Using network analysis to explore the effects of road network on traffic congestion and retail store sales

by

Junyi Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The physical road system plays a critical role in environmental and city planning. In the context of retail store site-selection, measures of accessibility and the ease and willingness of consumers to shop at a store can be essential to revenue generation and retail success. To quantify accessibility requires a detailed examination of the road networks and in many cases modelling to estimate potential traffic congestion that would inhibit accessibility. The application of network theory to assess the accessibility of road segments and land parcels is non-existent. Research on the effects of the structure of the road network, via network analysis, can facilitate identifying potential congestion issues and subsequently the effects of congestion on commercial performance (e.g., retail sales). The application of network analysis to a road network is distinctive from applications in other disciplines (e.g., sociology, ecology), since, among other network attributes, the road network is a low-dimension, link-centroid, and relatively static system with time-variant traffic flow. In addition to conceptually interrogating the difference between social and road networks for network analysis, the presented research results show the relationship among different network metrics and simulated traffic congestion and the strength of the relationship between network metrics and retail sales relative to socio-demographic and site-location characteristics.

Acknowledgements

Foremost I would like to express my gratefulness towards my supervisor Dr. Derek Robinson. Without your guidance and patience through the years, I can never finish this work.

Secondly, I would like to thank my other committee member Dr. John McLevey. Thank you for having good discussions and supporting me in this thesis.

Lastly, I want to thank my friends, colleagues, and family. Thanks, Zhengyang, for encouraging and inspiring me during the hard times. Thanks, Bo and Li, for sharing ideas and comments. Thanks, HaPi! For being such an adorable companion. And I thank my parents, for your selfless love and caring.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction to location theory and the role of the road network in site selection

1.1. Background

1.1.1. Location theory and retail

The history of location theory spans from ancient periods to the Renaissance, through the first industrial revolution, and continues to develop in contemporary times. In the Shang dynasty of ancient China (1556 BC to 1046 BC), feng-shui philosophy was developed as a systematic location theory for landscape arrangement and site selection of capitals, towns, and buildings (Hong, Song, & Wu, 2007); in the seventeenth century, location selection became an analytical science when Evangelista Torricelli devised a solution to Pierre de Fermat's geometric median problem, which finds the optimal location to minimize the distance to given points in a Euclidean space (Krarup & Vajda, 1997); and in the nineteenth century, location selection was formulated as the land rent theory in The Isolated State by von Thunen (1826). It was not until the progression of social sciences before First World War that location theory was formally brought to the stage of economic geography by Alfred Weber, who reintroduced location theory and contributed to industrial location models. Contemporary concepts of location theory have been broadly implemented in many disciplines (e.g., economics, biology and ecology; Martin & Roper, 1988); however, facility site-selection remains one of the principle fields of location study.

Retail site-selection is an application of location theory (Goodchild, 1984). A retail store's location is critical to a retail business for it determines the store's accessibility, defines the market area, and is exposed to the community's atmosphere (Huff, 2003). According to the North American Industry Classification System (NAICS), the retail sectors comprises merchandises from twelve subsectors, from food, cloth, and hobby stores, to auto parts, furniture, and home improvement supplies dealers, including small shops and big-box chain stores, and even non-store retailers. (Statistics Canada, 2016). While each retailer aims to maximize business profitability by allocating stores as intermediates between warehouses and prospective customers, the allocation strategy various among the types of retail. One could argue that the location decision for small merchandises is flexible and has more alternatives relative to the relatively inelastic decisions of big-box retailers who make large, fixed, and often long-term investment in store location choices. Moreover, stores in the home improvement sector require

higher driving accessibility since customers are more likely to drive for shopping, and have less incentive to be located at the city centres because of the high rent, parcel size and availability, operational costs, and customers are willing to travel farther for large and less frequent purchases.

Most retail site selection decisions primarily depend on the knowledge and subjective judgements of owner, corporation, or real-estate decision-makers (Evans, 2011; Fowler, 2016). However, the development of site-selection techniques and increased computational capacity have enabled a large variety of models, algorithms, and formulations to solve the site-selection problem in the retail sector using a number of different criteria.

Centuries after von Thunen (1826) first formalized location theory, the list of site-selection criteria has diversified (Weber, Current, & Benton, 1991; Hoffman & Schniederjans, 1994; Badri, 1999; Sarkis & Sundarraj, 2002; MacCarthy & Atthirawong, 2003). Facility site-selection is divided into two stages: country or region selection and community or site selection (Brown & Gibson, 1972; Hoffman & Schniederjans, 1994).

During regional selection, criteria are given consideration based on factors in a broad context, for example, the regional socio-economic, political, technical, natural and market factors. Specifically, the economic factors reflect monetary policy and economic conditions in a country (e.g., tax rates and average wage); the social factors reveal the social atmosphere (e.g., demography and crime rate); the policy factors include market regulations, restrictions, and incentives (e.g., potential trade barriers and tax relief); technical factors indicate the regional technology level (e.g., innovation cycle and cost); the natural factors include regional climate and natural disaster frequency and extent; and the market environment factors include competition and market potential, among other factors (Hoffman & Schniederjans, 1994).

At the community or site selection stage, criteria should be focused on site attributes such as the parcel size, operation and construction cost, accessibility, visibility, neighbourhood environment, local demand, competitor density and prestige. These criteria can be classified into three groups. The first group includes critical (or location) criteria, which directly impact or restrict the facility's location regardless of the existence of other conditions. For example, a retail

store is allocated close to the market demand points, and a manufacturing facility requires adequate labor supply nearby. The second group consists of objective criteria, which can be measured by crisp and precision-based monetary terms. Such criteria include labor, construction, and transportation costs. The last group includes subjective criteria with qualitative definitions in linguistic description. Unusually, such criteria may not be directly and precisely evaluated by numeric values. For example, shopping distances can be assessed by "farther" and "closer" (McGuirt, et al., 2014).

Depending on the available data, the data input for a site-selection problem is either precision-based or fuzzy-based (Liang & Wang, 1991; Weber, Current, & Benton, 1991). Precision-based data has an explicit numeric formation, such as the construction or travel cost, while fuzzy-based data is usually expressed via linguistic variables and describes a conceptual status, such as preference, willingness and likelihood. Additionally, the fuzzy data can be transferred into precise numbers. For example, when expert or stakeholder opinions are integrated in a site-selection problem, the data input is often intuitive and prone to inconsistencies. In this case, methods like analytical hierarchy process (AHP) procedure can be used to convert fuzzy input into crisp numeric data (Zadeh, 1965; Liang & Wang, 1991; Rangone, 1996; Siddiqui, Everett, & Vieux, 1996; Charnpratheep, Zhou, & Garner, 1997; Cheng, Chan, & Huang, 2002; Mahler & De Lima, 2003).

1.1.2. Site-selection problems and techniques

The overarching goal of a retail site-selection problem is often to minimize cost and maximize revenue to maximize profitability. Therefore, store site-selection problem can be broken into three secondary goals (Brown & Gibson, 1972). The first goal of site-selection is to minimize the production cost. It is one of the main concerns in the early works of facility site-selection (Greenhut, 1956; Hoover, 1967). This site-selection strategy integrates fixed production costs like rent (Thünen, 1826) and labor (Weber A. , 1909) and is mainly achieved by analytical models.

The second site-selection goal is to minimize distribution cost. It is mostly rooted in service providers or vendors whose main concern is the travel time and distance between the facility and the customers (Brown & Gibson, 1972). Many popular site-selection problems were

proposed based on travel costs, such as the single facility problem, the multi-facility problem, the p-median problem[1] and the p-centre[2] problem (Cooper, 1964; Hakimi, 1964).

The third goal of retail site-selection is to maximize revenue. Methods of estimating customer expenditure or market potential are often based on the assumption of homo economicus, where the rational customers choose shopping behavior for their maximum benefits. A comprehensive site-selection strategy should integrate all the criteria including cost and revenue to optimize store profitability (Greenhut, 1956; Brown & Gibson, 1972).

Beyond the three site-selection strategies, some facility location problems require an optimal coverage of demand locations, especially in the public sector, such as fire stations, and hospitals. Such location problem is formulated as a covering problem (Toregas, Swain, ReVelle, & Bergman, 1971; Church & ReVelle, 1974).

To facilitate site location selection problems, mathematical models, heuristic algorithms, and multi-criteria/multi-attribute decision making (MCDM/MADM) methods have been proposed (Onut, Efendigil, & Kara, 2010). The mathematical models are feasible if precision-based data is available. For example, integer programming is a basic mathematical optimization technique. It uses analytical models (e.g., linear and non-linear regression) as objective functions and constraints to find an optimal solution (Hillsman, 1984; Kao & Lin, 1996; Nema & Gupta, 1999; Chang & Wei, 2000).

Heuristic algorithms are applied to find the best approximate solution when a location problem does not have a known solution in a polynomial time (i.e., NP-complete). Meta heuristics is a high-level procedure that coordinates known solutions (including heuristic) to avoid local

[1] P-median is an optimization strategy of facility location problem where the objective is to optimize the total utility by minimizing the total travel distance from the demand points to the facilities.

[2] P-centre is an optimization strategy of facility location problem where the objective is to optimize individual benefit by minimizing the maximum distance from the demand points to the facility.

optimization, the most well adopted methods include the genetic algorithm[3] (Holland, 1992), the Tabu search[4] (Glover, 1986; Hansen, 1986), the ant colony optimization[5] (Dorigo, Maniezzo, & Colorni, 1991) and simulated annealing[6] (Hebb, 1949; Minsky & Papert, 1969; Rumelhart, Smolensky, McClelland, & Hinton, 1986). Heuristic algorithms have been implemented in facility site-selection problems, for example, waste site selection (Ramu & Kennedy, 1994; Bautista & Pereira, 2006; Al-Jarrah & Abu-Qdais, 2006).

The MCDM/MADM method facilitates site selection by ranking and making a selection from site alternatives via an explicit evaluation against the criteria. It is especially effective when the criteria are conflicting, the inputs have inconsistent forms, or the data input is fuzzy (e.g., inputs have different units or qualitative formats; Liang & Wang, 1991).

1.2. Motivation

A road infrastructure system is an indispensable part of a contemporary urban system. It provides a physical connection between locations and is also the principal mediator of urban sustainability (Ford, Barr, Dawson, & James, 2015). As a vital part of the urban system, the road network is related to many aspects of the economy, such as industrial production, land prices, and resident quality-of-life (Verhoef, 2010; Boyle, Barrilleaux, & Scheller, 2014) via the efficiency of transportation (Frost & Spence, 1995; Gutierrez, Gonzalez, & Gomez, 1996; Gutierrez, 2001; Vickerman, Spiekermann, & Wegener, 1999).

One of the critical criteria for retail location is accessibility. For smaller retailers and retailers at the city centers, pedestrian accessibility might be a main concern; while for the big-

[3] The genetic algorithm is a heuristic method that uses the Darwinian evolution theory, where the population is updated by the best solution in each iteration via akin to natural selection, it also enables potential solutions to be explored through genetic crossover and mutation.
[4] The Tabu search is a meta-heuristic neighborhood search method that finds the best solution in the neighbours of the current solution in each iteration.
[5] Colony optimization is a meta-heuristic algorithm that was inspired by the pheromone in ant path seeking. In each iteration, the solution was updated with the bias from “pheromone” between elements.
[6] Simulated annealing is a meta-heuristic algorithm that finds the global optimum in a discrete space in a notion of “slow cooling”, which is achieved by a slow decrease in the probability of the acceptance of worse solutions.

box[7] retailers, especially retailers in the home improvement sector, driving accessibility is the main factor. Moreover, the efficiency of the transportation system determines the overall cost of a supply chain via distribution and logistics processes.

Travel cost has been a critical factor in site-selection problems since von Thunen proposed the rent model. However, travel cost should not be regarded as a simple function of distance (Hoover, 1967), but a function of the complex attributes of the road network (Allen, Liu, & Singers, 1993; Hull, Silva, & Bertolini, 2012). The principal objective of retail site-selection problems is to minimize travel cost between a location and the consumers via an optimization of store accessibility (Hakimi, 1964; Cooper, 1964; Goodchild, 1984; Arentze, Borgers, & Timmermans, 1996). Road infrastructure, which is the basis of travel behaviours, should have impacts on traffic flow and congestion. However, these impacts have not been explicitly incorporated in retail location problems.

The incorporation of network theory in the assessment of the effects of the road network on store revenue via accessibility and congestion is rare but is necessary to facilitate market research and to support retail store site selection. In site-selection problems, the spatial representation of locations and distances can be discrete, continuous, or network-based (Love, Morris, & Wesolowsky, 1988). Therefore, it is more appropriate to use road infrastructure as a network representation in a retail site-selection problem.

In previous site-selection or planning studies, the descriptions of road network patterns were macroscopic and subjective, for example, the use of terms “dense” or “major” (Marshall, 2005). Quantitative studies about road network structure should be integrated into site-selection problems; however, its integration has primarily been conducted in the field of transportation engineering (Möller & Schroer, 2014). Specifically, the infrastructure topology and geometry is used in traffic modelling, and some of the studies involved the implementation of network theory (Möller & Schroer, 2014).

[7] Big-box stores are located in large-scale buildings which footage normally exceed 50,000 square feet (Basker, Klimek, & Hoang Van, 2012).

Network theory has a wide application in the analysis of technological, social, informational and biological networks (Newman, 2010). Network measures reflect the relations between the entities and the information flow through the network. In road network analysis, studies have focused on heterogeneity (entropy), importance (centrality), connection pattern (Xie & Levinson, 2007; De Montis, Barthélemy, Chessa, & Vespignani, 2007), and fractal dimension (Rodin & Rodina, 2000; Lu & Tang, 2004).

However, the previous implementations of network theory in retail analysis mainly involved market organization, commercial activity (Jensen, 2006), and supply chain (Wagner & Neshat, 2010) rather than the impacts of road network on retail revenue generation (Coughlan & Grayson, 1998). The connection between retail and road network structure has been identified. For example, retail density and land-use intensity are found to have relationships with road network measures like centralities (Porta, et al., 2009), which suggests that network theory would provide a novel method for road network analysis and retail location analysis.

1.3. Research objectives

This Master's thesis was established based on a site-selection project for a big-box retail chain store, in which market potential and site suitability have been estimated and evaluated (Balulescu, 2015; Caradima, 2015). The overarching goal of this Master's thesis is to improve the decision-making capacity of site-selection problem for a retail store by investigating the relationship of different network metrics with simulated traffic congestion and the relative effect of these results (i.e., network attributes) on retail sales compared to socio-demographic and site-location characteristics. Within this context, this project aims to answer the following questions:

1) To what extent are network metrics correlated to traffic congestion?

2) Does the incorporation of road network metrics improve retail store sales modelling and if so what is their contribution relative to demographic or suitability analysis variables?

To answer these questions, this thesis starts with a description of the unique properties of road networks, followed by the methodology and results that provide insights into the relationship of network metrics with simulated traffic congestion. Then these results are used to

statistically evaluate the strength of the relationship between network attributes and retail sales relative to socio-demographic and site-location characteristics.

1.4. Thesis structure

This thesis is structured as two separate and independent papers (Chapter Two and Chapter Three). This chapter provided an introduction to location theory and the relationship between site selection and the road network. The history of site-selection studies and the role and importance of including the road network structure in retail site-selection was discussed to point out the gap in the explicit linkage of location and network analysis. It established two research questions to understand the relationships of road network metrics with traffic congestions and retail store sales. Chapter Two provides a primer on network theory and presents a list of quantitative network metrics applied to the City of Toronto road network and compared with simulated traffic data to identify the degree of correspondence with traffic congestion. Chapter Three presents an exploratory study of the use of road network metrics in store sales modeling comparing the effect of network metrics relative to demographic and suitability variables in three types of retail sales models to assess their relative impact on retail sales. Chapter Four presents a conclusion relevant to the improvement regained.

Chapter 2: Application of network analysis on road traffic congestion: A case study of City of Toronto Retail Store Road Networks

2.1. Introduction

The road infrastructure creates a physical network that connects locations and facilitates movement of everything from people to commerce and ideas. The structure of the road network affects the efficiency of the movement of a society (e.g., transportation) and therefore has a mutual effect on many local-regional aspects of markets and, more broadly, the economy (Frost & Spence, 1995; Gutierrez, Gonzalez, & Gomez, 1996; Gutierrez, 2001; Vickerman, Spiekermann, & Wegener, 1999). Such effects can be reflected via economic indicators, for example, industrial production, retail sales, land prices, and residents' quality-of-life (Verhoef, 2010; Boyle, Barrilleaux, & Scheller, 2014). The physical road infrastructure system also plays a critical role in environmental planning and city planning. Some have even ventured to label the transportation system as the principal mediator of urban sustainability (Ford, Barr, Dawson, & James, 2015).

Particularly, within the context of retail supply chain, the efficiency of transportation largely affects the success of retail stores. As retail sectors distribute goods from warehouses to stores and serve individual consumers or business entities, the travel cost caused by the separation of opportunities, services, and market demand in a distribution and logistics process has a significant impact on the overall cost of a supply chain (Allen, Liu, & Singers, 1993; Beamon, 1998; Hull, Silva, & Bertolini, 2012). From the customers' perspective, the efficiency of the road network near a store determines the store's accessibility, which influences the level of ease and willingness of consumers' shopping journey (Páez, Mercado, Farber, Morency, & Roorda, 2010; Öner, 2015) and therefore affects a customer's potential expenditure and store's attractiveness (Teller & Reutterer, 2008). Although an urban transportation system also serves for pedestrian and public accesses, driving access is mainly discussed in this thesis for the retail big-box stores in the home improvement sector.

The site-selection problem is formulated as the optimization of a store's accessibility given the dispersed demands for a retail store (Goodchild, 1984). Traditionally, the objective of site-selection is either to optimize the total utility by minimizing the total travel distance from the demand points to the facilities (i.e., p-median problem; Hakimi, 1964; Cooper 1964) or to

optimize individual benefit by minimizing the maximum distance from the demand points to the facility (i.e., p-centre; Hakimi, 1964). More complex spatial interaction models may integrate the behavior of consumers or the characteristics of store sites (e.g., Huff's model; Li & Liu, 2012). Road network is the basis of travel behaviors and the traffic conditions have marginal effect on a store's accessibility. However, road network and traffic conditions have not been extensively incorporated in retail site-selection problems.

The implementation of network theory leads to a novel method of road system analysis in the context of retail site selection. There are previous implementations of network theory in complex system analysis. For example, in the field of transportation engineering, network analysis has been applied on infrastructure topology for traffic modelling (Möller & Schroer, 2014); in the retail sector, it has been adopted to global and regional market analysis for market optimization (Iori, De Masi, Precup, Gabbi, & Caldarelli, 2008). Nevertheless, the practice of assessing the impact of road network structure on store revenue is virtually non-existent. Such study is necessary to facilitate market research and to support retail store site-selection.

A road network is distinctive from many other networks in sociology, ecology, or information technology. The representation of road network is a flat, edge-centroid, and static system that contains time-variant traffic information. For example, a national road network can be 99% planar with only a small number of overpasses due to the constraints on construction and maintenance costs. (Newman, 2010). It implies that road network studies, unlike other network studies, should not focus on degree distribution. Moreover, the road network heterogeneity is represented by road segments via the road operational or functional levels, traffic volume, and degree of congestion (Xie & Levinson, 2007). So analysis of a road network should focus on edge properties. Furthermore, a road system has a predominantly static structure but also has dynamic traffic information. Hence, road network studies should consider the variances of measurements in different time periods.

The application of network theory on road infrastructure structure will further facilitate accessibility assessment and benefit business operations from the merits of simplicity and cost-efficiency during retail store location decisions. This chapter aims to establish links between

network metrics and traffic congestion measures to answer the research question: to what extent do road network properties correlate with traffic congestion? We will describe methodology and results that provide insights into the relationship between network metrics and simulated traffic congestion.

2.2. Network Analysis Methodologies

For decades, network analysis has been widely adopted in the fields of social studies, biology, and information sciences among others (Wasserman & Faust, 1994; Robins, 2015); however, its implementation in research about road infrastructure is limited. If assuming the drivers are rational and have the ability to select the shortest path, some observation methods (e.g., network metrics) and analysis techniques (e.g., general linear model) could be adopted from other disciplines in road network analysis. Particularly, a road network can be assessed by either local measures (i.e., individual-based measures) or global measures (i.e., network structure measures).

Local measures reflect the role that an individual node or edge plays in a network. For example, centrality is a set of measurements that describe the importance of a network element. Specifically, betweenness and load centrality are centrality metrics that measure the vitality of a node or an edge via the probability of passing the element during a traversal search. The higher betweenness or load centrality value indicates the higher frequency the node or edge is on "shortest paths" in a network. Therefore, the removal of such element will potentially partition the network. Closeness centrality is another example of centrality that measures the importance of a network element. It values edges or nodes that are closer to other edges or nodes regarding the network distances. Other centrality measures exist, for example, degree centrality. Degree centrality is calculated as the node degree standardized by the frequency of a node. In a road network, nodes with a high degree centrality are recognized as transportation hubs, and nodes with degree of one are the road dead-ends.

Global measures represent the regional character of a road network. For example, average edge length is a primary feature that distinguishes a network. Like a highway system, it is characterised by short edges compared to other networks, such as the airline network and the continental internet (Gastner & Newman, 2006). Beyond length, road density reflects the

crowdedness of a network. The measure of road density may reflect and impact other social-economic factors, like gasoline consumption (Su, 2011). Network entropy is another example of global measure that indicates the assortativity of road heterogeneity (Xie & Levinson, 2007). And fractal dimension is a global network measure that reflects the formation and density distribution of a regional road network (Lu & Tang, 2004).

