks, J. R. (2004). The role of spatial analysis in demographic research. In M. G. a. D. Janelle (Ed.), *Spatially Integrated Social Science: Examples in Best Practice*. New York: Oxford University Press.

Wegener, M. (2004). Overview of Land-Use Transportation Models. In A. Hensher & K. Button (Eds.), *Transport Geography and Spatial Systems* (pp. 127-146). Kidlington, UK: Pergamon/Elsevier Science.

Wickramasuriya, R., Chisholm, L., Puotinen, M., Gill, N., & Klepeis, P. (2013). A method to dynamically subdivide parcels in land use change models. *International Journal of Geographic Information Science, 27*(8), 1497-1513.

Wickramasuriya, R., Chisholm, L. A., Puotinen, M., Gill, N., & Klepeis, P. (2011). An automated land subdivision tool for urban and regional planning: Concepts, implementation and testing. *Environmental Modelling & Software, 26*(12), 1675-1684. doi: 10.1016/j.envsoft.2011.06.003

Winters, M., Brauer, M., Setton, E. M., & Teshke, K. (2010). Built Environment Influences on Healthy Transportation Choices: Bicycling versus Driving. *Journal of Urban Health: Bulletin of the New York Academy of Medicine 87*(6).

Wiseman, N., & Patterson, Z. (2016a). *Modeling Road Network Type*. Paper presented at the Proceedings of the 95th Annual Meeting of the Transportation Research Board, Washington, DC.

Wiseman, N., & Patterson, Z. (2016b). Testing block subdivision algorithms on block designs. *Journal of Geographical Systems, 18*(1), 17-43. doi: 10.1007/s10109-015-0222-6

Xie, Y., & Batty, M. (2003). Integrated Urban Evolutionary Modeling. *UCL Working Paper Series*. Retrieved from Center for Advanced Spatial Analysis website: http://www.bartlett.ucl.ac.uk/casa

Appendix A

Dichotomous keys for road network type classification

Gridiron

Gridiron Description

- Parallel square or short rectangular blocks forming a continuous mesh
- Characteristic feature is high street connectivity and accessibility
- Slight street misalignment is ok
- Some missing streets are ok
- Different length blocks are ok

Border Cases

- Fragmented Parallels are essentially incomplete Gridirons
- How much discontinuity must there be before something is classified as Fragmented Parallels?

Gridiron – Fragmented Parallels (Dichotomous key)

1a) Blocks mostly short (1:2/1:3 width to length ratio)……… 2

1b) Blocks mostly long (1:4/1:5 width to length ratio)…........ Fragmented

2a) Proportion of 4-way intersections >= 0.75………………. Gridiron

2b) Proportion of 3-way intersections >= 0.75………………. Fragmented

2c) Proportion of 3-way and 4-way intersections roughly equal..3

3a) Blocks mostly parallel (>=⅔ by area)................................ Gridiron

3b) Some blocks perpendicular (>=⅓ by area)....................... Fragmented

Fragmented Parallels

Fragmented Parallels Description

- Blocks mainly rectangular and parallel or at 90 degree angles
- Blocks at 90 degree bends (ie: L, S, T, U)
    - Can have rounded edges if they're at 90 degrees
- Characteristic feature is limited access/high proportion of 3-way intersections
- Includes grids with high proportion (=> 75%) of 3-way intersections
    - If number of 3-way and 4-way intersections are ~equal, see Gridiron vs. Fragmented decision tree
- Conspicuous street misalignment
- Long block lengths (1:4, 5) - signifies a conscious effort to limit access to middle of block and discourage through traffic

Border Cases

- Distinction between fragmented parallels and warped parallels can be made based on whether or not walkability/direct connectivity is being sacrificed for the sake of visual interest in the form of non-straight sight lines
- Can still be considered fragmented if some streets are warped (<¼)
- Need to recognize when warped streets are a design feature (warped parallel) vs. when they're there to accommodate an irregularity in the landscape (ie: park or curved arterial) (fragmented)
- Sometimes accommodating an irregularity becomes the basis for a design (ie: not correcting an irregularity in an arterial with streets further within the block)

Special Cases

- If it has a lollipop can it really be considered fragmented? (yes, if <= 2)
- Series of terraces within encircling streets (if at right angles)
- If local roads have no intersecting streets between arterials

<u>Fragmented Parallels – Warped Parallels (Dichotomous key)</u>

1a) Blocks are straight…………..Fragmented

1b) Blocks partly warped……….2

2a) Only bounding streets are warped………………Fragmented

2b) Some interior streets are warped……….3

3a) Warped streets are bent <= 15 degrees…………Fragmented

3b) Some Warped streets are bend > 15 degrees or curved…………..4

4a) Straight blocks comprise =>75% of blocks………….Fragmented

4b) Warped blocks compose > 25% of blocks…………...Warped

Warped Parallels

Warped Parallels Description

- Purpose of design is to limit the visual length of streets
- Blocks are mostly curved or bent at non-90 degree angles (ie: warped)
- Blocks follow a grain (multiple parallel streets aligned in a general direction)
- Blocks tend to be curved to the contours of arterial/collector streets
- Some streets lining open space like parks, so houses overlooking parkland
- Lollipops on a stick are considered warped if stick is warped (ie: curved or bent at non 90 degree angle)

Border Cases

- If the only alignment is between an inner and outer blocks (loops & lollipops)
- A lollipop is often present in the center and then more than one block is wrapped around it (warped)
- Through traffic is possible, but rarely direct (warped) vs. access across entire length of patch is only possible through collectors (loops & lollipops)
- Distinction between design to achieve a shorter visual length of streets (warped) vs. design to create privacy, isolation between blocks (loops & lollipops)

Warped Parallels – Loops and lollipops – New Urbanism (Dichotomous key)

1a) Access from one far corner to another is only through bounding streets…. Loops and lollipops
1b) Access from one far corner to another through interior streets…………. 2

2a) Greatest number of aligned blocks = 3………...Loops and lollipops
2b) Greatest number of aligned blocks => 6………...Warped
2c) Greatest number of aligned blocks >3 AND <6…3

3a) Ratio of loops and cul-de-sacs to blocks => ¼…. 4

3b) Ratio of loops and cul-de-sacs to blocks < ¼…… 5

4a) Most lollipops connect with pedestrian paths or alleys…… New Urbanism

4b) Most lollipops don't connect with pedestrian paths or alleys… Loops and lollipops

5a) Cul-de-sac count <= 2 features……………………............... Warped

5b) Cul-de-sac count > 2 features………………………………... 5

6a) Loops and lollipops form and interconnected pedestrian network….New Urbanism

6b) Loops and lollipops don't form an interconnected pedestrian network.. Warped

Mixed

Mixed Description

- Includes different road network types that happen to be in the same patch
- Also includes new-urbanism blocks where elements of different road types are combined to form an aesthetic, walkable and drivable community
- These will be excluded from the model

Mixed – Uniform (Dichotomous Key)

1a) Elements of an adjacent type on grid-dendritic spectrum <= 1/4 of total….. Uniform
1b) Elements of an adjacent type on grid-dendritic spectrum > ¼ of total……… 2

2a) Number of elements of a different type not adjacent on spectrum <= 2……. Uniform
2b) Number of elements of a different type not adjacent on spectrum > 3……… 3

3a) Ratio of different elements to dominant elements < 15%.............................. Uniform
3b) Ratio of different elements to dominant elements > 15%.............................. Mixed