2.2.1. Centralities[8]

Network centrality measures have been implemented in traffic flow and congestion modelling (Holme, 2003; Kazerani & Winter, 2009; Gao, Wang, Gao, & Liu, 2013). However, such implementation has not yet been extended to the assessment of road network accessibility. Accessibility is a function of traffic flow and road network structure. Traditionally, accessibility has been measured by travel cost via distance, time or network connectivity (Litman, 2003). However, for a road network in a metropolitan area (e.g., City of Toronto), computing a cost matrix would incur substantial computational overhead due to the large size of the network. Instead, centrality measures facilitate an efficient, detailed and comprehensive accessibility assessment scheme. In a broader context, measuring network centralities is the preliminary approach to investigating the role of network analysis in traffic congestion research and retail site selection. The network centralities explored in this study include betweenness centrality, load centrality, closeness centrality, and degree centrality.

2.2.1.1. Betweenness Centrality

Betweenness centrality reveals the importance of a network element according to the number of the dominated paths (Figure 2-1). The betweenness of a node is calculated as the fraction of the passing shortest paths over the total number of the shortest paths in the network[9] (Equation 2-1). A node or edge with high betweenness centrality passes many shortest paths in the network. Therefore, its removal may partition (i.e., split) the network into multiple components.

[8] The algorithms used in this study were implemented by Hagberg in NetworkX load.py.
[9] Applying a shortest path algorithm on a network would produce a list of shortest paths that connect each pair of nodes in the network; among these paths, some pass through the node/edge of interest, they are referred as the "passing shortest paths".

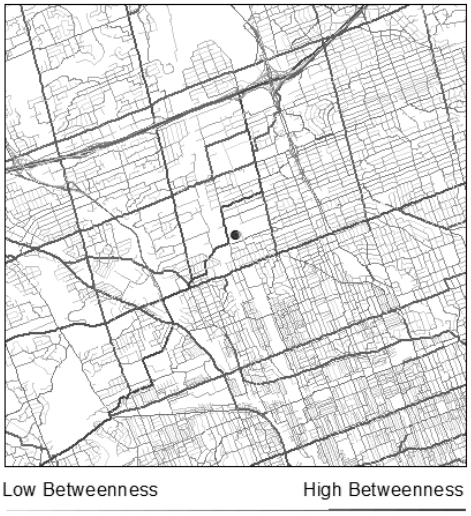


Figure 2-1 Betweenness centralities of roads, screenshot of road network in the City of Toronto.

According to the definition given by Everett and Borgatti (1999), the betweenness centrality of node *v* in network *G* is denoted as:

$$c_v^b = \sum_{\substack{s,t \in V \\ s \neq t}} \frac{\sigma(s,t|v)}{\sigma(s,t)} \qquad \text{Equation 2-1}$$

where *V* is the node collection of network G; *s, t* are two nodes in *V*; (*s, t*) is a path from *s* to *t*; node *v* is a node on the path (*s, t*); $\sigma(s,t)$ is the number of shortest paths[10] between source node *s* and target node *t*; among these shortest paths, $\sigma(s,t|v)$ is the number of paths that pass through node *v*.

Brandes (2008) developed an algorithm of betweenness centrality calculation by a recursively accumulation of "dependency". A "dependency" variable is defined as:

$$\delta(s,t|e) = \frac{\sigma(s,t|e)}{\sigma(s,t)} \qquad \text{Equation 2-2}$$

where *e* is an edge on the path (*s, t*). After summation for all target nodes, the "one-sided dependency" from a source *s* to a single edge *e* is defined as:

$$\delta(s|e) = \sum_{t \in V} \delta(s,t|e) \qquad \text{Equation 2-3}$$

[10] The shortest paths were identified using Dijkstra's algorithm.

This expression could be exploited as:

$$\delta(s|e) = \sum_{\substack{w:\ for\ e(v,w)\in E,\\ v\ is\ the\ predecessor\ of\ w}} \frac{\sigma(s,v)}{\sigma(s,w)}(1+\delta(s,w)) \qquad \text{Equation 2-4}$$

where node *v* is the predecessor of node *w* in a single source shortest path from source node *s*. The initial value of $\delta(v)$ for all nodes is 0. Therefore, in each iteration the furthest edge has $\delta(e)$ of 0; and the nearest edge gets the largest value of $\delta(e)$. Then in the reverse discovery order in the shortest path algorithm, repeat this calculation on all nodes as source and accumulate $\delta(s|e)$ for each edge. The final result is the betweenness centralities for each edge.

2.2.1.2. Load Centrality

Load centrality is a variation of betweenness centrality and shares similar definition and calculation (Figure2-2). Load centrality is defined as the ratio of the shortest paths that pass through a node in a network, indicating the influence of a node over the network via shortest paths. Since it was coined by Freeman in 1977, load centrality has been implemented by various algorithms. Particularly, Goh et al. (2001) proposed an efficient algorithm that works on large graphs. This algorithm was refined by Newman in his paper later in the same year (thence load centrality is also referred as Newman betweenness centrality in some literatures).

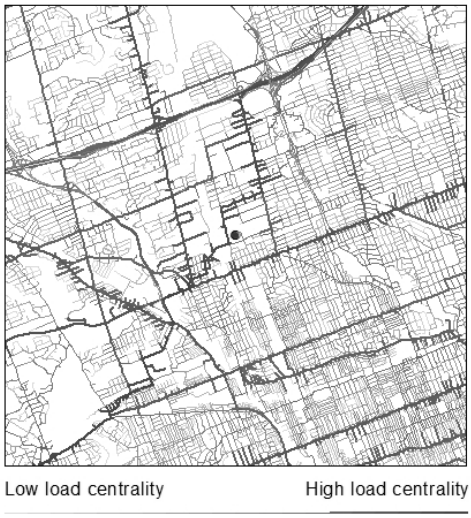


Figure 2-2 Load centralities of roads, screenshot of road network in the City of Toronto.

The calculation of load centrality shares the same formation with that of betweenness centrality. In network *G*, the load centrality of node *v* is denoted as:

$$c_v^l = \sum_{\substack{s,t \in V \\ s \neq t \\ s,t \neq i}} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$ Equation 2-5

where *V* is the node collection of network G; *s, t* are two nodes in *V*; (*s, t*) is a path from *s* to *t*; node *v* is a node on the path (*s, t*); $\sigma(s,t)$ is the number of shortest paths between source node *s* and target node *t*; among these shortest paths, $\sigma(s,t|v)$ is the number of paths that pass through node *v*.

Different from the calculation of betweenness centrality, the accumulation term is called "load"; it was defined by Newman (2001) as follow:

$$\varphi(s|v) = \sum_{\substack{w:\ for\ e(v,w)\in E, \\ v\ is\ the\ predecessor\ of\ w}} \frac{\varphi(s|w)}{\sigma(s|w)}$$ Equation 2-6

where node *v* is the predecessor of node *w* in a single source shortest path from source node *s*. Then load $\varphi(s|v)$ is accumulated for each node in the spinning tree from the farthest end to the source *s*. For calculation purpose, $\varphi(v)$ for each *v* is assigned an initial value of 1, and after the accumulation in each single source iteration, 1 is deducted from $\varphi(s|v)$ for each node. Therefore, both the leaves and the root of one spinning tree have load values of 0.

To compare with other edge-based network metrics, edge load centrality was calculated as the average the load centrality values of the two endpoints:

$$c^l_{\substack{(v,w) \\ (vw)\in E}} = \frac{1}{2}(c_v^l + c_w^l)$$ Equation 2-7

2.2.1.3. Closeness Centrality

Closeness centrality indicates the influence of a node based on the geodetic distance. It is measured by the mean distance from a node to the rest of the nodes in a graph (Figure 2-3).

Newman (2001) defined the closeness centrality of a node *u* as the reciprocal of the sum of the distances from node *u* to all other *n-1* nodes:

$$c_u^c = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$ Equation 2-8

where *n* is the number of nodes in the graph, *v* is a node in the node collection, and *d(v, u)* is the distance of the shortest path between node *v* and node *u*. A large closeness value means the node is comparably closer to other nodes and thus is considered to be more centralized. The edge closeness centrality was calculated as the average of the closeness values of the two endpoints:

$$c^c_{(v,w)} = \frac{1}{2}(c^c_v + c^c_w) \qquad \text{Equation 2-9}$$

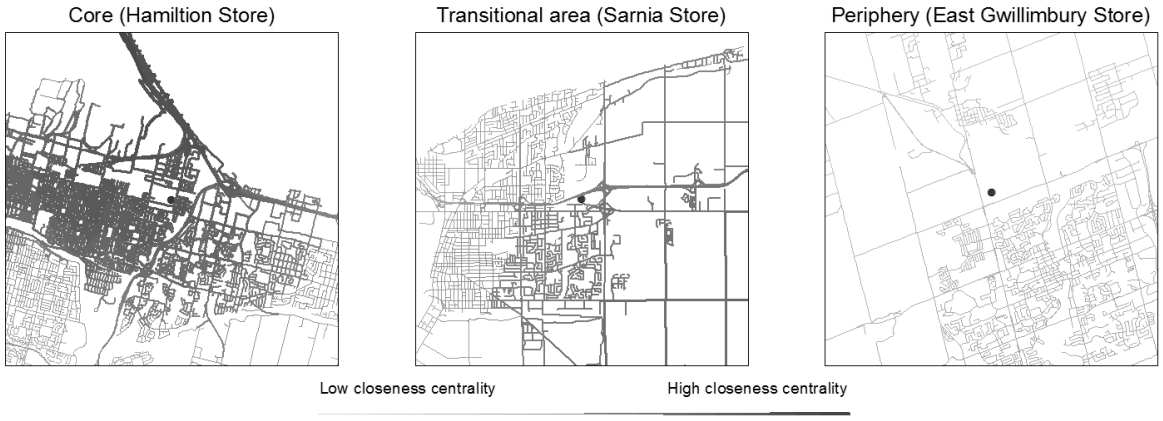


Figure 2-3 Closeness centralities of roads, screenshot of road network in the City of Toronto.

2.2.1.4. Degree Centrality

Degree centrality is the simplest centrality measure which is calculated as the number of edges that connects to a node (Figure 2-4). In this study, degree centrality is normalised by *(n-1)*, which is the maximum possible degree of a node (Newman, 2010):

$$c^n_i = \frac{d_i}{n-1} \qquad \text{Equation 2-10}$$

where d_i is the degree of node *i*, and *n* is the total number of nodes in the network. The node centrality in a road network indicates the complexity of an intersection. A high degree centrality value is detected at an intersection that connects many roads, where traffic flow comes from multiple possible directions and potential congestions may occur. Again, node degree centrality can be transformed to edges to compare with edge-based metrics.

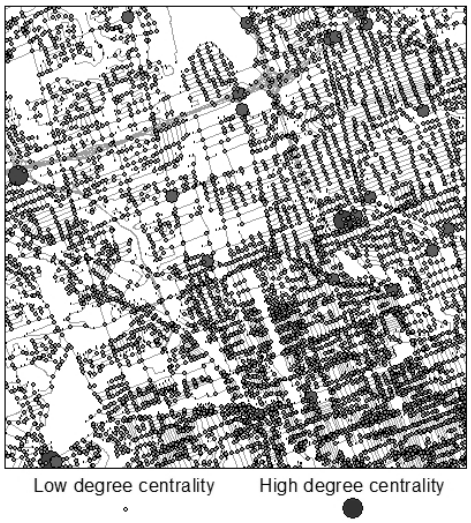


Figure 2-4 Degree centralities of roads, screenshot of road network in the City of Toronto.

2.2.1.5. Exceptions and Errors

The real-world road network can be complex and centrality calculation in a road network may contain exceptions or errors. For example, value 0 may be assigned to betweenness which infer that the edge (i.e., road segment) does not pass any traffic in a network. Betweenness centrality is accumulated during a Breadth First Search[11] (BFS), so that non-tree-edges (back, forward, and cross edges), form circles in a road network and yield a betweenness value of 0 (Figure 2-5). However, practically there is hardly a road segment that has no traffic flow. Therefore, such value was excluded from the results before further analysis in this presented study.

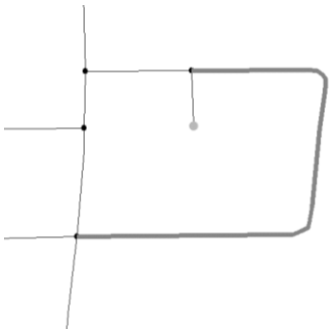


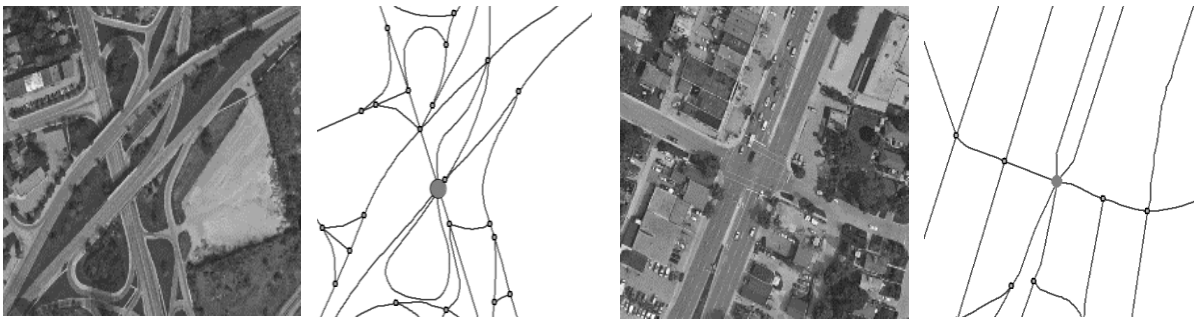
Figure 2-5 Instances of centrality calculation exception. In the screenshot of City of Toronto road network, edge betweenness centrality of 0 on a circle is highlighted in bold line.

[11] Implemented in Dijkstra's algorithm in this presented research.

Errors are often detected at overpasses or intersections in road network generation and they can lead to erroneous degree centrality values. For example, many of the multi-arm intersections detected at overpasses turn out to be errors of planarization. The overpass in Figure 2-6 (a) involves roads from different levels; however, it was recognised as one intersection during network generation. In this case, the measured degree is the sum of the degrees of the two intersections from different levels.

Specification is another type of error that also induces false high degree centrality. For instance, according to the satellite image, the six-arm intersection in Figure 2-6 (b) is a two-way street with barriers intersecting with another two-way street without a barrier. The error of specification misleads to increased number of possible traffic directions. Theoretically, there should be $P_6^2 = 30$ possible directions for traffic flows at a two-way six-arm intersection. However, at this shown intersection, traffic is forbidden for at least ten directions (with U-turn restriction). Therefore, it is more reasonable to consider this intersection as a four-way junction with a degree of four.

The error of overestimation of node degree was manually corrected using a geographic information system (i.e., ArcMap). In the planarization problem, roads were merged to eliminate intersections that were induced by errors; in specification problem, extra intersection was created for the divided roads.



(a) *(b)*

Figure 2-6 Instances of error: overestimation of degree centralities

Note: (a) Planarization: The node with a degree of seven is a combination of two intersections at two levels with degree of three and four. (b) Specification: The road from north-east to south-west is divided by barriers, and traffic is not permitted to go into the oncoming direction; however,

as the road is specified as two separate roads, the intersection can be mistaken as a six-arm junction. Satellite image source: Google Maps

2.2.2. Global network metrics

2.2.2.1. Network Entropy

Entropy is a global measure that represents the degree of uncertainty in a system (i.e., randomness and noise; Shannon, 1948). A system *X* contains objects from different categories, which are indexed by *i*. The proportion of the objects of interest in category *i* is represented as p_i, whose range is [0, 1]. Then the entropy of the system is expressed as:

$$H(X) = -\sum_{i=1}^{m} p_i log_2(p_i) \qquad \text{Equation 2-11}$$

The value of entropy ranges from 0 to $log_2 m$. A low entropy value indicates system homogeneity, it is observed in a system when a large proportion of the objects of interest fall within a single category and other categories cumulatively have a small proportion of the objects of interest in the system. Specifically, if a category covers nearly all or none of the objects of interest, the term $p_i log_2(p_i)$ approaches value 0. And thus the aggregation of the objects of interest in one category would yield a summation value closes to 0. In contrast, a higher value of entropy implies a higher degree of heterogeneity, which can be observed when p_i is close to 1/m in all categories.

The concept of entropy is not new to network analysis. Balch (2000) has studied the diversity of an artificial society by measuring system entropy, and Xie and Levinson (2007) calculated the entropies of modelled road networks to compare their heterogeneity. In this presented research, entropy was calculated based on road length and road functional classifications. Thereby, the value of entropy would indicate the degree of diversity of road types in a road network (Figure 2-7). According to the designed road functionalities, a road network with low entropy that contains a significant portion of local streets has the advantage of accessibility to property parcels. And a network with low entropy that is dominated by highways or arterial roads has better traffic movement with high speed limits and large road capacities. In contrast, the road networks with high entropy values have an assortment of road types and are more complex and balanced between property accessibility and traffic movement.

Figure 2-7 Entropy and road functional types.

2.2.2.2. Fractal Analysis

The patterns of many natural objects and human activities are visually complex; however, the disorder and irregularity underlying these real-world phenomena may show a scale-invariant repeating pattern. "Fractal dimension" was developed to describe this pattern using the concept of "dimension" in Euclidean space. The theory of fractal originated in 1970's, when Mandelbrot (1967) first coined the term "fractal" to describe the shape of British coastline. It has attracted the attention of geographers, especially in the field of city and city system studies (Batty & Longley, 1987). After Thomson (1977) first conducted fractal measurements on transportation systems and related them to urban form, the fractal measurements of the transportation system have proved to be associated with other urban subsystems in economic, institutional, and social processes system (Thomson, 1977; Benguigui & Daoud, 1991; Lu & Tang, 2004).

In Euclidean space, geometry measures follow a proportional relationship:

$$L^{\frac{1}{1}} \propto S^{\frac{1}{2}} \propto V^{\frac{1}{3}} \qquad \text{Equation 2-12}$$

where *L* stands for length, *S* is the surface area, and *V* is volume, representing one, two, and three-dimensional measures, respectively. This relationship can be extended to spaces of any dimensions:

$$L^{\frac{1}{1}} \propto M^{\frac{1}{d}}$$ Equation 2-13

where *M* can be replaced by any of *L, S, V*, or any other spaces; and *d* stands for dimension. If *d* is a non-integer, then the object is called fractal and *d* is an instance of fractal dimension (*D*).

For a road network with an area of *S*, its total length *L(S)* has a fractal dimension of *D* and follows:

$$L(S)^{\frac{1}{D}} \propto S^{\frac{1}{2}}$$ Equation 2-14

Assuming the network has a circle sampling area, then $S \propto r^2$, where *r* is the radius of the circle. Substitute S by r^2, and Equation 2-14 can be simplified as:

$$L(r) \propto r^D$$ Equation 2-15

Or, further specify the length-radius fractal dimension *D* as D_L:

$$D_L \propto log_r L(r)$$ Equation 2-16

If we define road network density $\rho(r) \propto \frac{L(r)}{r^2}$ as the length of road per unit area, the density is proportional to $r^{D_{L-2}}$:

$$\rho(r) \propto r^{D_L-2}$$ Equation 2-17

In a road network, the fractal dimension reflects the distribution and development stage of the road networks (Figure 2-8). If $D_L < 2$, the density of the road network decreases from the core to the periphery, indicating the urban space allows further development; if $D_L = 2$, the road network is evenly distributed in the area, and the space is filled up densely by the transportation system; if $D_L > 2$, the density of road network increases from the core to the periphery, which implies the circle centre is not a centre of transportation system or human settlement.

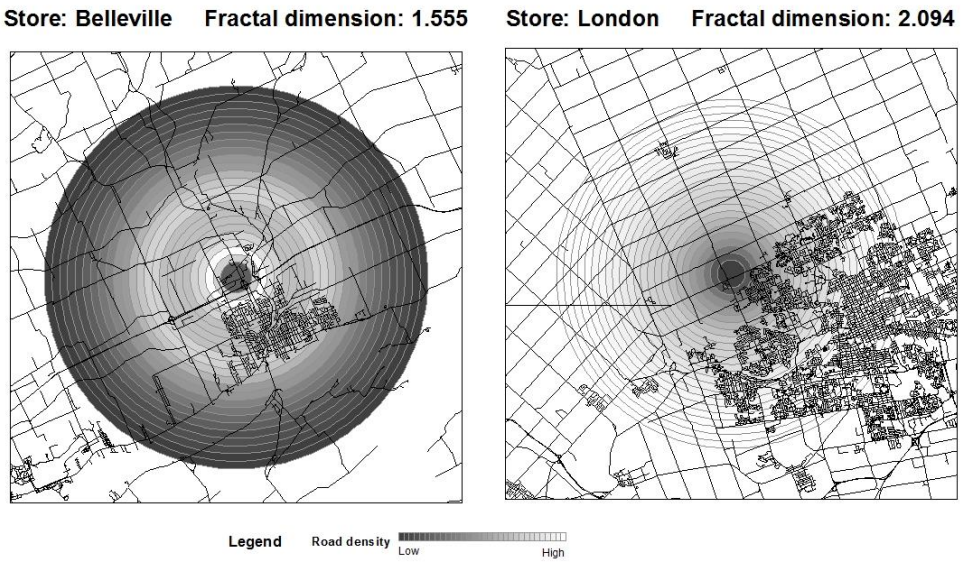


Figure 2-8 Fractal dimension and road network density.

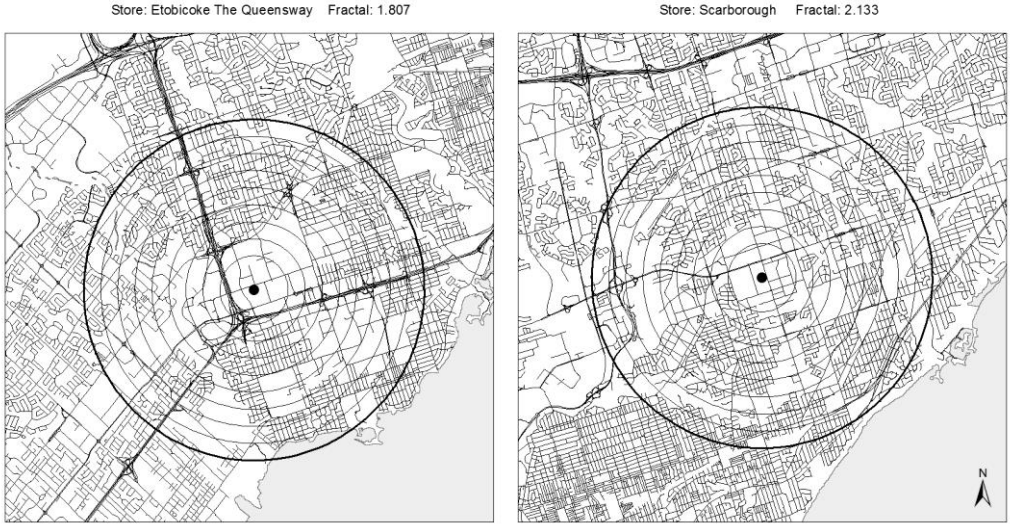


Figure 2-9 Examples of fractal dimension measurement

Note: In the Etobicoke road network (on the left), there are complex highways passing by the store (i.e., the centroid), and also some open spaces distributed across the sample area. Therefore, the road density decreases from the sample core as measured by a fractal dimension lower than 2. In contrast, the Scarborough road network (on the right) is more evenly distributed in general. There are two major roads at the periphery on the west and the east-west, and a dense road cluster at the south of the periphery. Thence, the fractal dimension is higher than 2.

In this presented research, the fractal dimension of a road network was measured based on the total length of the road network and the radius of the sample area via Ordinary Least Squares regression. The road networks were clipped by a set of concentric circles around each store in ArcMap (Figure 2-9). The radius of the smallest circle was 800 metres, and the radius increment by 400 metres, until the outermost circle reached a radius of 10,000 metres. Ideally, each store should be sampled by a set of 24 concentric circles. However, some stores are located along the waterfront, and part of the corresponding sampling area covers open water where the road network does not exist. Therefore, the number of sampled concentric circles were shrunk for such stores' road networks to exclude water body from samples.

2.3. Case Study: Ontario Road Networks and Traffic

Traffic congestion potentially has negatively effects on a store's accessibility and attractiveness. Whereas, the high (and slow) traffic volume that causes congestion may also increase store visibility to the drivers and passengers. Moreover, traffic congestion is often related to high market density (Wheaton, 1998), which can benefit a retail store due to agglomeration and spillover effects. Therefore, the impact of traffic congestion on retail revenue is controversial. To reveal the impacts of road network structure on retail performance, we first investigated the relationship between the road network and the traffic congestion in a case study situated in the census division of City of Toronto, Ontario, Canada.

2.3.1. Study area

City of Toronto is the provincial capital of Ontario, Canada, covering 636 km^2 (Figure 2-10). It is the most populated metropolitan region in Canada, with a population of 2,576,025 in 2011 (Statistics Canada, 2011). The large population coincides with a high demand for mobility. The road system of City of Toronto has a total length of 6,604 kilometres. It carries over one million people during daily commutes between home and work (Statistics Canada, 2011). Due partially to the volume of traffic, traffic congestion is a reality for many Toronto residents. In 2011, the average commuting time from home to work was over 30 minutes (City of Toronto, 2013).

City of Toronto road network derived from Ontario Road Network (ORN) 2010 published by Natural Resources Canada was used in this study. The ORN data comprises road geometry and

a set of attributes such as road length and road classification. During calculations, a 50-kilometre buffer was applied to the census division to eliminate edge effects, and the network measures were reported only within the study area's boundary.

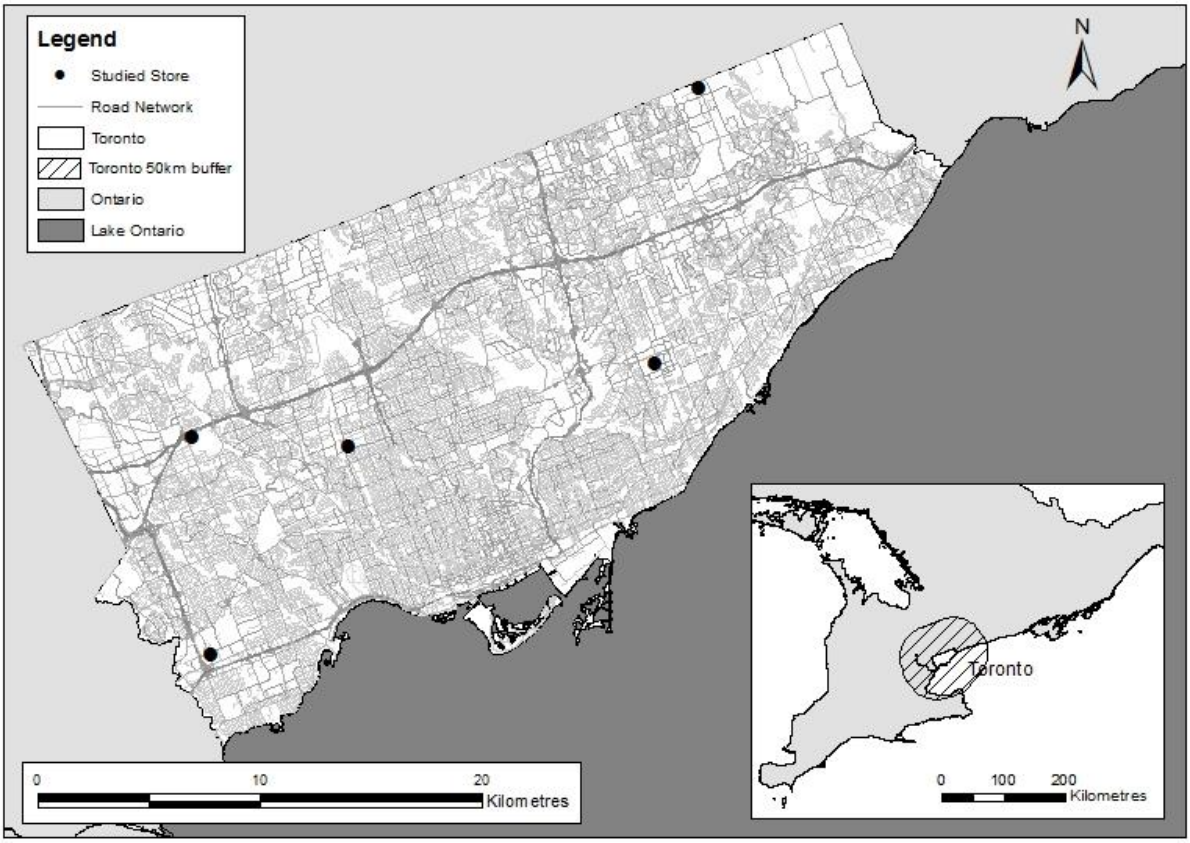


Figure 2-10 City of Toronto census division in Ontario

2.3.2. Road Traffic Congestion Analysis

Traffic congestion is a state of a traffic condition with the occurrence of an impeded traffic flow, reduced travel speed, and delayed travel time. Aside from road-side exceptions, such as incidents or constructions, traffic congestion is most commonly seen at intersections or at the connections of different types of roads (Xie & Levinson, 2007). Also, reduced travel speed is often observed on the "major" and "busy" roads that undertake heavy traffic flow. Moreover, areas with dense human settlements or business are prone to large traffic volume. Such intersections, road segments, or areas that are vulnerable to traffic flow can be identified via road network analysis

(Dunphy & Fisher, 1996; De Montis, Barthélemy, Chessa, & Vespignani, 2007; Xie & Levinson, 2007; Kazerani & Winter, 2009; Gao, Wang, Gao, & Liu, 2013).

The occurrence of congestion can be defined by the relationships among three basic macroscopic elements of a traffic state: velocity (speed), flux (flow), and density (May, 1990). The relationship between speed, flow, and density has been formalized in many traffic stream models (Underwood, 1961; Greenberg, 1959; Pipes, 1967; Ceder, 1976; van Aerde & Rakha, 1995). Among others Greenshield's macroscopic stream model provides a simple and straightforward explanation of the interrelationship of the three factors (Figure 2-11).

On each road, free-flow speed is the maximum achievable driving speed on a road segment, and once the traffic density exceeds the free flow density, the actual driving speed decreases with increase in traffic density (Figure 2-11 (a)). Also, there is a threshold (often known as the critical point) where traffic flow reaches the maximum achievable flow (known as capacity) and congestion occurs. In an urban road system, traffic density is usually controlled by traffic signals to keep traffic flow below the capacity. Therefore, practically the increase of traffic flow corresponding to the decrease of speed (Figure 2-11 (b)) and the increase of density (Figure 2-11 (c)).

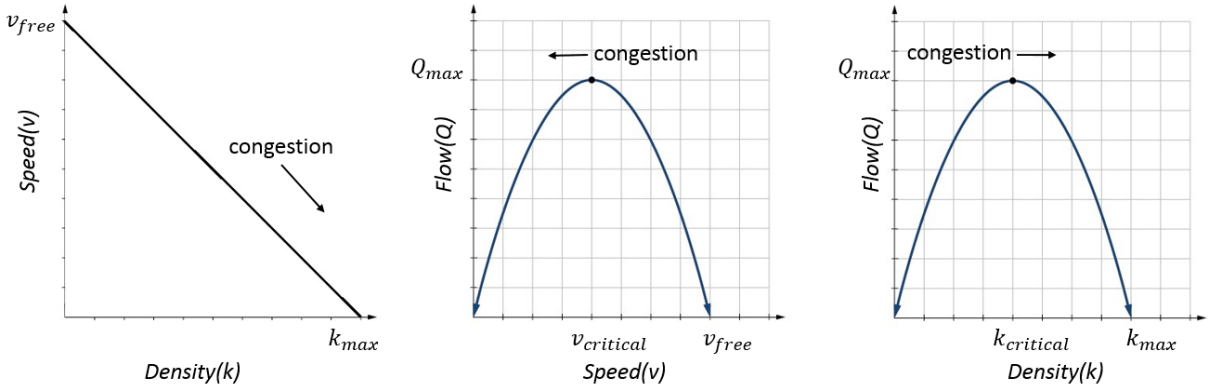


Figure 2-11 Greenshield's macroscopic stream model

Notably, the relationships among traffic flow, speed and density are not monotonic. Before reaching the critical point, flow-speed relationship is negtive and flow-density relationship is positive; once the critical point is passed, the relationships reverse and eventually reaches a

status of congestion where the flow is very low but density is high. Therefore, low speed or high density may indicate a traffic congestion more directly, since a low flow can be observed either on roads with little traffic (low density and high speed) or overwhelmed traffic (high density and low speed).

Due to the proprietary nature of traffic information, data provided by government, companies, and other organizations are void of congestion information (e.g., Google Traffic data, Ontario Ministry of Transportation Annual Average Daytime Traffic). Therefore, traffic simulation data were solicited from the Travel Modelling Group (TMG) from the University of Toronto (Travel Modelling Group, 2015). The simulation uses road-segment-based traffic data such as speed, flow, and capacity. Three congestion indicators (speed difference, volume overflow, and road occupancy time) were calculated based on traffic simulation.

Speed difference was calculated as the difference between free-flow traffic speed and the simulated speed:

$$\delta_v = v_{free} - v_{simulated} \qquad \text{Equation 2-18}$$

A larger speed difference value implies a higher possibility of traffic congestion.

Volume overflow was represented by the ratio between volume and capacity (VCR):

$$VCR = \frac{Volume}{Capacity} \qquad \text{Equation 2-19}$$

VCR ranges from 0 to 1. A VCR value closer to one indicates that the traffic volume is approaching to the capacity and a congestion may occur.

Lastly, occupancy ratio[12] was calculated as the ratio between traffic flow and speed:

[12] Occupancy ratio shares the same formula with traffic density. The occupied time for a unit length road segment is the product of volume and the travel time for each vehicle to pass the road segment. If given a time slot T, the traffic volume (i.e., the number of vehicles passing a point on the road) within T is $\mathrm{T} \times \mathrm{flow}$. The travel time for each vehicle to pass a point on the

$$Occupancy = \frac{volume}{speed} \quad \text{Equation 2-20}$$

A high value of occupancy indicates a high possibility of congestion and vice versa.

The road-segment-based traffic-simulation data were acquired for four time periods (Table 2-1). They are morning (06:00 to 09:00, denoted as "AM"), mid-day (09:00 to 15:00, denoted as "MD"), afternoon (15:00 to 19:00, denoted as "PM"), and evening (19:00 to 24:00, denoted as "EV"). Morning (AM) and afternoon (PM) contain peak traffic hours (Transportation Services, 2014) and there are high possibilities for congestion according to the statistics on congestion measures (Table 2-1).

Table 2-1 Averages of congestion measures in four time slots

Time slot	Congestion		
	δv	VCR	Occupancy
AM	3.87	0.45	9.57
MD	0.36	0.16	2.53
PM	2.86	0.39	7.64
EV	0.06	0.1	1.63

2.3.3. Measuring the relationship between road network metrics and traffic congestion

2.3.3.1. Network centralities and traffic congestion

The segment-based centrality metrics[13] and congestion measures in the four time periods were compared in a simple Ordinary Least Squares Linear Regression model, where network centrality measurements were independent variables and congestion measures were dependent variables.

road is $\frac{\text{vehicle length}}{\text{speed}}$. Therefore, the occupied time is $\text{volume} \times \frac{\text{vehicle length}}{\text{speed}}$. We assume that vehicle length is identical then the occupancy can be simplified as Equation 2-20.

[13] The measure of average road length is also a local network measure. However, the road network is segmented with random lengths despite of location. An average of segments length cannot capture the spatial heterogeneity of a road system, and therefore was not used in this presented study.

Moreover, indicator variables were included to improve the exploratory power of a single network metric of congestion. Variables included were noted in existing literature as having an effect on simulated congestion, which were factors like road class (McNally & Ryan, 1992), road capacity (Wilson, 1983), free-flow speed (Hadiuzzaman, Qiu, & Lu, 2012), and volume-delay function (i.e., VDF; Engelson & van Amelsfort, 2011). A model of the following form was used:

$$V_c = \alpha + \beta_n \times N + \gamma_1 I_1 + \cdots + \gamma_i I_i + \cdots + \gamma_n I_n \qquad \text{Equation 2-21}$$

where V_c is the congestion variable, α is the intercept, β_n is the coefficient for the network metric N, γ_i is the coefficient for the indicator variable I_i and n is the number of categories in the indicator variable.

The goodness of fit was tabulated by indicator variable. And network metrics were assessed for their ability to affect congestion according to adjusted R-squared and significance tests.

2.3.3.2. Global network metrics and traffic congestion

While both of the local measures (e.g., centralities) and congestion measures are segment-based, entropy and fractal dimension are global measures that cannot be directly compared to segment-based variables. Hence, to assess the effects of entropy or fractal on traffic congestion, the segment-based congestion measures were scaled up by taking global mean values.

Specifically, congestion was summarized at the following four spatial scales: adjacent roads, network community, 5-kilometer neighbourhood, and 19 minutes network-drive-time service area (Figure 2-12). Moreover, to compare with fractal dimension, congestion measures were summarized at the sampling area (i.e., the largest concentric circle that was used in fractal calculation). A high regional mean value of a congestion measure implies that the regional road network is more likely to incur traffic congestion relative to other regions with a lower mean value.

Global network metrics and congestion measures were compared based on observations near the five targeted retail stores in the City of Toronto. With the limited sample size, a

correlation analysis was performed to determine if there is a relationship between global network metrics and congestion measures according to the Pearson correlation coefficient.

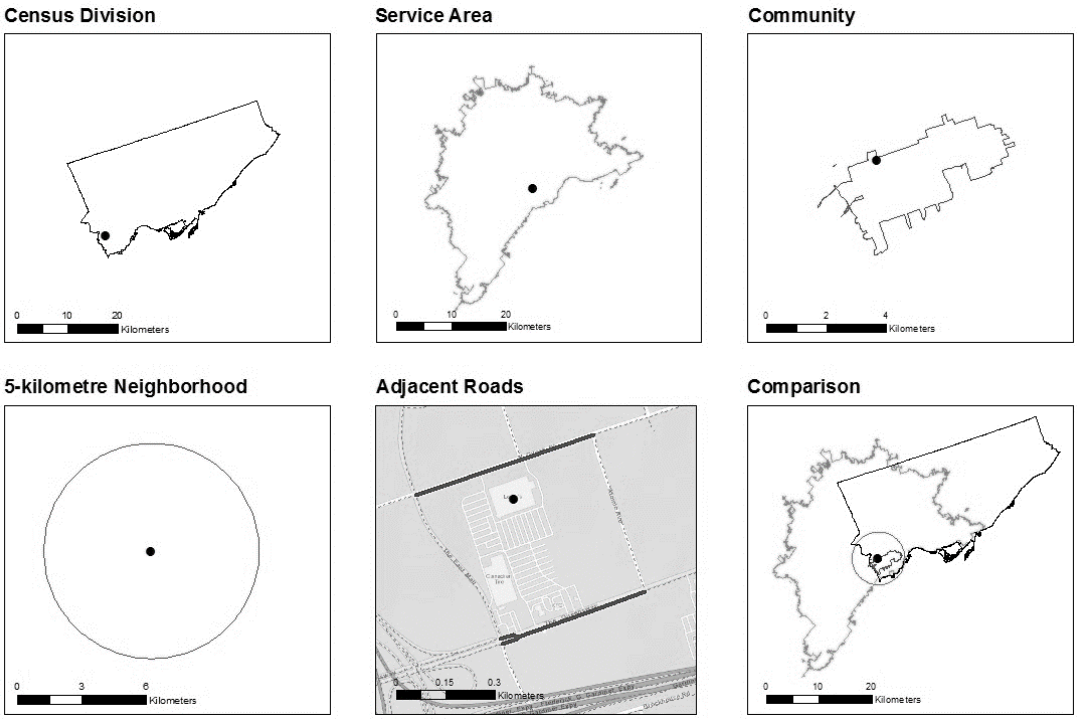


Figure 2-12 City of Toronto census division area and spatial scales for network analysis

2.3.4. Results

2.3.4.1. Relation of network centralities to congestion

Ordinary Least Squared Regressions were used to determine if a relationship exists between network centrality metrics (i.e., betweenness, load, closeness, degree maximum, degree standard deviation, and degree mean) and congestion measures (i.e., speed difference, VCR, and occupancy) during the four different periods of the day (morning, mid-day, afternoon, and evening) using 26 samples. The best predictor and indicator variables for each congestion metric were determined according to the p-values of the predictor and adjusted R-squares of the model.

a. *Speed difference (δv)*

All of the edge centrality metrics were significant at $p < 0.01$ when regressed against speed difference (δv), with each indicator variable included, over all time periods (Table 2-2). In contrast,

degree centrality statistics do not always hold a significant relationship with congestion measures[14].

The goodness of fit, as measured by adjusted R-squared, were low overall between speed difference and network metrics – with models included significant predictors at 0.01 level yielding a range of adjusted R-squared of 0.13 – 0.24 in the morning, 0.05 – 0.13 in the mid-day, 0.12 – 0.22 in the afternoon, and 0.01 to 0.03 in the evening. The closeness metric offered the highest goodness of fit in all time periods. Free-flow speed outperformed other indicator variables in terms of the goodness of fit, implying that partitioning the road network by free-flow speed would improve speed difference estimation via centralities.

Table 2-2 OSL regression of speed difference against network measures (Adjusted R-squares and significance)
AM: morning; MD: mid-day; PM: afternoon; EV: evening
**** P < 0.01; ** P < 0.05; * P < 0.1*

Time Slot	Centrality	Categorical Variables			
		Road Class	Road Capacity	Free-flow Speed	VDF
AM	Betweenness	0.16 ***	0.18 ***	0.19 ***	0.19 ***
	Closeness	0.20 ***	0.22 ***	0.24 ***	0.22 ***
	Load	0.16 ***	0.18 ***	0.19 ***	0.19 ***
	Degree max.	0.13 *	0.15 ***	0.16 **	0.16 ***
	Degree std.	0.13 ***	0.15 ***	0.16 ***	0.16 ***
	Degree mean	0.13	0.15	0.16	0.16
MD	Betweenness	0.05 ***	0.06 ***	0.08 ***	0.13 ***
	Closeness	0.06 ***	0.06 ***	0.08 ***	0.13 ***
	Load	0.05 ***	0.06 ***	0.08 ***	0.13 ***
	Degree max.	0.05 *	0.05 ***	0.07 ***	0.12 ***
	Degree std.	0.05 *	0.05 **	0.07 **	0.12 *
	Degree mean	0.05	0.05 **	0.07 *	0.12 **
PM	Betweenness	0.15 ***	0.17 ***	0.18 ***	0.19 ***
	Closeness	0.17 ***	0.19 ***	0.22 ***	0.21 ***
	Load	0.15 ***	0.17 ***	0.18 ***	0.19 ***
	Degree max.	0.12 **	0.14 ***	0.15 ***	0.16 ***
	Degree std.	0.12 ***	0.14 ***	0.15 ***	0.16 ***
	Degree mean	0.12	0.14	0.15	0.16
EV	Betweenness	0.01 ***	0.01 ***	0.01 ***	0.02 ***
	Closeness	0.03 ***	0.02 ***	0.03 ***	0.03 ***
	Load	0.01 ***	0.01 ***	0.01 ***	0.02 ***
	Degree max.	0.01	0.01	0.01	0.01
	Degree std.	0.01	0.01 *	0.01	0.01
	Degree mean	0.01	0.01	0.01	0.01

[14] Specifically, none of the degree centralities was significant at P<0.05 in the evening; degree mean was not significant at p<0.01 with any indicator in any time slot; the significance of degree maximum and standard deviation varies with each indicator among the time slots.

Based on the goodness of fit and significance test ($p < 0.01$), the best model for predicting congestion via speed difference uses the closeness network metric and free-flow speed as an indicator variable in the morning. The equation from these metrics has the following form:

$$\delta_v = -11.986 + 644.7\,Closeness + 0.0\,ffs_{40} + 2.594\,ffs_{50} + 5.413\,ffs_{60} + 8.689\,ffs_{70} + 6.334\,ffs_{80} + 9.15\,ffs_{90} + 14.770\,ffs_{100} + 26.474\,ffs_{110} \quad \text{Equation 2-22}$$

where δ_v is the speed difference; Closeness is closeness centrality; and ffs_x is the free flow speed with a value of *x*.

b. *VCR*

In regression analysis of centrality metrics against VCR, all predictors except for the mean degree centrality were significant at 0.01 level with all indicator variables in all time slots (Table 2-3). The ranges of the goodness of fit were 0.29 – 0.43 in the morning, 0.27 – 0.40 in the mid-day, 0.29 – 0.42 in the afternoon, and 0.26 – 0.38 in the evening. The results show that there were stronger relationships between VCR and centrality metrics compared to those between speed difference and centrality metrics. VCR was best estimated by closeness across all time slots. While VDF provided the best indicator variable by generating the highest R^2, road capacity and free-flow speed yielded comparable results, but road class had the poorest performance. These results suggest segmenting the road network by VDF improves VCR prediction using centrality metrics.

Based on the goodness of fit and significance test ($p < 0.01$), the best models for predicting congestion via VCR uses the closeness network metric and VDF as indicator variable based on morning traffic simulation. The equation from these metrics has the following form:

$$VCR = -1.356 + 47.812\,Closeness + 0.0\,VDF_0 + 3.390\,VDF_{11} + 0.841\,VDF_{13} + 2.448\,VDF_{14} + 1.270\,VDF_{15} + 1.134\,VDF_{16} + 0.472\,VDF_{17} + 0.857\,VDF_{20} + 1.594\,VDF_{21} + 0.847\,VDF_{22} + 1.270\,VDF_{30} + 1.079\,VDF_{40} + 0.519\,VDF_{41} + 0.746\,VDF_{42} + 0.804\,VDF_{43} + 0.813\,VDF_{50} + 0.629\,VDF_{51} \quad \text{Equation 2-23}$$

where VCR stands for Volume-Capacity Ratio; Closeness is closeness centrality; and VDF_x is the VDF with index of *x*.

Table 2-3 OSL regression of VCR against network measures
(Adjusted R-squares and significance)
AM: morning; MD: mid-day; PM: afternoon; EV: evening
**** P < 0.01; ** P < 0.05; * P < 0.1*

Time Slot	Centrality	Categorical Variables			
		Road Class	Road Capacity	Free-flow Speed	VDF
AM	Betweenness	0.33 ***	0.37 ***	0.34 ***	0.40 ***
	Closeness	0.38 ***	0.41 ***	0.41 ***	0.43 ***
	Load	0.32 ***	0.37 ***	0.34 ***	0.39 ***
	Degree max.	0.29 ***	0.33 ***	0.29 ***	0.36 ***
	Degree std.	0.29 ***	0.33 ***	0.29 ***	0.36 ***
	Degree mean	0.29	0.33 ***	0.29 **	0.36 ***
MD	Betweenness	0.29 ***	0.34 ***	0.34 ***	0.39 ***
	Closeness	0.31 ***	0.36 ***	0.37 ***	0.40 ***
	Load	0.29 ***	0.34 ***	0.34 ***	0.39 ***
	Degree max.	0.27 ***	0.33 ***	0.30 ***	0.37 ***
	Degree std.	0.27 ***	0.33 ***	0.30 ***	0.37 ***
	Degree mean	0.27 **	0.33 ***	0.30 **	0.37 ***
PM	Betweenness	0.33 ***	0.37 ***	0.35 ***	0.40 ***
	Closeness	0.37 ***	0.41 ***	0.41 ***	0.42 ***
	Load	0.32 ***	0.37 ***	0.35 ***	0.40 ***
	Degree max.	0.29 ***	0.33 ***	0.30 ***	0.37 ***
	Degree std.	0.29 ***	0.33 ***	0.30 ***	0.37 ***
	Degree mean	0.29	0.33 ***	0.30 **	0.37 ***
EV	Betweenness	0.28 ***	0.31 ***	0.29 ***	0.36 ***
	Closeness	0.31 ***	0.34 ***	0.34 ***	0.38 ***
	Load	0.28 ***	0.31 ***	0.29 ***	0.36 ***
	Degree max.	0.27 ***	0.29 ***	0.26 ***	0.35 ***
	Degree std.	0.27 ***	0.29 ***	0.26 ***	0.34 ***
	Degree mean	0.27 ***	0.29 ***	0.26 ***	0.34 ***

c. *Occupancy Ratio*

Occupancy ratio reflects the traffic crowdedness by occupied time of a road segment. When Occupancy was regressed against network centrality metrics in the four time slots, each edge centrality (betweenness, load, and closeness) was significant at 0.01 level, but degree centrality statistics were not always significant (Table 2-4). Specifically, degree maximum was not significant with road class in the morning and afternoon, degree standard deviation was significant with all indicators in all time slots, and degree mean was significant at 0.05 level in the mid-day and evening with three indicators except for road class.

The goodness of fit of Occupancy estimation has ranges of 0.23 - 0.34, 0.29 – 0.44, 0.25 – 0.37, and 0.31 – 0.45 during morning, mid-day, afternoon, and evening, respectively. The goodness of fit is comparable to that in VCR estimation. Again, closeness was the best estimator

for Occupancy in all time slots although other edge centralities produced very close goodness of fit; VDF outperformed other indicators and improved the goodness of fit the most.

Table 2-4 OSL regression of occupancy against network measures (Adjusted R-squares and significance)
AM: morning; MD: mid-day; PM: afternoon; EV: evening
**** P < 0.01; ** P < 0.05; * P < 0.1*

Time Slot	Centrality	Categorical Variables			
		Road Class	Road Capacity	Free-flow Speed	VDF
AM	Betweenness	0.27 ***	0.32 ***	0.31 ***	0.33 ***
	Closeness	0.29 ***	0.34 ***	0.34 ***	0.34 ***
	Load	0.26 ***	0.32 ***	0.31 ***	0.33 ***
	Degree max.	0.23	0.29 ***	0.27 ***	0.27 ***
	Degree std.	0.23 ***	0.29 ***	0.27 ***	0.27 ***
	Degree mean	0.23	0.29 *	0.27	0.27
MD	Betweenness	0.31 ***	0.38 ***	0.37 ***	0.43 ***
	Closeness	0.32 ***	0.40 ***	0.39 ***	0.44 ***
	Load	0.30 ***	0.38 ***	0.37 ***	0.43 ***
	Degree max.	0.29 ***	0.37 ***	0.35 ***	0.42 ***
	Degree std.	0.29 ***	0.37 ***	0.35 ***	0.42 ***
	Degree mean	0.29	0.37 ***	0.35 ***	0.42 ***
PM	Betweenness	0.28 ***	0.34 ***	0.33 ***	0.35 ***
	Closeness	0.30 ***	0.36 ***	0.36 ***	0.37 ***
	Load	0.28 ***	0.34 ***	0.33 ***	0.35 ***
	Degree max.	0.25 *	0.31 ***	0.29 ***	0.33 ***
	Degree std.	0.25 ***	0.31 ***	0.29 ***	0.33 ***
	Degree mean	0.25	0.31 **	0.29	0.33 **
EV	Betweenness	0.32 ***	0.39 ***	0.35 ***	0.44 ***
	Closeness	0.35 ***	0.41 ***	0.39 ***	0.45 ***
	Load	0.32 ***	0.39 ***	0.35 ***	0.44 ***
	Degree max.	0.31 ***	0.37 ***	0.33 ***	0.42 ***
	Degree std.	0.31 ***	0.37 ***	0.33 ***	0.42 ***
	Degree mean	0.31	0.37 ***	0.32 ***	0.42 ***

Therefore, the best models for predicting congestion via Occupancy uses the closeness network metric and VDF as an indicator variable on evening traffic data. The equation from these metrics has the following form:

$$Occupancy = -4.19 + 147.92\,Closeness + 0.0\,VDF_0 + 17.58\,VDF_{11} + 4.93\,VDF_{13} + 2.42\,VDF_{14} + 2.44\,VDF_{15} + 0.52\,VDF_{16} + 1.44\,VDF_{17} + 2.49\,VDF_{20} + 5.01\,VDF_{21} + 2.47\,VDF_{22} + 4.03\,VDF_{30} + 3.68\,VDF_{40} + 1.79\,VDF_{41} + 1.88\,VDF_{42} + 1.11\,VDF_{43} + 2.29\,VDF_{50} + 1.68\,VDF_{51} \quad \text{Equation 2-24}$$

where Occupancy is the Occupancy Ratio; Closeness is closeness centrality; and VDF_x is the VDF with index of *x*.

In summary, congestion estimations via VCR and occupancy had better performance than the estimation via speed difference. Closeness was the most suitable predictor regarding goodness of fit and significance and VDF was the best indicator variable overall. Traffic during peak hours (morning and afternoon slots) were best estimated via VCR by closeness and indicator variable with adjusted R-squared of 0.43 and 0.42. Traffic during off-peak hours (mid-day and evening) are the best estimated via occupancy by closeness with adjusted R-squared of 0.44 and 0.45.

Notably, network metrics that focused on edge attributes typically outperformed metrics that used node centrality statistics. The limitation on the exploratory power of node degree statistics in the regression analysis could be owing to the little variance in node degree in the road network.

2.3.4.2. Relation of global network metrics to congestion

The correlation analysis between entropy and congestion, or fractal and congestion, did not yield significant results (Table 2-5). The resulting p-values had a minimum value of 0.15 (between entropy at community level and speed difference measured in the evening) and an average of 0.72, suggesting that entropy or fractal dimension of road network are not significantly correlated to traffic congestion.

Table 2-5 Correlation tests between entropy, fractal, and congestion

Pearson correlations coefficient (r) and p-values (in *italic*)
AM: morning; MD: mid-day; PM: afternoon; EV: evening

Network Measure	Scale	Congestion Measure											
		δv				VCR				Occupancy			
		AM	MD	PM	EV	AM	MD	PM	EV	AM	MD	PM	EV
Entropy	Adjacent Roads	**0.13**	0.06	**0.18**	**0.39**	0.06	**0.41**	**0.22**	**0.31**	**0.23**	**0.44**	**0.24**	**0.33**
		0.84	*0.93*	*0.77*	*0.52*	*0.93*	*0.49*	*0.72*	*0.61*	*0.71*	*0.46*	*0.69*	*0.58*
	5km Neighborhood	-0.38	-0.16	-0.40	-0.60	0.07	**0.13**	0.08	-0.28	**0.13**	**0.27**	**0.26**	-0.11
		0.53	*0.79*	*0.51*	*0.28*	*0.91*	*0.83*	*0.90*	*0.64*	*0.83*	*0.66*	*0.68*	*0.86*
	Community	-0.29	-0.23	-0.27	-0.75	**0.14**	**0.13**	**0.11**	-0.12	0.00	**0.16**	0.07	-0.03
		0.64	*0.71*	*0.67*	*0.15*	*0.83*	*0.84*	*0.86*	*0.85*	*1.00*	*0.80*	*0.92*	*0.96*
	Service Area	**0.13**	-0.03	0.06	-0.15	**0.16**	**0.18**	**0.16**	-0.03	0.05	**0.12**	0.06	-0.05
		0.83	*0.96*	*0.92*	*0.81*	*0.79*	*0.78*	*0.80*	*0.97*	*0.94*	*0.85*	*0.92*	*0.94*
Fractal	Fractal Area	-0.10	**0.34**	-0.13	**0.31**	-0.54	-0.54	-0.58	-0.14	-0.28	-0.54	-0.47	-0.23
		0.88	*0.58*	*0.83*	*0.61*	*0.35*	*0.35*	*0.30*	*0.82*	*0.64*	*0.35*	*0.43*	*0.71*

Note: positive correlation (r>=0.1) in bold with shading; negative correlation (r<=-0.1) with shading; no correlation (-0.1<r<0.1) in hallow.

For correlations between entropy and congestion measures, the results contained high p-values and an inconsistency across spatial scales, time slots, or among congestion measures. Although, some correlations existed within certain spatial scales or time slots. For example, at adjacent road level, entropy was positively correlated (correlation coefficient > 0.1) with congestion for 10 out of the 12 tests. Or during mid-day, entropies was positively correlated with VCR and Occupancy at all scales. However, negtive correlations with speed difference were observed at the 5km neighborhood and community scales across all time slots. Therefore, entropy cannot be used as a strong predictor of congestion as measured by speed difference, VCR, or Occupancy.

The correlations between fractal dimension and congestion measures were more consistent than those attained for entropy (Table 2-5). The mean values of VCR and Occupancy were slightly negatively correlated with fractal dimension in all time slots, implying traffic congestion is more likely to occur at cores of road networks. For example, the road network around the Etobicoke store yielded the highest mean VCR value (1.242) and the smallest fractal value (1.807) among the tested road networks; while the road network around the Scarborough store had a higher fractal dimension (2.133) and less congestion (0.975) with a less dense core (Figure 2-9 in section 2.2.2.2., Chapter 2).

2.4. Discussion

In this research we identified the relationship between traffic congestion measures and network metrics in a case study of the City of Toronto road network. Positive correlations between congestion measures and road centrality metrics were identified. Specifically, during peak hours, congestion was best estimated via VCR with adjusted R-squared values of 0.43 (AM) and 0.42 (PM); during off-peak hours, congestion was best estimated via occupancy ratio with adjusted R-squared values of 0.44 (MD) and 0.45(EV). These results indicate that the important road segments which are identified by high centrality values have greater possibilities for the occurrence of traffic congestion. This finding suggests that centrality metrics are able to serve as indicators of traffic congestion. Moreover, we identified that closeness centrality was the best estimator of congestion among other centrality metrics. And the inclusion of VDF as an indicator variable in the regression with traffic simulation yielded the best quality of the results.

However, the results of correlation tests between global network metrics (i.e., entropy and fractal) and congestion measures were ambiguous. First, correlation results contained high p-values (with minimum of 0.15, average of 0.75), suggesting there was no or a weak relationship between global network metrics and congestion measures. Second, the Pearson correlation coefficients between global network metrics and congestion measures were not consistent across different spatial scales or time slots. Therefore, no concrete conclusion about the relationships between entropy and congestion, or fractal and congestion can be drawn from the results of this presented research. However, both global network metrics reflect the formation and spatial configuration of a road network, and their potential underlying correlations to congestion or other socio-economic factors should be expected in further studies.

2.4.1. Limitations

2.4.1.1. Observation techniques

The ambiguity in the correlation results between global network metrics and congestion is partially induced by the limited sample size (i.e., only five sample locations were tested by global network metrics in this research). Having small sample size would increase the probability of introducing type 2 errors into a correlation test. According to the Law of Large Numbers[15], increasing sample size in the future studies may provide better estimation of the distribution of exploratory and response variables, and therefore improve the detection of the correlation between global network metrics and congestion measures. However, because entropy are measured at certain fixed spatial scales (e.g., service area or community), and network metrics may loss integrity across two cities, increasing sample size would require extending the study area to other census division areas. In contrast, increasing sample size of road network fractal dimension measurement can be achieved by applying a moving window on each road segment for fractal dimension calculation. However, this method would result in a high computational overhead.

[15] Law of Large Numbers: the convergence of the results' average to the expected value for a random process as the number of trials increases.

Another possible reason for the lack of correlation between statistical network metrics and traffic congestion could be the generalization in the statistics of segment-based values. When calculating the correlation, road-segment-based congestion measures (i.e., speed difference, VCR and occupancy ratio) were summarized by averages at the corresponding spatial scales of entropy and fractal dimension measurements. Although the method of summarizing network data by mean value is simple and direct, the statistic contains bias which is induced from unequal road segment length and spatial heterogeneity. Instead, spatial statistical methods are recommended for the comparison between segment-based traffic data and global network metrics in further studies. For example, the Network Kernel Density Estimation approach (Xie & Yan, 2008) uses a Kernel function to create a probability model for each individual point on the road network. It provides a method for the integration of global network metrics by using an identical parameter in the model.

While traffic data were obtained from the Ministry of Transportation in the form of Average Annual Daily Traffic variables, the acquisition of real-world congestion data is both difficult and costly. To overcome this issue, simulation results are widely adopted (Mahmassani & Chang, 1986; Minderhoud & Bovy, 2001). The present research used the traffic simulation data provided by the Travel Modelling Group from the University of Toronto. Despite the inaccuracy and limitations of simulated data, we used traffic simulation data in this study as they capture the movement in a transportation system and provided comprehensive spatial and temporal coverage, which would be difficult to attain from a private company (e.g., Google Traffic). Real-world and real-time traffic data may be available in the near future from initiatives like the Vehicular Ad hoc Network (VANET). The VANET is a widely discussed system which collects comprehensive real-time traffic data and facilitates intelligent traffic management (Harri, Filali, Bonnet, & Fiore, 2006; Piorkowski, et al., 2008; Nzouonta, Rajgure, Wang, & Borcea, 2009); although this technique is still pending for mass application due to market penetration process (Sam & Raj, 2014) and authentication challenges (Studer, Shi, Bai, & Perring, 2009), high quality traffic data should be expected in the near future with the development of technology and market acceptance.

2.4.1.2. Study design

A network model uses nodes and edges to represent entities, processes, and their underlying structure in the real world (Newman, 2010). The interactions and reciprocations among entities or processes are inevitably included in a network model. Therefore, road network metrics, like centrality metrics, inherently contains dependent data and violate the assumption of independence in a standard statistical model (Robins, Pattison, Kalish, & Lusher, 2007; Valente, Coronges, Lakon, & Costenbader, 2008). Some statistical models have been proposed to investigate the reciprocation between network elements. For example, exponential random graph models (ERGMs) uses Markov chain Monte Carlo (MCMC) method to estimate the probability of the formation of edges in the observed network (Robins, 2011).

Road network in geographic information system (GIS) is an instance of geometric network that describes the physical structure of a road system (Okabe & Sugihara, 2012). So the entities in the network also follow the First Law of Geography: everything is related to everything else, but near things are more related than distant things (Tobler, 1970). In a road network, traffic flow is inclined to be affected by the nearby traffic conditions. Black (1992) pointed out the necessity of using spatial autocorrelation in network analysis, and proposed an approach of calculating network Moran's I. The incorporation of spatial autocorrelation in road network and transportation analysis is highly recommended for future research to explore the clustering, dispersion, or random pattern of road network structure and traffic congestion.

2.4.2. Contributions and future directions

The presented research established a link between network metrics and traffic congestion measures. The findings in this research enabled a cost-efficient congestion estimation method which requires a publicly accessible input (i.e., road infrastructure network). This method is especially suitable for large network analysis. It benefits researchers, commercial firms, and organizations whose access to detailed traffic data is limited. The direct implementation of network metrics is to facilitate traffic planning via congestion prediction for transportation engineers and city planners.

Road network is an indispensable component of an urban system, and there are mutual impacts between the transportation system and human settlements/economic activities. Beyond traffic estimation, the identified relationships between road network metrics and traffic congestion measures can also be implemented to explore economic phenomenon, for example, retail stores sales modelling.

2.5. Conclusion

Traffic system is an essential part of an urban environment. In the retail sector, the efficiency of a road network and the retail store's accessibility affect a store's attractiveness and hence are crucial to a store's performance. The analysis of the impacts of road network structure on traffic congestion is necessary in retail site-selection decision making process. However, there is little published evidence about such implementation. This study presented a novel method based on network theory to analysis road network structure and link network metrics with traffic congestion measures.

The presented results confirm that traffic congestion measures are positively correlated with road network centrality metrics, meaning that the occurrence of congestion is often corresponding to the importance of road segments or intersections. While no clear correlation between network entropy and congestion was identified, the impacts of the assortativity of road types on traffic congestion remains a question for future research. Lastly, congestion had a weak and negative correlation with road network fractal dimension, suggesting that density changes in a road network may affect congestion.

The application of network theory to assess the accessibility of road segments and land parcels is previously non-existent. Network analysis on the effects of the road network structure provides a cost-efficient and convenient method for identifying potential congestion and subsequently the effects of congestion on other socio-economical activities.

Chapter 3: Estimating the effects of road network metrics on retail store sales modelling

3.1. Introduction

Many retail strategies are highly elastic, such as market communications, pricing strategy, and product assortment (Levy, Weitz, & Grewal, 1998; Kapferer, 2012). In contrast, store location is relatively inelastic and often represents a long-term investment (e.g., 99 year lease and building costs). Store location is chosen among many non-controllable elements, for example, demand distribution, market area, accessibility, and competition (Huff, 2003). While brick and mortar retail retain the majority of sales (Statistics Canada, 2017), the growing proportion of sales attributed to e-commerce suggests that a new hybrid retail approach is on the horizon. In the face of online competition, only the most accessible locations can retain the offline market (Qi, 2015). The survived stores function as not only traditional physical retail sites but also showrooms of the online market, and it would possible only if the showrooms have easy access. Therefore, site-selection decision was always one of the main concerns of decision-makers in the retail sector and will remain to be a critical problem in the future.

In site-selection problems, the overarching goal is to simultaneously allocate spatially dispersed (and heterogeneous) demands to potential facility locations to optimize an objective (Goodchild, 1984). Specifically, retail site selection primarily aims to maximize profitability by allocating stores as intermediates between central facilities and prospective customers. Traditionally, retail site selection largely relied on the knowledge and experience of decision-makers using simple checklists or analogues comprising criteria identified at successful stores (Clarkson, Clarke-Hill, & Robinson, 1996; O'Malley, Patterson, & Evans, 1997; Evans, 2011; Wood & Reynolds, 2012). Frequently, these criteria and similar methods were subjectively defined and composed without objective statistical or spatial analytical approaches (Baumgartner & Steenkamp, 2011).

While these simple approaches remain frequently adapted, there is an increasing use of analytical methods such as regression, discriminant, and decision tree analyses based on empirical data. Moreover, more complex spatial interaction and optimization methods have been integrated into site-selection decision-making process (Mendes & Themido, 2004; Canbolat,

Chelst, & Garg, 2007; Duthie, Brady, Mills, & Machemehl, 2010). For example, Geographic Information Systems (Clarke, Bennison, & Pal, 1997), gravity models (Benoit & Clarke, 1997), and Artificial Neural Networks (Hernandez & Bennison, 2000).

Store accessibility is an important factor during retail site-selection (Goodchild, 1984; Arentze, Borgers, & Timmermans, 1996; Onut, Efendigil, & Kara, 2010). According to the North American Industry Classification System (NAICS), the retail sectors comprises merchandises such as motor vehicle and parts dealers, furniture and home furnishings stores, electronics and appliance stores, building material and garden equipment and supplies dealers, food and beverage stores, health and personal care stores, gasoline stations, clothing and clothing accessories store, sporting goods, hobby, book and music stores, general merchandise stores, miscellaneous stores, and non-store retailers (Statistics Canada, 2016). Driving accessibility is critical to some retailers, especially the home improvement stores, where the customers usually shop by vehicle.

While the essential approach in site-selection is to minimize travel cost between facilities and consumers, accessibility determines the marginal cost relative to travel distance (Cooper, 1964; Hakimi, 1964; Arentze, Borgers, & Timmermans, 1996). Factors that impact site accessibility include the access to roads or public transport, the level of transport, the quality of ingress and egress, and the availability of parking space (Arentze, Borgers, & Timmermans, 1996; Onut, Efendigil, & Kara, 2010). Practically, a store that provides convenient access can be more attractive to consumers. And consequentially, two neighboring analogous stores can generate significant difference in revenue due to varying accessibility.

A transportation network, particularly a road network in an urban area, is the base of traffic activities and determines the accessibilities of the parcels along the road network. However, in previous city planning studies, the description and use of road network patterns has been subjective and somewhat ambiguous as they have typically lacked quantitative measurements or converged on a set of standard measurements (Marshall, 2005; Xie & Levinson, 2007). The previous chapter illustrated a network analysis approach to quantify road segment

accessibility via traffic congestion. It provides a set of standard measurements for which different locations, regions, or service areas may be compared.

Traditionally, in the retail sector, store sales was mainly estimated based on socio-economic factors (Meade & Sarkis, 1998). This chapter builds off the previous chapter to determine the significance of road network on store sales modelling. Two research questions will be answered in the following sections: Do network metrics outperform demographic or suitability variables in retail store sales modelling? Will incorporating road network metrics improve retail store sales modelling? Regression and mathematical models will be used as sales modelling methods to investigate the research questions.

3.2. Methods

3.2.1. Study area

Ontario is located in east-central Canada, bordering the United States and four of the five Great Lakes (Figure 3-1). It is the largest province by population in Canada with about 12.85 million people (Statistics Canada, 2011), which is 38.5 percent of the total population in Canada and is 1.6 times of that of the second largest province, Quebec. Ontario is also one of the largest economic entities in Canada. Through 2011 to 2014, Ontario contributed approximately 37 percent Canada's gross domestic product (GDP) with a steady growth over the four-year period (Table 3-1). Meanwhile, the retail sector plays an important role in Ontario marketplace. With the thriving economy, the retail trade (North American Industry Classification System (NAICS), 44-45) GDP in Ontario had an average annual growth rate of 3.5% with 1.04 billion GDP annual increment from 2012 to 2014. Notably, the annual growth rate of home improvement stores (identified by NAICS 444) in Ontario from 2012 to 2014 was 4.5%, which is higher than that of the overall retail sector (3.5%).

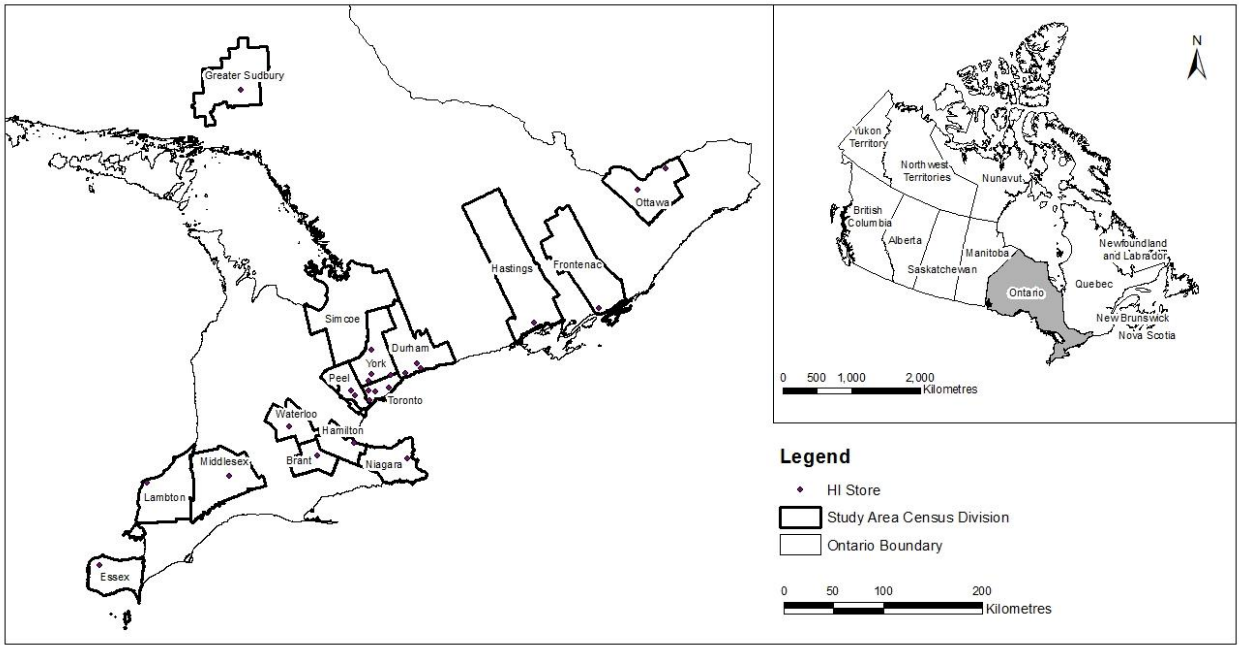


Figure 3-1 Ontario census divisions that contain stores of interest

Note: Ontario is located in east-central Canada. 26 home improvement (HI) store distributed in 16 census divisions were studied.

Table 3-1 Annual GDP at basic prices in Canada and Ontario
Reported in annual million dollars (2007 Chained dollars)

NAICS	Geography	Year			
		2011	2012	2013	2014
All industries	Ontario	570,633	578,794	585,642	600,094
	Canada	1,528,454	1,557,004	1,590,457	1,629,775
	ON-Contribution	37%	37%	37%	37%
	ON-Growth		1%	1%	2%
Retail trade [44-45]	Canada	80,685	80,995	83,627	86,651
	Ontario	29,738	29,562	30,455	31,684
	ON-Contribution	37%	36%	36%	37%
	ON-Growth		-1%	3%	4%
Building material and garden equipment and supplies dealers [444]	Canada	5,557	5,469	5,495	5,633
	Ontario	1,896	1,854	1,944	2,023
	ON-Contribution	34%	34%	35%	36%
	ON-Growth		-2%	5%	4%

Source: Statistics Canada, Table 379-0030 - Gross domestic product (GDP) at basic prices, by North American Industry Classification System (NAICS), provinces and territories, annual (dollars).

3.2.2. Data

Twenty-six home improvement retail stores were distributed in sixteen census divisions in Ontario. The sales and location information of the sampled stores were acquired and the road networks of the studied census divisions were derived from the Ontario Road Network (2011). A set of road network metrics were used in conjunction with store sales data (2013) to reveal the relationship between road network and store revenue. Meanwhile, demographic information and suitability criteria were developed from data sources from Statistics Canada, Ontario Ministry of Natural Resources, and Ontario Ministry of Transportation (Balulescu, 2015; Caradima, 2015). The derived variables were used in parallel or affiliation with road network metrics in store sales modelling.

3.2.2.1. Road Network Metrics

In the previous chapter, the road networks were measured by centralities, entropy, and fractal dimension. Among these measures, centralities are local measurements based on individual edges or nodes, while entropy, fractal, and density are global measurements that characterize the structure of a regional road network. To compare the local and global road metrics with point-based store sales, network metrics were summarized at multiple scales for each store location (Table 3-2).

Table 3-2 Road network metrics and statistics

<table>
<tr><th rowspan="3">Spatical Scale</th><th colspan="9">Network Metric</th></tr>
<tr><th colspan="3">Global</th><th colspan="6">Local</th></tr>
<tr><th>Fractal</th><th>Entropy</th><th>Density</th><th>BC</th><th>LC</th><th>CC</th><th>NDC</th><th>NCC</th><th>NLC</th></tr>
<tr><td>Census division</td><td rowspan="5">N/A</td><td rowspan="3">entropy</td><td rowspan="3">density</td><td colspan="6" rowspan="4">mean & std</td></tr>
<tr><td>Service area</td></tr>
<tr><td>5-km neighborhood</td></tr>
<tr><td>Community</td></tr>
<tr><td>Adjacent roads</td><td rowspan="2">N/A</td><td rowspan="2">N/A</td></tr>
<tr><td>Fractal area</td><td>fractal</td><td colspan="6">N/A</td></tr>
</table>

Note: BC: betweenness centrality; LC: load centrality; CC: closeness centrality; NDC: node degree centrality; NCC: node closeness centrality; NLC: node load centrality.

Aside from the fractal area used for fractal dimension calculation, five spatial scales were used for road network centrality statistics, including census division area, 19-minute-drive service area, 5-km neighbourhood, community, and adjacent roads. Specifically, sixteen census divisions

in the south-western Ontario were selected for containing stores of interest; twenty-six 19-minute-drive service areas were calculated based on network distance (Caradima, 2015); twenty-six communities were identified by strong network connections [16]; twenty-six store neighborhoods with five kilometres buffer areas; and adjacent roads were the roads that provide direct access to a store or the corresponding plaza (Figure 2-12 in section 2.3.3.2., Chapter 2).

The statistics on network metrics produced 69 variables (Table 3-2). The large amount of data input would potentially create a computational overhead and reduce model efficiency. To reduce the number of network metrics in modelling and determine the statistically significant metrics, a stepwise regression was performed with a threshold of p-value<0.12 to select a subset with proper size. Nine network metrics were selected (Table 3-3): entropy at community level (*ETP*), mean of closeness centrality at 5km neighborhood area (CC_{avg}), standard deviation of closeness centrality at community level (CC_{std}), sum of node closeness centrality at community level (NCC_{sum}), mean of node closeness centrality at community level (NCC_{avg}), standard deviation of betweenness centrality at community level (*BC*), mean of node load centrality at adjacent roads (*NLC*), sum of degree centrality at service area (DC_1), and sum of degree centrality at 5km (DC_2).

3.2.2.2. Demographic attributes

Balulescu (2015) proposed five demographic variables and one site variable for retail store sales modelling (Table 3-3). These demographic variables were immigrant population (*Imm*), average dwelling value (D_V), count of dwelling owner (D_O), dwelling counts (D_V), and households with income over CAD 100,000 (*Inc*).Demographic data were derived from the 2011 Census and National Household Survey (NHS) and were calculated in a 19 minutes network-drive-time service area. Statistics Canada conducts a national survey every five years. In 2011, the long mandatory census was replaced by a combination of a short census and the NHS, which is a detailed voluntary survey. The census data covers topics of population and dwelling counts, age and sex, families, households and marital status, structural type of dwelling and collectives, and language.

[16] Community detection was implemented by "community" algorithm in NetworkX package.

The NHS data includes immigration, income and housing, etc. (Statistics Canada, 2011). And the proposed site variable was store area (*S*), which as calculated by digitizing store foot print.

3.2.2.3. Suitability criteria

Caradima (2015) identified nine site and situational criteria for retail site suitability. The criteria include topography, traffic, transportation, market, and expenditure (Table 3-3; for more details see Appendix A). The data were derived from primary datasets such as digital elevation model (DEM; Ontario Ministry of Natural Resources), annual average daily traffic (AADT; Ontario Ministry of Transportation), Ontario road network (ORN, Ontario Ministry of Transportation), retail store information, and census data. Most of the primary data are publicly available; however, retail store location and store attributes were geo-coded and digitized in ArcGIS with Google Maps API.

Table 3-3 Variable, symbol, and description

Group	Variable Name	Symbol	Description
Network	Entropy	ETP	Entropy at community level.
	Closeness centrality mean	CC_{avg}	Mean of closeness centrality at 5km neighborhood area.
	Closeness centrality standard deviation	CC_{std}	Standard deviation of closeness centrality at community level.
	Node closeness centrality sum	NCC_{sum}	Sum of node closeness centrality at community level.
	Node closeness centrality mean	NCC_{avg}	Mean of node closeness centrality at community level.
	Betweenness centrality	BC	Standard deviation of betweenness centrality at community level.
	Node load centrality	NLC	Mean of node load centrality at adjacent roads.
	Degree centrality at service area	DC_1	Sum of degree centrality at service area.
	Degree centrality at 5km	DC_2	Sum of degree centrality at 5km.
Demographic	Immigrants	Imm	Total population identified as immigrant in the service area.
	Average dwelling value	D_V	Average value of dwelling in the service area.
	Dwelling owner	D_O	Count of owned dwellings in the service area.
	Store area	S	Area of a retail store footprint in square feet.
	Dwelling counts	D_C	Count of dwellings in the service area.
	Income over CAD 100,000	Inc	Count of households with income over CAD 100,000.
Suitability	Site maximum slope	b	Maximum value of the parcel's slope.
	Traffic visibility	v	Defined base on distance from the major highways and the traffic volume.
	highway accessibility	r	Travel time from a parcel to the nearest highway access point (i.e., ramp).
	distance to distribution centre	d	The network distance to the nearest distribution centre.
	market representation	l	Location quotient of home improvement retail a dissemination area.
	density of competitors	d_c	The number of competitors per unit area in the service area.
	density of retail stores	d_r	The number of retailers per unit area in the service area.
	Potential expenditures	e_p	Estimated expenditure without competitors in the service area in a Huff's model.
	Competitive expenditures	e_c	Estimated expenditure with competitors in the service area in a Huff's model.

3.2.3. Model selection

3.2.3.1. Categories of predictors

The aforementioned predictor groups (i.e., network metrics, demographic variables and suitability criteria) were used in isolation and combination for store sales modelling. The nomenclature used in this presented study was consistent among all tested models. N stands for

network metrics, D stands for demographic variables, and S stands for suitability criteria. Therefore, the isolated variable groups were denoted as N, D, or S; the combined variable groups were denoted as ND, NS, DS, or NDS (Table 3-4).

Table 3-4 Categories of predictors of sales models

Index	Categories of predictors
N	Network metrics
D	Demographic variables
S	Suitability criteria
ND	Network metrics and demographic variables
NS	Network metrics and suitability criteria
DS	Demographic variables and suitability criteria
NDS	Network metrics, demographic variables, and suitability criteria

3.2.3.2. Predictor selection

Predictors were selected from each variable group using cross-validation as variable evaluation scheme. The number of potential combination of variables varies with the number of predictors and the number of the variable candidates in each group. Take network metrics for example, there were $C_9^1 = 9$ one-predictor combinations, $C_9^2 = 36$ two-predictor combinations, $C_9^3 = 84$ one-predictor three-predictor combination, etc.

Both 10-fold (denoted as 10-F in the following content) and Leave-P-Out (denoted as LPO in the following content) cross-validation schemes were used for comparison. During a 10-F cross validation, the input dataset is split into ten groups, then one group is selected as test group and the remaining groups are used as training data. This process is repeated iteratively until all groups have been tested. The LPO cross-validation has similar mechanism but has a test group of size p (p=2 in this presented study, so it is also denoted as L2O). The test group(s) is (are) selected using an exhaustive enumeration (scikit-learn developers, 2017). In this presented research, the LPO produced $C_{26}^2 = 325$ validation comparisons.

At the validation stage, the trained models were fitted by the test data and were scored by the mean squared error (MSE). A smaller MSE indicates less information loss and better sales modelling. Therefore, variable combinations with small MSE will be selected as model inputs.

3.2.4. Store sales modelling methods

Three sales modelling methods are presented in this chapter to investigate the significance of road network, demographic, and suitability factors on store sales and to test if the inclusion of road network metrics improves sales modelling. The first modelling method is backwards Ordinary Least Square (OLS) regression; the second modelling method is Partial Least Square (PLS) regression; the last modelling method is mathematical modelling (MM) via an artificial intelligence enabled software[17].

3.2.4.1. Backwards OLS regression model

Stepwise regression is a semi-automated process for model building and variable subset selection (Hengl, Heuvelink, & Stein, 2004). It is an effective coefficient estimation method in a general linear model when the number of predictors is large and the data are limited. Backwards stepwise regression determines the significances of variables based on a sequence of t-test and R-squared values, then uses a greedy variable selection algorithm to remove the variable with p-values below the threshold (0.1 in the presented research) in backwards eliminations. The resulting model contains only statistically significant variables in a store sales modelling.

3.2.4.2. PLS regression model

The OLS models regressed a set of predictors against the sales data but did not detect nor eliminate multicollinearity among predictors. To detect multicollinearity, correlation test was performed among predictors. Also, unlike an OLS model, PLS reduces multicollinearity among predictors by project predictors and response to an orthogonal space.

Considering that some of the components add little explanatory power to a model, a leave-one-out cross-validation was used for component reduction. During the validation process, PLS starts from a model with a single component, and one observation is omitted from modelling. Then the resulting model is fitted to the test data to generate residual and R-squared. The process is repeated until all observations have been omitted for one time, prediction residual sum of squares (PRESS) and predicted R-squared values are calculated as the average of the test results.

[17] Eureqa by Nutonian.

Then another component is added into the model and the cross-validation procedure is repeated until all models (all components have been added) have been validated. The model with the lowest PRESS and the highest predicted R-squared will be chosen. Moreover, the variables are rescaled to standardize the deviations to 1, therefore the results would be unbiased regarding the scales of variables.

3.2.4.3. Mathematical models

Predetermined algorithms and hypotheses are required in traditional data analysis; however, it is not unusual to have a large set of predictors and implicit relationships with the dependent variables. Instead of manually specifying the functional form of model, a mathematical modelling software (i.e., Eureqa) provides a data-driven analysis that searches for the best-fit model. This artificial intelligence enabled software is capable of iteratively testing a wide choice of algorithmic building blocks, including addition and subtraction, trigonometry, and exponential, among others. The method has the advantage of generating highly fitted models; however, the result is not always interpretable because the empirical reasoning is omitted from the modelling process.

3.2.4.4. Model assessment

The models were assessed regarding complexity (number of coefficients), information loss (sum of squared errors (SSE), Akaike information criterion (AIC), and mean squared error (MSE)), and goodness-of-fit (R-squared and adjusted R-squared). Notably, mathematical modelling may produce non-linear models, where the uses of R-squared and adjusted R-squared are controversial (Spiess & Neumeyer, 2010). Although they may not reflect the explanatory power of non-linear models, R-squared was calculated to indicate and compare model residual with the following formula:

$$R^2 = 1 - \frac{SSE}{SST} \qquad \textit{Equation 3-1}$$

where SSE is the sum of squared error, and SST is the total sum of squared.

3.3. Results

3.3.1. Model selection

A model with lower average MSE during cross-validation yields less information loss in sales modelling. In a group of models with the same number of predictors, the best model is identified by the minimum MSE and the worst model is identified by the maximum MSE. Sometimes the two cross-validation schemes (LPO and 10-F) ranked the model differently (Table 3-5).

The minimum MSE of models varies according to the predictor groups and the number of predictors (Figure 3-2). Among models based on network metrics, additional predictors improve model performance; while among models based on demographic variables or suitability criteria, increasing the number of predictors did not always decrease MSE. For demographic variables, the lowest MSE observed via L2O was on a model with three predictors (4.19E13) and the lowest MSE observed via 10-fold was on a model with five predictors (4.18E13). For suitability criteria, models with two (4.60E13 via L2O, 4.69E13 via 10-Fold), three (4.52E13 via L2O, 4.71E13 via 10-Fold), and four (4.55E13 via L2O, 4.77E18 via 10-Fold) predictors yielded lower MSE than models of other sizes.

Table 3-5 MSE of the best models of each variable group in cross-validation
Reported in squared million dollars

Number of Variables	Network		Demographic		Suitability	
	L2O	10-Fold	L2O	10-Fold	L2O	10-Fold
1	41.52	46.07	50.61	51.65	48.30	50.80
2	36.52	42.17	44.07	45.51	46.01	46.93
3	28.45	29.45	41.95	41.84	45.18	47.07
4	24.61	24.22	45.02	43.41	45.53	47.67
5	20.49	21.27	48.14	41.80	47.54	48.96
6	14.53	14.47	53.13	45.90	51.43	50.19
7	10.31	10.62	-	-	55.80	54.56
8	7.50	7.34	-	-	60.66	57.35
9	6.62	7.19	-	-	69.29	65.21

Note: See Appendix C for more details about predictor selections.

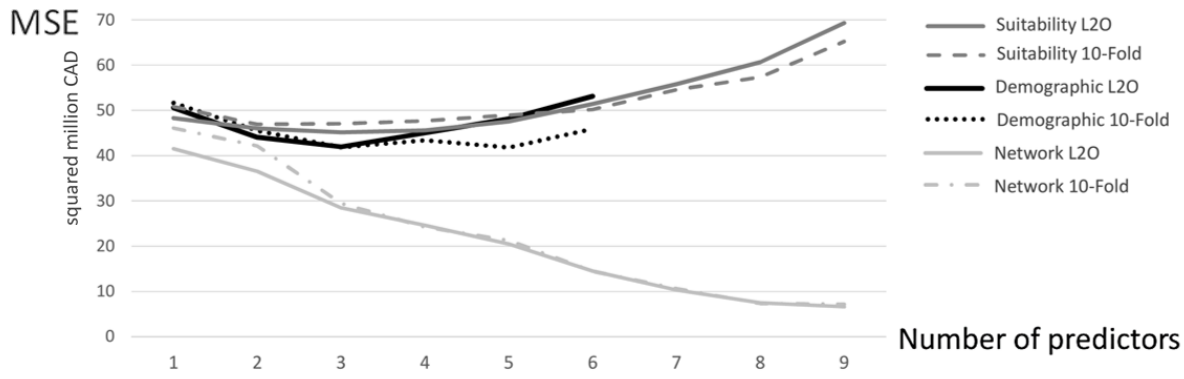


Figure 3-2 Variable selection via cross-validation with the highest MSE of each variable group.

In social research, the number of predictors should match sample size to yield an unbiased regression result (VanVoorhis & Morgan, 2007). When the number of predictors is small, Harris (2001) suggested that sample size should be 50 larger than the number of predictors. And Green (1991) suggested that sample size should be 50 larger than eight times of the number of predictors. However, the available sales data was limited to 26 stores in this study. Hence, the number of predictors should be minimized and the exploratory power should be retained as possible.

Although the MSEs of models with network metrics reduced with the increase of the number of predictors, the MSEs of models with demographic or suitability variables were comparable when the number of predictors was 2, 3, or 4. Therefore, the model size was limited to two predictors in this study.

Another problem encountered in model selection was that the best models recognized by L2O and 10-F cross-validations were inconsistent (Appendix C). For example, the combination of *ETP* and CC_{std} was recognized as the best in L2O with an MSE of 3.65E13 and ranked as the second best in 10-F with an MSE of 4.27E13; while the combination of NCC_{avg} and DC_1 was recognized as the best in 10-F with an MSE of 4.22E13 but only ranked as the fourth best model in L2O with an MSE of 4.35E13. Considering the overall performance, the model with *ETP* and CC_{std} was more stable than the other model and was therefore selected for further analysis. Moreover, *Imm* and D_V in demographic variables and e_p and e_c in suitability criteria outperformed other models in both L2O and 10-F cross-validations.

3.3.2. Linear regression models

The OLS regression models were established based on isolated or combined predictor groups. The results showed the influence of each predictor on store sales (Table 3-6). The Pearson coefficient indicated the degree of effect and the p-value represented the significance of effect.

Table 3-6 Backward stepwise regression model and variable selection
Coefficients in **bold**, p-values in *Italic*

Predictor		OLS-N		OLS-D		OLS-S		OLS-ND		OLS-NS		OLS-DS		OLS-NDS	
		coef.	p	coef.	p	coef.	p	coef.	p	coef.	p	coef.	p	coef.	p
Network	*ETP*	**-1.48E+07**	*0.01*					**-1.13E+07**	*0.02*	**-1.28E+07**	*0.01*			**-1.32E+07**	*0.01*
	CC_{std}	**-2.50E+12**	*0.03*					**-3.07E+12**	*0.01*	**-2.52E+12**	*0.02*			**-3.55E+12**	*0.00*
Demographic	*Imm*			**-11.58**	*0.03*			**-10.86**	*0.02*			**-18.76**	*0.01*		
	D_V			**44.60**	*0.04*			**46.20**	*0.02*			**35.20**	*0.09*	**41.00**	*0.03*
Suitability	e_c					**-0.01**	*0.03*			**-0.01**	*0.04*			**0.00**	*0.03*
	e_p					**1.44**	*0.04*			**1.16**	*0.06*	**0.90**	*0.08*		
Constant		**59523945**	*0*	**19406816**	*0.003*	**18980303**	*0.005*	**43436080**	*0*	**46995868**	*0*	**12804930**	*0.067*	**49748610**	*0*
Coefficients		2		2		2		4		4		3		4	
SSE/sqr million		741.12		894.45		916.43		558.86		606.90		774.71		576.16	
AIC		810.03		814.92		815.55		808.07		810.22		813.75		808.87	
R-Sq		0.34		0.20		0.18		0.50		0.46	0.46	0.31	0.31	0.49	0.49
R-Sq(adj)		0.28		0.14		0.11		0.41		0.36	0.36	0.22		0.39	0.39

3.3.2.1. OLS with isolated predictor groups

In model OLS-N, both entropy at community (*ETP*) and closeness centrality standard deviation (CC_{std}) were included at a significance level of 0.05. And both were negatively correlated with annual store sales.

Entropy (*ETP*) represents the heterogeneity of road categories at a community level. A high *ETP* value indicates a high assortativity of road categories, and a low *ETP* value implies that the road network is dominated by a single category of road segments.

At a community level, the standard deviation of closeness centrality (CC_{std}) indicates the variance of closeness centrality among a road network. A regional road network can be divided into three parts: the “centroid”, which has high closeness centrality; the “periphery”, which has low closeness centrality; and the “connection”, where the variance of closeness centrality is high. A high CC_{std} is observed in community road networks distributed in the “connection” part of a regional network where the variance of closeness centrality is large; community road networks with small CC_{std} values are at either the “centroid” or “periphery” of a regional network where the CC_{std} is more stable and has less variance.

Spiess and Neumeyer (2010) suggested that big-box retail stores are ideally allocated to industrial zones with easy highway access where both land price and accessibility are optimized. Such areas are more commonly seen outside of city centres. It is corroborated by the regression results of OLS-N that high store sales is more likely to be observed in an area with less road assortativity and at the periphery of an urban area.

In model OLS-D, both immigrant population (*Imm*) and dwelling value (D_V) were included at the significance level of 0.05. The coefficient of *Imm* was -11.58, indicating a negative effect on store sales. However, the impacts of immigrants on the economy are ambiguous in the literature (Bodvarsson & Van den Berg, 2006; Bodvarsson, Van den Berg, & Lewer, 2008). On one hand, immigrants bring retail demand to the local marketplace, but on the other hand, the exogenous immigrated labors may lead to a wage fall or even out-migration, and the retail market will shrink especially when the labor demand is not wage elastic. In this study, the negative effect of immigrant population on store sales implies that the impacts of immigration population on the shrink of retail market is stronger than that on the increase of retail demand.

Meanwhile, dwelling value (D_V) showed a positive impact at sales of home improvement retail stores. Past studies have shown that the correlation between neighborhood demographic and retail activity may differ by retail sectors. Meltzer and Schuetz (2012) found that in New York City, the retail establishment is significantly denser and more diverse in higher income neighbourhoods, while the lower income neighbourhoods have more accesses to the necessities, which may be lower in both quality and cost. For example, the average dwelling value was slightly lower in neighbourhoods with a supermarket in a study conducted in Edmonton, Canada (Smoyer-Tomic, et al., 2008). Compared with supermarkets and grocery stores, home improvement retail stores provide products and services that are of higher cost and require less frequent purchases so that customers would be more likely to travel farther. Therefore, the incitation of allocating a home improvement store to a low-income neighborhood should be lower than other retailers.

In model OLS-S, the expenditures are the allocated demand to a potential store location (Balulescu, 2015). The impacts of competitive expenditure (e_c) and potential expenditure

without competition (e_p) on store sales had different directions: the competitive expenditure had a slightly negative impact on sales with a coefficient of -0.0086, while the potential expenditure showed a positive impact with a coefficient of 1.436. Since stores with higher demand are more likely generate higher sales, the negative impact of competitive expenditure can be a result of its collinearity with potential expenditure. Also, potential expenditure showed a stronger influence on store sales and therefore might have an effect on parameter estimation of the other predictor. However, both expenditure variables may have more complex relations to store sales, given the large residuals during a linear regression.

The sizes of predictor sets of models based on isolated variable groups were identical, all of these models had two predictors. However, there were differences in information loss among the models. Model OLS-N had lower SSE (741.12) and AIC (810.03) than that of OLS-D (SSE: 894.45, AIC: 814.92) or OLS-S (SSE: 916.43, AIC: 815.55). Also, OLS-N yielded the highest adjusted R-squared of 0.28, while the adjusted R-squared of OLS-D and OLS-S were significantly lower (0.14 and 0.11). Therefore, when included as a single predictor group in regression on sales, network metrics were more influential on sales than demographic and suitability variables.

3.3.2.2. OLS with combined predictor groups

The OLS models with combination of predictor groups retained the same (or similar) coefficients and the corresponding confidence levels with those in models with isolated predictor group (Table 3-6).

The assessments of OLS models with combined predictor groups showed that the inclusion of network metrics in sales modelling had improved the model quality. Firstly, the information loss of OLS-ND (SSE: 558.86, AIC: 808.07) and OLS-NS (SSE: 606.90, AIC: 810.22) were lower than that of either OLS-D or OLS-S. In contrast, without network metrics, model OLS-DS (SSE: 774.71, AIC: 813.75), which was established on the combination of demographic variables and suitability criteria, showed less improvement relative to OLS-D or OLS-S regarding SSE or AIC. Lastly, model OLS-NDS, which included all three categories of predictors (but *Imm* and e_c were omitted in backward selection) yielded the second lowest information loss (SSE: 576.16, AIC: 808.87) among the seven OLS models.

The inclusion of network metrics in sales modelling also improved the goodness of fit. When combined with network metrics, OLS-ND and OLS-NS were greatly improved from OLS-D and OLS-S regarding adjusted R-squared (0.41 and 0.36 respectively). Whereas, without network metrics, OLS-DS generated an adjusted R-squares of 0.22, where the improvement from OLS-D or OLS-S was less significant. Moreover, OLS-NDS included three groups of predictors and yielded the second highest adjusted R-squared of 0.39.

3.3.3. PLS Regression models

Strong correlations (Pearson correlation coefficient was greater than 0.80 at significance level of 0.01) were identified among all pairs of demographic and suitability predictors (Table 3-7; see Appendix B for correlations among the full variable set), indicating multi-collinearity exists in model OLS-D, OLS-S, OLS-DS, and OLS-NDS. Correlation among all the other variables combinations were below 0.35 and insignificant, except for the correlation between CC_{std} and D_V, which was significant at 0.1.

Table 3-7 Pearson correlation coefficient between predictors
**** P < 0.01; ** P < 0.05; * P < 0.1*

Group	Variable	*ETP*	CC_2	*Imm*	D_V	e_c
Network	*ETP*					
	CC_{std}	-0.31				
Demographic	*Imm*	-0.04	0.183			
	D_V	-0.23	0.339 *	0.854 ***		
Suitability	e_c	-0.07	0.081	0.968 ***	0.821 ***	
	e_p	-0.16	0.113	0.894 ***	0.824 ***	0.939 ***

PLS was then implemented to eliminate multi-collinearity (Table 3-8). During a PLS regression, the predictors and response were projected to a multi-dimensional orthogonal space and redundancy was removed via dimension reduction instead of predictor selection. Therefore, all input predictors were included in PLS results. Moreover, standardized coefficients were calculated to indicate the unbiased contributions of predictors in the models. They were calculated corresponding to the standardized predictors and response, which were derived by rescaling variable variances to one.

Table 3-8 PLS regression models

Predictor		PLS-N		PLS-D		PLS-S		PLS-ND		PLS-NS		PLS-DS		PLS-NDS	
		Coef	Std Coef.	Coef	Std Coef.	Coef	Std Coef.	Coef	Std Coef.	Coef	Std Coef.	Coef	Std Coef.	Coef	Std Coef.
Network	*ETP*	-14832900	-0.5602					-11330900	-0.4280	-12797300	-0.4833			-11038700	-0.4169
	CC_{std}	-2.50E+12	-0.4130					-3.07E+12	-0.5061	-2.52E+12	-0.4159			-3.26E+12	-0.5385
Demographic	*Imm*			-12	-0.8576			-10.8624	-0.8042			-12	-0.8640	-7.83958	-0.5804
	D_V			45	0.8026			46.1737	0.8304			42	0.7620	45.3097	0.8149
Suitability	e_c					0	-1.2417			-0.00707033	-1.0266	0	-0.4553	-0.00374462	-0.5437
	e_p					1	1.2010			1.15981	0.9701	1	0.5653	0.435098	0.3639
Constant		59523900	0.0000	19406816	0.0000	18980303	0.0000	43436100	0.0000	46995900	0.0000	13668624	0.0000	40109000	0.0000
Coefficients		2		2		2		4		4		4		6	
SSE/sqr million		741.12		894.45		916.43		558.86		606.90		815.29		515.65	
AIC		811.51		816.40		817.03		808.17		810.31		817.99		810.08	
R-Sq		0.34		0.20		0.18		0.50		0.46		0.27		0.54	
R-Sq(adj)		0.28		0.14		0.11		0.41		0.36		0.14		0.40	

3.3.3.1. PLS with isolated predictor groups

The directions of signs and magnitudes of coefficients of the PLS models with isolated predictor groups were comparable to those in OLS models. Model PLS-N was established on network metrics and both of *ETP* and CC_{std} had negtive effects on sales. In model PLS-D, *Imm* contributed negatively while D_V contributed positively. In model PLS-S, e_c and e_p still had different impacts on sales

All PLS models with isolated predictor groups had the identical size of two predictors. In terms of information loss, the models established on demographic (PLS-D, SSE: 894.45, AIC: 816.40) and suitability (PLS-S, SSE: 916.43, AIC: 817.03) predictors produced higher SSE and AIC than the model based on network metrics (PLS-N, SSE: 741.12, AIC: 811.51). The results of information loss showed a significant difference of PLS-N with PLS-D and PLS-S. Also, model PLS-N outperformed PLS-D and PLS-S regarding the goodness-of-fit. PLS-N had an adjusted R-squared value of 0.28, while the adjusted R-squared of PLS-D and PLS-S were 0.14 and 0.11. Therefore, when included as a single predictor group in PLS regression on sales, network metrics were more influential on sales than demographic and suitability factors.

3.3.3.2. PLS with combined predictor groups

The PLS models with combined predictor groups kept the similar coefficients with that in PLS models with isolated predictor groups. Among the three groups of predictors, demographic variables (*Imm* and D_V) had more significant impacts on sales since their standardized coefficients were larger than the other predictors in absolute values. And according to the

standard coefficients in model PLS-NS, suitability criteria (e_c and e_p) were more important than network metrics; while in model PLS-NDS the two groups of predictors played comparable roles.

Each of the PLS models with two predictor groups had the size of four, and model PLS-NDS had a size of six. The largest information loss was produced by model PLS-DS (SSE: 815.29, AIC: 817.99), followed by PLS-NS (SSE: 606.90, AIC: 810.31) and PLS-ND (SSE: 558.86, AIC: 808.17). Model PLS-NDS had the least information loss (SSE: 515.65, AIC: 810.08) among the four PLS models.

The goodness of fit corresponded to information loss of these PLS models. PLS-ND had the best fit with an adjusted R-squared of 0.41, followed by PLS-NDS (adjusted R-squared: 0.40) and PLS-NS (adjusted R-squared: 0.36). The worst goodness of fit was produced by model PLS-DS, which had an adjusted R-squared of 0.14.

The performance of PLS models was very similar to the OLS models according to complexity, information loss, and goodness-of-fit. PLS-ND had the best performance in sales modelling with less complexity, better fitting and less information, and PLS-NDS generated comparable results.

3.3.4. Mathematical modelling

Seven non-linear models have been developed corresponding to the seven predictor groups via a mathematical modelling software (Eureqa by Nutonian; Table 3-9; for details see Appendix Table C-5).

Table 3-9 Mathematical modelling summary

	MM-N	MM-D	MM-S	MM-ND	MM-NS	MM-DS	MM-NDS
Coefficients	6	9	9	6	7	8	6
MSE/sqr million	7.62	8.92	13.74	7.88	9.22	3.98	5.16
AIC	787.64	804.54	815.78	788.52	796.37	778.82	777.48
R-sq	0.82	0.79	0.68	0.82	0.79	0.91	0.88

Because mathematical modelling uses non-linear modelling blocks, the number of coefficients may not correspond to the number of predictors. The number of coefficients ranged from six to nine among the seven models. Also, model quality did not always match model

complexity. For example, MM-S had nine coefficients but yielded the largest information loss (MSE: 13.74, AIC: 815.78) among the seven models, while MM-NDS had only six coefficients and produced the least information loss (MSE: 5.16, AIC: 777.48).

Among the models with isolated predictor groups, MM-N produced the best quality of sales modelling with the lowest information loss (MSE: 7.62E12, AIC: 787.64) and the highest goodness of fit (R-squared of 0.82). However, in contrast with OLS and PLS models, adding network metrics into mathematical modelling did not always improve the quality of models. Model MM-DS (MSE: 3.98, AIC: 778.82, R-squared: 0.91) outperformed MM-ND (MSE: 7.88, AIC: 788.52, R-squared: 0.82) and MM-NS (MSE: 9.22, AIC: 796.37, R-squared: 0.79). Except for the uncertainty contained in the differences in model complexity and model building time, the higher performance of demographic and suitability variables in mathematical modelling might be because of nonlinear relations or interactions among the socio-economic variables and sales were captured by the models. Moreover, model MM-NDS combined three groups of predictors and reduced the number of coefficients but also decreased the model quality regarding MSE (5.16), AIC (777.48), and R-squared (0.88).

3.4. Discussion

This presented study finds that across all models tested (i.e., OLS, PLS, and mathematical modelling), road network metrics played a very important role in retail store sales modelling. While most retail site location analyses are simple regression or suitability analysis, this study has shown that the road network metrics derived from network analysis outperformed traditional demographic and suitability variables. Specifically, among the OLS models, OLS-N had better fit (adjusted R-squared: 0.28) and less information loss (SSE: 741.12, AIC: 810.03) compared to OLS-D (adjusted R-squared: 0.14, SSE: 894.45, AIC: 814.92) or OLS-S (adjusted R-squared: 0.11, SSE: 916.43, AIC: 815.55). The results in PLS were very similar to that in OLS, the network metrics outplayed the other two variable categories. And network metrics remained the best overall performance in mathematical modelling, where the network metrics yielded R-squared of 0.82 and AIC of 787.64 while the demographic and suitability variables had R-squared of 0.79 and 0.68 and AIC of 804.54 and 815.78 respectively.

The inclusion of network metrics in sales modelling improved the quality of the model, especially OLS and PLS models. When integrated with network metrics, OLS-NS, OLS-ND and OLS-NDS had great improvements from the OLS models based on isolated predictor groups while OLS-DS had a less significant improvement. Similarly, among the PLS models based on combined predictor groups, PLS-ND, PLS-NS, and PLS-NDS had significantly higher adjusted R-squared values and less information loss than PLS-DS, which did not improve much from PLS-D not PLS-S. In more complex modelling approaches (i.e., mathematical and non-linear modelling), the road network metrics remained highly influential but the traditional metrics continued to play dominant roles.

3.4.1. Parameter estimation with small sample size and collinearity

Two challenges encountered in the presented research to investigating the effects of network, demographic, and suitability variables on store sales, which were small sample size and multicollinearity. Collinearity is the interdependency between a pair of predictors in a regression (Farrar & Glauber, 1967), and multicollinearity exists when more than two predictors are correlated. In an OLS regression, the parameters are estimated based on the observations of the predictors and the response. Therefore, the accuracy of parameter estimation is determined by (1) the quality of the sample observations (e.g., are they unbiased and do they reflect the real distribution of the total population?); (2) sample size (e.g., is the sample size large enough regarding the number of parameters?); and (3) experimental design (e.g., are the predictors well designed to exclude collinearity?). However, in social studies, the sample size is often limited and the collinearity (or multicollinearity) is commonly inherent (Baguley, 2012).

In a regression model with collinearity, parameter estimation is more imprecise and unstable as there will be a large variance of the estimated parameter with the absence or presence of other predictors. In an extreme case where there is a perfect correlation between two predictors, the inclusion of the second predictor does not create additional information to the model so that the associated parameters cannot be determined (or say, there are infinity solutions for parameters estimation). In a less severe case where the correlation is high between two predictors, the parameters can be determined but with a large inflation (which is reflected by variance inflation factor, i.e., VIF). Generally, although the existence of collinearity does not

affect the overall performance of a regression model regarding predictive power, it is often difficult to interpret the effect of an individual predictor if the parameter of the predictor is not precise.

In this study, high correlations existed among the three predictor groups. Especially, there were high pair-wise correlations (Pearson correlation coefficient > 0.8 at 0.01 level) among many of the demographic variables and suitability criteria (Appendix B). Furthermore, because of restrictions on the availability of confidential sales data, the sample size is limited in this study (sales data available from only 26 stores of interest). On the other hand, despite the sampling method, the probability of bias increases in parameters estimation when the sample size is too small (e.g., less than 50). For example, the true correlation between competitive expenditure (e_c) and potential expenditure (e_p) was not captured in the sample. Among the 162,692 land parcels in the southwestern Ontario, the correlation between e_c and e_p is 0.72, which is smaller than the measured correlation in this study (0.939).

Collinearity is an inherent problem in most social studies (Baguley, 2012). It is often tolerable if the degree of collinearity is mild or the variable of interest is not involved (Farrar & Glauber, 1967). In other cases, significant collinearity can be eliminated by (1) removing a variable from the correlated pair or(2) using dimension redundancy methods, such as Partial Least Squares (PLS) or Principal Component Analysis (PCA). The implementation of PLS is presented in this chapter, and a comprehensive comparison among these methods is displayed in Appendix D. Whereas, none of these methods seem effective in reducing collinearity while maintaining good explanatory power in this study. Therefore, this study provides an exploratory evaluation of the effects of selected road network, demographic, and suitability factors on store sales. Subsequent investigations of the effects of certain factors require adequate sales data (or sales estimation as an alternative) in future studies.

3.4.2. Contributions and future directions

This chapter identified the strong effects of road network metrics relative to demographic and suitability variables on store sales. The findings suggest that road network metrics are more influential in store sales modeling and should be incorporated in future studies to improve model

explanatory power. This presented study is based on the data provided by a home improvement retail chained store in Ontario, Canada, who is also the direct beneficiary of this study. Therefore, extending and adapting of the findings in this study to other businesses or other regions need further examination. The findings also have implications for city-planners in research on the relationship between road network and land-use intensity, and can be eventually used to facilitate land-use zoning and retail planning.

3.5. Conclusion

The structure of road infrastructure affects a retail store's performance via store accessibility and therefore should be incorporated in retail site selection approach. This study explored the explanatory power of explicit road network metrics on store sales. Across all models tested in this study, road network metrics played a very important role in predicting site location success. They outperform demographic and suitability variables, which are traditionally used in site location analyses. Moreover, the inclusion of road network metrics improved the quality of store sales modeling especially in OLS and PLS models. In more complex modelling approaches (i.e., mathematical and non-linear modelling) the network metrics remained highly influential but the traditional metrics continued to play a dominant role.

However, this study is just an exploratory of the use of road network metrics in store performance modeling in Ontario based on the implementation of network theory in quantitative road network analysis. Further research is recommended for the adaption of the findings in this study in other businesses or regions.

Chapter 4: Conclusion

This Master's thesis confirmed the influence of road network structure on traffic congestion and retail store sales, and in store sales modelling, road network metrics has outperformed other tested socio-economic variables and improved model performance.

In Chapter Two, a novel method based on network theory was presented to analyze road network structure and to establish a link between network metrics and traffic congestion measures. Specifically, network metrics, including centralities (e.g., betweenness, closes, load and degree centrality), entropy, and fractal dimension, were derived for the City of Toronto road network. OLS or correlation test was used to investigate the relationship between road network metrics and congestion measures. Congestion measures were found to be positively correlated with network centralities, especially closeness centrality, during peak travel hours. However, there was no solid conclusion for the impacts of road entropy and fractal dimension on traffic congestion since their relationships were either ambiguous or very weak in this study.

Chapter Three furthered this study by exploring the predictive power of road network metrics on retail store sales. Road network metrics were compared to traditional site location analysis variables (i.e., demographic and suitability variables) in linear and non-linear store sales models. When incorporated as an isolated predictor group, road network metrics outperformed other variables in store sales modelling; when combined with other variables, road network metrics also improved sales models regarding increased exploratory power and decreased information loss. Although, in a more complex mathematical modelling approach, traditional variables continued to play a dominant role, network metrics remained highly influential.

4.1. Limitations

One of the major limitations encountered in this study was a small sample size due to restricted access to confidential sales data. Specifically, in the second chapter, there were only five stores of interest in City of Toronto census division. Consequentially, the global network metrics (i.e., entropy and fractal dimension) had only five observations and their correlations with traffic congestion were ambiguous. In the third chapter, twenty-six stores in Ontario were available for sales data. The sample size was so small that the distribution of variables could be biased.

According to the Law of Large Numbers, increasing sample size allows better estimations of the distributions of total population and therefore improves the performance of sales modelling. However, store sales data are often confidential information and have restricted access. Alternatively, modelled sales data can be used if quality and reliability of sales are not the main concerns in a study.

Another issue encountered in this presented study was the inclusion of dependency, interaction, and correlation in models. It violated the assumption of independency and caused imprecise and unstable model estimations. Therefore, such relationships among variables should be explicitly addressed in modelling. Specifically, network centrality of road segments inherently contains dependency and interaction as it represents the relationship between network entities. To address the reciprocation among network metrics in the road network, statistical models like exponential random graph models (ERGMs) are recommended. Also, as measures of the structure of spatial features (i.e., the road network), network centralities and other network metrics are prone to have spatial autocorrelations. To address this issue, residual of empirical models should be tested by network Moran's I (Chun & Griffith, 2011), and if the spatial correlation is significant, spatial regression models (e.g., spatial lag, spatial error, geographically-weighted regression) should be used instead (Anselin, 2002). Moreover, strong collinearity was detected among variables in sales models in Chapter Three. To reduce the degree of collinearity, experimental design should be refined by expanding variable selection and enlarging sample size.

4.2. Contributions and future directions

This thesis presents an exploration of the implementation of network theory in quantitative road network analysis and the use of road network metrics in store sales modelling. It provides a method for city-level road network analysis especially when the access to detailed traffic data is limited. The correlations among road network, store sales, and other socio-economic factors reflect that road network structure may be capable of capturing the variance of the socio-economic variables via spatial configuration and relative locations. Therefore, network analysis of road system can be used to not only facilitate traffic congestion estimation but also locate retail stores. However, the findings of this study are restricted to Ontario, Canada and require further tests before being adapted to other businesses and regions.

Road network metrics were found to have strong correlation with traffic congestion and very influential on retail store sales in this study. Therefore, road network analysis is recommended to future retail site-selection studies. Adequate data sources for real traffic data and real sales data (or estimations as an alternative) are required in future studies to eliminate the potential bias in data. Meanwhile, a comprehensive selection of variables would be beneficial for preventing collinearity in sales modelling via variable selection. Furthermore, spatial statistical methods are recommended to incorporate the spatial heterogeneity of road networks.

References

Al-Jarrah, O., & Abu-Qdais, H. (2006). Municipal solid waste landfill siting using intelligent system. *Waste management, 26*(3), 299--306.

Allen, B. W., Liu, D., & Singers, S. (1993). Accesibility measures of US metropolitan areas. *Transportation Research Part B: Methodological, 27*(6), 439-449.

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural economics, 27*(3), 247--267.

Arentze, T. A., Borgers, A. W., & Timmermans, H. J. (1996). An Efficient Search Strategy for Site-Selection Decisions in an Expert System. *Geographical Analysis, 18*(2), 126--146.

Badri, M. A. (1999). Combining the analytic hierarchy process and goal programming for global facility location-allocation problem. *International journal of production economics, 62*(3), 237--248.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Palgrave Macmillan.

Balch, T. (2000). Hierarchic social entropy: An information theoretic measure of robot group diversity. *Autonomous robots, 8*(3), 209-238.

Balulescu, A. M. (2015). Estimating retail market potential using demographics and spatial analysis for home improvement in Ontario. ON: University of Waterloo.

Batty, M., & Longley, P. (1987). Fractal-based description of urban form. *Environment and planning B: Planning and Design, 14*(2), 123--134.

Baumgartner, H., & Steenkamp, J. B. (2011). Retail Site Selection. *The SAGE Dictionary of Quantitative Management Research, 31*(2), 271.

Bautista, J., & Pereira, J. (2006). Modeling the problem of locating collection areas for urban waste management. An application to the metropolitan area of Barcelona. *Omega, 34*(6), 617--629.

Beamon, B. M. (1998). Supply chain design and analysis:: Models and methods. *International journal of production economics, 55*(3), 281--294.

Benguigui, L., & Daoud, M. (1991). Is the suburban railway system a fractal? *Geographical Analysis, 23*(4), 362-368.

Benoit, D., & Clarke, G. P. (1997). Assessing GIS for retail location planning. *Journal of retailing and consumer services, 4*(4), 239--258.

Black, W. R. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis, 24*(3), 207--222.

Bodvarsson, O. B., & Van den Berg, H. F. (2006). Does immigration affect labor demand? Model and test. In *The Economics of Immigration and Social Diversity* (pp. 135--166). Emerald Group Publishing Limited.

Bodvarsson, O. B., Van den Berg, H. F., & Lewer, J. J. (2008). Measuring immigration's effects on labor demand: A reexamination of the Mariel Boatlift. *Labour Economics, 15*(4), 560--574.

Boyle, A., Barrilleaux, C., & Scheller, D. (2014). Does Walkability Influence Housing Prices? *Social Science Quarterly, 95*(3), 852-867.

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks, 30*(2), 136-145.

Brown, P. A., & Gibson, D. F. (1972). A quantified model for facility site selection-application to a multiplant location problem. *AIIE transactions, 4*(1), 1--10.

Canbolat, Y. B., Chelst, K., & Garg, N. (2007). Combining decision tree and MAUT for selecting a country for a global manufacturing facility. *Omega, 35*(3), 312--325.

Caradima, B. (2015). Multi-criteria suitability analysis and spatial interaction modeling of retail store locations in Ontario, Canada. ON: University of Waterloo.

Ceder, A. (1976). A deterministic traffic flow model for the two-regime approach. *Transportation Research Record, 567*, 16--30.

Chang, N.-B., & Wei, Y. L. (2000). Siting recycling drop-off stations in urban area by genetic algorithm-based fuzzy multiobjective nonlinear integer programming modeling. *Fuzzy Sets and Systems, 114*(1), 133--149.

Charnpratheep, K., Zhou, Q., & Garner, B. (1997). Preliminary landfill site screening using fuzzy geographical information systems. *Waste management & research, 15*(2), 197--215.

Cheng, S., Chan, C. W., & Huang, G. H. (2002). Using multiple criteria decision analysis for supporting decisions of solid waste management. *Journal of Environmental Science and Health, Part A, 37*(6), 975--990.

Chun, Y., & Griffith, D. A. (2011). Modeling network autocorrelation in space--time migration flow data: an eigenvector spatial filtering approach. *Annals of the Association of American Geographers, 101*(3), 523--536.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *32*(1), 101--118.

City of Toronto. (2013, 06 26). *2011 National Household Survey: Backgrounder.* Retrieved 01 2016, 01, from toronto.ca: https://www1.toronto.ca/city_of_toronto/social_development_finance__administration/files/pdf/nhs-backgrounder-labour-education-work-commuting.pdf

Clarke, I., Bennison, D., & Pal, J. (1997). Towards a contemporary perspective of retail location. *International Journal of Retail \& Distribution Management, 25*(2), 59--69.

Clarkson, R. M., Clarke-Hill, C. M., & Robinson, T. (1996). UK supermarket location assessment. *International Journal of Retail \& Distribution Management, 24*(6), 22--33.

Cooper, L. (1964). Heuristic methods for location-allocation problems. *Siam Review, 6*(1), 37--53.

Coughlan, A. T., & Grayson, K. (1998). Network marketing organizations: Compensation plans, retail network growth, and profitability. *International Journal of Research in Marketing, 15*(5), 401--426.

De Montis, A., Barthélemy, M., Chessa, A., & Vespignani, A. (2007). The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design, 34*(5), 905-924.

Dorigo, M., Maniezzo, V., & Colorni, A. (1991). The ant system: An autocatalytic optimizing process. Technical report.

Dunphy, R., & Fisher, K. (1996). Transportation, congestion, and density: new insights. *Transportation Research Record: Journal of the Transportation Research Board*, 89--96. doi:http://dx.doi.org/10.3141/1552-12

Duthie, J., Brady, J., Mills, A., & Machemehl, R. (2010). Effects of on-street bicycle facility configuration on bicyclist and motorist behavior. *Transportation Research Record: Journal of the Transportation Research Board, 2190*, 37--44.

Engelson, L., & van Amelsfort, D. (2011). The role of volume-delay functions in forecast and evaluation of congestion charging schemes, application to Stockholm. *European Transport Conference 2011.*

Evans, J. R. (2011). Retailing in perspective: the past is a prologue to the future. *The International Review of Retail, Distribution and Consumer Research, 21*(1), 1--31.

Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of mathematical sociology, 23*(3), 181--201.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92--107.

Ford, A. C., Barr, S. L., Dawson, R. J., & James, P. (2015). Transport accessibility analysis using GIS: Assessing sustainable transport in London. *ISPRS International Journal of Geo-Information, 4*(1), 124-149.

Fowler, K. (2016). Exploring the use of managerial intuition in retail site selection. *The Service Industries Journal, 36*, 183--199.

Frost, M. E., & Spence, N. A. (1995). The rediscovery of accessibility and economic potential: the critical issue of self-potential. *Environment and Planning A, 27*(11), 1833--1848.

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environment and Planning B: Planning and Design, 40*(1), 135--153.

Gastner, M. T., & Newman, M. E. (2006). The spatial structure of networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 49*(2), 247-252.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & operations research, 13*(5), 533--549.

Goh, K. I., Kahng, B., & Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Physical Review Letters, 87*(27), 278701.

Goodchild, M. F. (1984). LACS: A Location-Allocation Mode for Retail Site Selection. *Journal of Retailing, 60*, 84--100.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate behavioral research, 26*(3), 499--510.

Greenberg, H. (1959). An analysis of traffic flow. *Operations research, 7*(1), 79--85.

Greenhut, M. L. (1956). Plant location in theory and in practice; the economics of space.

Gutierrez, J. (2001). Location, economic potential and daily accessibility: an analysis of the accessibility impact of the high-speed line Madrid--Barcelona--French border. *Journal of transport geography, 9*(4), 229--242.

Gutierrez, J., Gonzalez, R., & Gomez, G. (1996). The European high-speed train network: predicted effects on accessibility patterns. *Journal of transport geography, 4*(4), 227--238.

Hadiuzzaman, M., Qiu, T. Z., & Lu, X. (2012). Variable speed limit control design for relieving congestion caused by active bottlenecks. *Journal of Transportation Engineering, 139*(4), 358--370.

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research, 12*(3), 450--459.

Hansen, P. (1986). The steepest ascent mildest descent heuristic for combinatorial programming. *Congress on numerical methods in combinatorial optimization, Capri, Italy*, (pp. 70--145).

Harri, J., Filali, F., Bonnet, C., & Fiore, M. (2006). VanetMobiSim: generating realistic mobility patterns for VANETs. *Proceedings of the 3rd international workshop on Vehicular ad hoc networks* (pp. 96-97). ACM.

Harris, R. J. (2001). *A primer of multivariate statistics.* Psychology Press.

Hebb, D. O. (1949). *The organization of behavior.*

Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma, 120*(1), 75--93.

Hernandez, T., & Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail \& Distribution Management, 28*(8), 357--367.

Hillsman, E. L. (1984). The p-median structure as a unified linear model for location—allocation analysis. *Environment and Planning A, 16*(3), 305--318.

Hoffman, J. J., & Schniederjans, M. J. (1994). A two-stage model for structuring global facility site selection decisions: the case of the brewing industry. *International Journal of Operations & Production Management, 14*(4), 79--96.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* MIT press.

Holme, P. (2003). Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems, 6*(02), 163--176.

Hong, S.-K., Song, I.-J., & Wu, J. (2007). Fengshui theory in urban landscape planning. *Urban ecosystems, 10*(3), 221--237.

Hoover, E. M. (1967). Some programmed models of industry location. *Land Economics, 43*(3), 303--311.

Huff, D. L. (2003). Parameter estimation in the Huff model. *ESRI, ArcUser*, 34--36.

Hull, A., Silva, C., & Bertolini, L. (2012). *Accessibility instruments for planning practice.* Porto: COST Office Porto.

Iori, G., De Masi, G., Precup, O., Gabbi, G., & Caldarelli, G. (2008). A network analysis of the Italian overnight money market. *Journal of Economic Dynamics and Control, 32*(1), 259-278.

Jensen, P. (2006). Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E, 74*(3), 035101.

Kao, J. J., & Lin, H. Y. (1996). Multifactor spatial analysis for landfill siting. *Journal of environmental Engineering, 122*(10), 902--908.

Kapferer, J.-N. (2012). *The new strategic brand management: Advanced insights and strategic thinking.* Kogan page publishers.

Kazerani, A., & Winter, S. (2009). Can betweenness centrality explain traffic flow. *Proceedings of the 12th AGILE International Conference on GIS.*

Krarup, J., & Vajda, S. (1997). On Torricelli's geometrical solution to a problem of Fermat. *IMA Journal of Management Mathematics, 8*(3), 215--224.

Levy, M., Weitz, A. B., & Grewal, D. (1998). *Retailing management.* Irwin/McGraw-Hill New York.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography, 32*(2), 591--600.

Liang, G.-S., & Wang, M.-J. (1991). A fuzzy multi-criteria decision-making method for facility site selection. *The International Journal of Production Research, 29*(11), 2313--2330.

Litman, T. (2003). Measuring transportation: traffic, mobility and accessibility. *73*(10), 28.

Love, R. F., Morris, J. G., & Wesolowsky, G. O. (1988). *Facilities location: models and methods.* New York: North-Holland.

Lu, Y., & Tang, J. (2004). Fractal dimension of a transportation network and its relationship with urban growth: a study of the Dallas-Fort Worth area. *Environment and Planning B, 31*(6), 895-912.

MacCarthy, B. L., & Atthirawong, W. (2003). Factors affecting location decisions in international operations--a Delphi study. *International Journal of Operations & Production Management, 23*(7), 794--818.

Mahler, C. F., & De Lima, G. (2003). Applying value analysis and fuzzy logic to select areas for installing waste fills. *Environmental monitoring and assessment, 84*(1-2), 129--140.

Mahmassani, H. S., & Chang, G.-L. (1986). Experiments with departure time choice dynamics of urban commuters. *Transportation Research Part B: Methodological, 20*(4), 297--320.

Maitra, S., & Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models, 79*.

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science, 156*(3775), 636-638.

Marshall, S. (2005). *Streets and Patterns.* London: Institute of Community Studies.

Martin, T. E., & Roper, J. J. (1988). Nest predation and nest-site selection of a western population of the Hermit Thrush. *Condor*, 51--57.

May, A. D. (1990). *Traffic flow fundamentals.* Prentice Hall.

McGuirt, J. T., Pitts, S. B., Ward, R., Crawford, T. W., Keyserling, T. C., & Ammerman, A. S. (2014). Examining the influence of price and accessibility on willingness to shop at farmers' markets among low-income eastern North Carolina women. *Journal of nutrition education and behavior, 46*(1), 26--33.

McNally, M. G., & Ryan, S. (1992). A comparative assessment of travel characteristics for neo-traditional developments. *University of California Transportation Center*.

Meade, L., & Sarkis, J. (1998). Strategic analysis of logistics and supply chain management systems using the analytical network process. *Transportation Research Part E: Logistics and Transportation Review, 34*(3), 201--215.

Meltzer, R., & Schuetz, J. (2012). Bodegas or bagel shops? Neighborhood differences in retail and household services. *Economic Development Quarterly, 26*(1), 73--94.

Mendes, A. B., & Themido, I. H. (2004). Multi-outlet retail site location assessment. *International Transactions in Operational Research, 11*(1), 1--18.

Minderhoud, M. M., & Bovy, P. H. (2001). Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention, 33*(1), 89--97.

Minsky, M., & Papert, S. (1969). *Perceptrons.* MIT press.

Möller, D. P., & Schroer, B. (2014). *Introduction to Transportation Analysis, Modeling and Simulation: Computational Foundations and Multimodal Applications.* Springer Publishing Company, Incorporated.

Natural Resources Canada;. (2012). *National Road Network Feature Catalogue Segmented View Edition 2.0.1.* Natural Resources Canada , Earth Sciences Sector Centre for Topographic Information . Retrieved from http://www.geobase.ca

Nema, A. K., & Gupta, S. K. (1999). Optimization of regional hazardous waste management systems: an improved formulation. *Waste Management, 19*(7), 441--451.

Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E, 64*(1), 016132.

Newman, M. E. (2010). *Networks: an introduction.* Oxford University Press.

Nzouonta, J., Rajgure, N., Wang, G., & Borcea, C. (2009). VANET routing on city roads using real-time vehicular traffic information. *IEEE Transactions on Vehicular Technology, 58*(7), 3609--3626.

O'Malley, L., Patterson, M., & Evans, M. (1997). Retailer use of geodemographic and other data sources: an empirical investigation. *International Journal of Retail \& Distribution Management, 25*(6), 188--196.

Okabe, A., & Sugihara, K. (2012). *Spatial analysis along networks: statistical and computational methods.* John Wiley & Sons.

Öner, Ö. (2015). Retail City: The Relationship between Place Attractiveness and Accessibility to Shops. *Available at SSRN 2549626*.

Onut, S., Efendigil, T., & Kara, S. S. (2010). A combined fuzzy MCDM approach for selecting shopping center site: An example from Istanbul, Turkey. *Expert Systems with Applications, 37*(3), 1973--1980.

Páez, A., Mercado, R. G., Farber, S., Morency, C., & Roorda, M. (2010). Relative accessibility deprivation indicators for urban settings: definitions and application to food deserts in Montreal. *Urban Studies, 47*(7), 1415-1438.

Piorkowski, M., Raya, M., Lugo, L. A., Papadimitratos, P., Grossglauser, M., & Hubaux, P. J. (2008). TraNS: realistic joint traffic and network simulator for VANETs. *ACM SIGMOBILE mobile computing and communications review, 12*(1), 31--33.

Pipes, L. A. (1967). Car following models and the fundamental diagram of road traffic. *1*(1), 21--29.

Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A., . . . Scellato, S. (2009). Street centrality and densities of retail and services in Bologna, Italy. *Environment and Planning B: Planning and design, 36*(3), 450--465.

Qi, Q. (2015). *商务部：“互联网+”时代 零售业仍大有前途 [Ministry of Commerce: In the era of "Internet Plus", retail sector is still promissing].* China Industrial Economy News.

Ramu, N. V., & Kennedy, W. J. (1994). Heuristic algorithm to locate solid-waste disposal site. *Journal of urban planning and development, 120*(1), 14--21.

Rangone, A. (1996). An analytical hierarchy process framework for comparing the overall performance of manufacturing departments. *International Journal of Operations \& Production Management, 16*(8), 104--119.

Robins, G. (2011). Exponential random graph models for social networks. *Handbook of Social Network Analysis. Sage*, Citeseer.

Robins, G. (2015). *Doing social network research: Network-based research design for social scientists.* Sage.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks, 29*(2), 173--191.

Rodin, V., & Rodina, E. (2000). The fractal dimension of Tokyo's streets. *Fractals, 8*(4), 413--418.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Sequential thought processes in PDP models. *Parallel distributed processing: explorations in the microstructures of cognition, 2*, 3--57.

Sam, D., & Raj, C. V. (2014). A Time synchronized hybrid vehicular Ad hoc network of roadside sensors and Vehicles for safe driving. *Journal of Computer Science, 10*(9), 1617--1627.

Sarkis, J., & Sundarraj, R. P. (2002). Hub location at Digital Equipment Corporation: A comprehensive analysis of qualitative and quantitative factors. *European Journal of Operational Research, 137*(2), 336--347.

scikit-learn developers. (2017, 1). *3.1. Cross-validation: evaluating estimator performance*. Retrieved from scikit learn: http://scikit-learn.org/stable/modules/cross_validation.html

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System technical Journal, 27*, 379-423.

Siddiqui, M. Z., Everett, J. W., & Vieux, B. E. (1996). Landfill siting using geographic information systems: a demonstration. *Journal of environmental engineering, 122*(6), 515--523.

Smoyer-Tomic, K. E., Spence, J. C., Raine, K. D., Amrhein, C., Cameron, N., Yasenovskiy, V., . . . Healy, J. (2008). The association between neighborhood socioeconomic status and exposure to supermarkets and fast food outlets. *Health & place, 14*(4), 740--754.

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC pharmacology, 10*(1), 6.

Statistics Canada. (2011). *NHS Profile.* Retrieved from Statistics Canada: https://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/index.cfm?Lang=E

Statistics Canada. (2016, 09 16). *North American Industry Classification System (NAICS) Canada 2012*. Retrieved April 2017, from Stantistics Canada: http://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=118464&CVD=118465&CPV=44-45&CST=01012012&CLV=1&MLV=5

Statistics Canada. (2017). *Table 080-0033 - Retail E-commerce sales, unadjusted, monthly (dollars).* CANSIM (database). Retrieved February 2017

Statistics Canada. (n.d.). *Table 379-0030 - Gross domestic product (GDP) at basic prices, by North American Industry Classification System (NAICS), provinces and territories, annual (dollars).* CANSIM (database). Retrieved 03 01, 2016

Studer, A., Shi, E., Bai, F., & Perring, A. (2009). TACKing together efficient authentication, revocation, and privacy in VANETs. *2009 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks* (pp. 1--9). IEEE.

Su, Q. (2011). The effect of population density, road network density, and congestion on household gasoline consumption in US urban areas. *Energy Economics, 33*(3), 445-452.

Teller, C., & Reutterer, T. (2008). The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services, 15*(3), 127--143.

Thomson, J. M. (1977). *Great Cities and Their Traffic* (Vol. 3). London: Gollancz.

Thünen, v. J. (1826). *Der isolierte Staat.* Beziehung auf Landwirtschaft und Nationalokonomie.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234--240.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research, 19*(6), 1363--1373.

Transportation Services. (2014, 9 25). Average Weekday , AM Peak Hour Traffic Volume, (Most Recent Counts from 2005-2013). Retrieved 2 25, 2017, from https://www1.toronto.ca/City%20Of%20Toronto/Transportation%20Services/Road%20safety/Files/pdf/2013volumemapam.pdf

Transportation Services. (2014, 9 25). Average Weekday , PM Peak Hour Traffic Volume, (Most Recent Counts from 2005-2013). Retrieved 2 25, 2017, from https://www1.toronto.ca/City%20Of%20Toronto/Transportation%20Services/Road%20safety/Files/pdf/2013volumemappm.pdf

Travel Modelling Group. (2015). *User's Guide to Running GTAModel V4.0.* University of Toronto, Faculty of Applied Science and Engineering, Transportation Research Institute.

Underwood, R. T. (1961). Speed, volume, and density relationships. In B. D. Greenshields, H. P. George, N. S. Guerin, M. R. Palmer, & R. T. Underwood, *Quality and Theory of Traffic Flow - A Symposium* (pp. 141-188). New Haven Bureau of Highway Traffic, Yale University .

Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures? *Connections, 28*(1), 16.

van Aerde, M., & Rakha, H. (1995). Multivariate calibration of single regime speed-flow-density relationships. *Proceedings of the 6th 1995 Vehicle Navigation and Information Systems Conference*, *334*, p. 341.

VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43--50.

Verhoef, E. (2010). *The economics of traffic congestion.* Edward Elgar Publishing.

Vickerman, R., Spiekermann, K., & Wegener, M. (1999). Accessibility and economic development in Europe. *Regional studies, 33*(1), 1--15.

Wagner, S. M., & Neshat, N. (2010). Assessing the vulnerability of supply chains using graph theory. *International Journal of Production Economics, 126*(1), 121--129.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Weber, A. (1909). *Ueber den standort der industrien.* Рипол Классик.

Weber, C. A., Current, J. R., & Benton, W. C. (1991). Vendor selection criteria and methods. *European journal of operational research, 50*(1), 2--18.

Wheaton, W. C. (1998). Land use and density in cities with congestion. *Journal of urban economics, 43*(2), 258--272.

Wilson, J. D. (1983). Optimal road capacity in the presence of unpriced congestion. *Journal of Urban Economics, 13*(3), 337--357.

Wood, S., & Reynolds, J. (2012). Leveraging locational insights within retail store development? Assessing the use of location planners' knowledge in retail marketing. *Geoforum, 43*(6), 1076--1087.

Xie, F., & Levinson, D. (2007). Measuring the structure of road networks. *Geographical analysis, 39*(3), 336-356.

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems, 32*(5), 396--406.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control, 8*(3), 338--353.

Appendix A – Descriptions of site suitability criteria

Table A-1 development of site suitability criteria

Criteria Category	Criteria Name	Criteria Definition and Calculation
Site variable	site maximum slope	The maximum value of the parcel's slope.
Traffic and transportation variables	traffic visibility	Visibility is correlated with distance from the major highways and the traffic volume. $$S_i = \left(1 - \frac{D_i}{D_{max}}\right) * \frac{T_i}{T_{max}}$$ S_i: the suitability of parcel I; D_i: the distance of parcel i to the nearest highway; D_{max}: the distance threshold of visibility; T_i: traffic volume of the adjacent highway; T_{max}: the highest traffic volume in the census division.
	highway accessibility	Travel time from a parcel to the nearest highway access point.
	distance to distribution centre	The network distance to the nearest distribution centre.
Market variables	market representation	Location quotient of a dissemination area. $$Location\ Quotient = \frac{c/C}{r/R}$$ c: the number of NAICS 444 retailers in a DA's trade area; C: the number of all retailers in the DA's trade area; r: the number of NAICS 444 retailors in Ontario; R: the number of all retailers in Ontario.
	density of competitors	The number of competitors per unit area in the trade area.
	density of retail stores	The number of retailers per unit area in the trade area.
	Potential expenditures	Estimated expenditure without competitors using Huff's model[18].
	Competitive expenditures	Estimated expenditure with competitors using Huff's model.

Note: Adapted from Caradima (2015)

[18] For details about Huff's model see Caradima (2015).

Appendix B – Full correlation table

Table B-1 Correlations among the full list of variables

**** P < 0.01; ** P < 0.05; * P < 0.1*

Group	Variable	Sales	ETP	CC_{avg}	CC_{std}	NCC_{sum}	NCC_{avg}	BC	NLC	DC_1	DC_2	Imm	D_V	D_O	S	D_C	Inc	b	v	r	d	l	d_c	d_r	e_c
Network	ETP	-0.43 **																							
	CC_{avg}	0.216	-0.29																						
	CC_{std}	-0.24	-0.31	0.248																					
	NCC_{sum}	-0.17	-0.07	0.677 ***	0.248																				
	NCC_{avg}	0.247	-0.33	0.994 ***	0.251	0.665 ***																			
	BC	0.123	0.088	-0.01	-0.24	-0.22	0.016																		
	NLC	0.254	0.036	-0.16	-0.17	-0.24	-0.12	0.805 ***																	
	DC_1	-0.11	-0.14	0.717 ***	0.115	0.383 *	0.73 ***	0.198	0.113																
	DC_2	0.016	-0.02	0.128	-0.24	-0.26	0.165	0.736 ***	0.552 ***	0.489 **															
Demographic	Imm	-0.17	-0.04	0.711 ***	0.183	0.798 ***	0.696 ***	-0.22	-0.41 **	0.472 **	-0.23														
	D_V	0.07	-0.23	0.628 ***	0.339 *	0.728 ***	0.624 ***	-0.32	-0.39 *	0.225	-0.42 **	0.854 ***													
	D_O	-0.15	-0.05	0.735 ***	0.175	0.817 ***	0.719 ***	-0.25	-0.43 **	0.489 **	-0.26	0.992 ***	0.849 ***												
	S	0.093	0.101	-0.22	-0.08	-0.2	-0.22	-0.28	-0.04	-0.2	-0.13	-0.14	-0.05	-0.16											
	D_C	-0.14	-0.05	0.739 ***	0.083	0.814 ***	0.724 ***	-0.21	-0.38 *	0.53 ***	-0.19	0.98 ***	0.826 ***	0.989 ***	-0.18										
	Inc	-0.15	-0.07	0.739 ***	0.191	0.821 ***	0.723 ***	-0.25	-0.41 **	0.493 **	-0.25	0.987 ***	0.871 ***	0.995 ***	-0.14	0.989 ***									
Suitability	b	-0.11	0.379 *	-0.12	-0.29	-0.01	-0.09	0.421 **	0.323	0.098	0.287	0.042	-0.05	0.024	-0.49 **	0.05	0.019								
	v	-0.34 *	0.614 ***	-0.16	-0.32	0.026	-0.2	-0.07	-0.1	-0.06	-0.02	0.025	-0.2	0.025	0.196	0.051	0.026	0.042							
	r	0.101	-0.51 ***	-0.02	0.174	0.011	-0.01	0.219	0.359 *	0.012	0.087	-0.13	-0.04	-0.13	-0.22	-0.11	-0.12	0.032	-0.43 **						
	d	-0.12	0.004	-0.27	0.032	-0.47 **	-0.24	0.483 **	0.595 ***	0.25	0.644 ***	-0.53 ***	-0.48 **	-0.54 ***	0.119	-0.49 **	-0.5 **	0.141	-0.09	0.156					
	l	-0.03	0.141	-0.75 ***	-0.3	-0.7 ***	-0.76 ***	0.194	0.239	-0.53 ***	0.176	-0.85 ***	-0.77 ***	-0.85 ***	0.116	-0.83 ***	-0.83 ***	0.022	0.062	-0.01	0.424 **				
	d_c	-0.03	-0.08	0.692 ***	-0.05	0.656 ***	0.682 ***	-0.02	-0.29	0.479 **	-0.03	0.837 ***	0.707 ***	0.866 ***	-0.26	0.893 ***	0.867 ***	0.085	-0.05	-0.05	-0.4 **	-0.7 ***			
	d_r	-0.05	-0.07	0.689 ***	-0.07	0.699 ***	0.686 ***	-0.12	-0.31	0.492 **	-0.08	0.896 ***	0.768 ***	0.9 ***	-0.18	0.938 ***	0.9 ***	0.055	-0.02	-0.05	-0.43 **	-0.78 ***	0.948 ***		
	e_c	-0.11	-0.07	0.727 ***	0.081	0.82 ***	0.716 ***	-0.22	-0.38 *	0.502 ***	-0.19	0.968 ***	0.821 ***	0.979 ***	-0.14	0.988 ***	0.979 ***	0.044	0.021	-0.1	-0.49 **	-0.81 ***	0.905 ***	0.938 ***	
	e_p	0.036	-0.16	0.753 ***	0.113	0.809 ***	0.751 ***	-0.36 *	-0.43 **	0.457 **	-0.28	0.894 ***	0.824 ***	0.917 ***	-0.17	0.935 ***	0.924 ***	-0.02	-0	-0.09	-0.52 ***	-0.83 ***	0.825 ***	0.884 ***	0.939 ***

Note: high correlations (correlation coefficient > 0.8 at significant level of 0.01) are shaded.

Appendix C – Model selection and model details

Table C-1 MSE of network models in cross-validation
Reported in squared million dollars

Number of Variables	Network L2O Best Groups		Network 10-Fold Best Groups		Network L2O Worst Groups		Network 10-Fold Worst Groups	
	variables	MSE	variables	MSE	variables	MSE	variables	MSE
1	ETP	41.52	ETP	46.07	NCC_sum	52.38	NCC_sum	55.11
2	ETP, CC_std	36.52	NCC_avg, DC1	42.17	DC1, DC2	56.21	NCC_sum, DC1	58.45
3	NCC_sum, NCC_avg, DC1	28.45	NCC_sum, NCC_avg, DC1	29.45	BC, DC1, DC2	66.17	NCC_sum, DC1, DC2	61.64
4	CC_std, NCC_sum, NCC_avg,	24.61	NCC_sum, NCC_avg, NLC,	24.22	NCC_sum, BC, DC1, DC2	70.97	NCC_sum, BC, DC1, DC2	67.87
5	ETP, CC_std, NCC_sum, NCC_avg, DC1	20.49	ETP, CC_std, NCC_sum, NCC_avg, DC1	21.27	CC_std, NCC_sum, BC, DC1, DC2	77.85	CC_std, NCC_sum, BC, DC1, DC2	69.97
6	CC_std, NCC_sum, NCC_avg, BC, NLC, DC1	14.53	CC_std, NCC_sum, NCC_avg, BC, NLC, DC1	14.47	CC_std, NCC_sum, BC, NLC, DC1, DC2	82.35	CC_std, NCC_sum, BC, NLC, DC1, DC2	71.32
7	ETP, CC_std, NCC_sum, NCC_avg, NLC, DC1, DC2	10.31	ETP, CC_std, NCC_sum, NCC_avg, NLC, DC1, DC2	10.62	ETP, CC_std, NCC_sum, BC, NLC, DC1, DC2	55.16	ETP, CC1, NCC_sum, NCC_avg, BC, DC1, DC2	50.14
8	ETP, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	7.50	ETP, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	7.34	ETP, CC1, CC_std, NCC_avg, BC, NLC, DC1, DC2	34.00	ETP, CC1, CC_std, NCC_avg, BC, NLC, DC1, DC2	32.61
9	ETP, CC1, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	6.62	ETP, CC1, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	7.19	ETP, CC1, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	6.62	ETP, CC1, CC_std, NCC_sum, NCC_avg, BC, NLC, DC1, DC2	7.19

Table C-2 MSE of demographic models in cross-validation
Reported in squared million dollars

Number of Variables	Demographic L2O Best Groups		Demographic 10-Fold Best Groups		Demographic L2O Worst Groups		Demographic 10-Fold Worst Groups	
	variables	MSE	variables	MSE	variables	MSE	variables	MSE
1	S	50.61	Imm,	51.65	DC	52.13	DC	52.62
2	Imm, DV	44.07	Imm, DV	45.51	Imm, DO	59.81	Imm, DO	58.40
3	DV, DC, Inc	41.95	DV, DC, Inc	41.84	Imm, DO, Inc	65.33	Imm, DO, Inc	63.42
4	DV, S, DC, Inc	45.02	DV, DO, DC, Inc	43.41	Imm, DO, S, Inc	69.93	Imm, DO, S, Inc	65.28
5	Imm, DV, DO, DC, Inc	48.14	Imm, DV, DO, DC, Inc	41.80	Imm, DO, S, DC, Inc	73.72	Imm, DO, S, DC, Inc	67.20
6	Imm, DV, DO, S, DC, Inc	53.13	Imm, DV, DO, S, DC, Inc	45.90	Imm, DV, DO, S, DC, Inc	53.13	Imm, DV, DO, S, DC, Inc	45.90

Table C-3 MSE of suitability models in cross-validation
Reported in squared million dollars

Number of Variables	Suitability L2O Best Groups		Suitability 10-Fold Best Groups		Suitability L2O Worst Groups		Suitability 10-Fold Worst Groups	
	variables	MSE	variables	MSE	variables	MSE	variables	MSE
1	v	48.30	d	50.80	ep	51.74	v	62.64
2	ep, ec	46.01	ep, ec	46.93	dc, dr	57.56	v, l	67.43
3	v, ep, ec	45.18	l, ep, ec	47.07	dc, dr, ec	62.43	v, l, dc	71.54
4	v, dc, ep, ec	45.53	l, dr, ep, ec	47.67	b, dc, dr, ec	69.44	b, v, l, dc	77.15
5	v, d, dr, ep, ec	47.54	d, l, dr, ep, ec	48.96	b, l, dc, dr, ec	73.92	b, v, l, dc, ec	82.62
6	v, r, d, dc, ep, ec	51.43	r, d, l, dr, ep, ec	50.19	b, d, l, dc, dr, ec	78.93	b, v, l, dc, dr, ec	87.27
7	v, r, d, l, dc, ep, ec	55.80	v, r, d, l, dc, ep, ec	54.56	b, r, d, l, dc, dr, ec	83.54	b, v, r, l, dc, dr, ec	91.54
8	v, r, d, l, dc, dr, ep, ec	60.66	v, r, d, l, dc, dr, ep, ec	57.35	b, v, r, d, l, dc, dr, ec	85.51	b, v, r, d, l, dc, dr, ec	92.49
9	b, v, r, d, l, dc, dr, ep, ec	69.29	b, v, r, d, l, dc, dr, ep, ec	65.21	b, v, r, d, l, dc, dr, ep, ec	69.29	b, v, r, d, l, dc, dr, ep, ec	65.21

Table C-4 PLS loading table

Variable		Model and Components																			
		PLS-N		PLS-D		PLS-S		PLS-ND				PLS-NS				PLS-DS		PLS-NDS			
		Comp1	Comp2	Comp1	Comp2	Comp1	Comp2	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp1	Comp2	Comp3	Comp4
Response	*Sales*	0.6700	0.0902	0.4625	0.4189	0.2563	0.7776	0.7138	0.2524	0.2226	0.2969	0.7033	0.1602	0.7567	0.1113	0.2800	0.5601	0.6866	0.1376	0.2741	0.4551
Network	*ETP*	-0.9834	0.4842					-0.9427	0.0854	0.6813	-0.2864	-0.9674	0.4700	0.5187	-0.2719			-0.8558	-0.2121	0.6771	-0.1173
	CC_{std}	-0.2882	-0.8750					-0.2886	-0.4893	-0.8424	0.7591	-0.2999	-0.9628	-0.1793	0.3304			-0.2763	-0.0236	-1.0755	0.7019
Demography	*Imm*			-1.4985	0.3779			-0.3600	0.8018	-0.8204	-0.0998					-1.0921	0.2454	-0.5142	0.5314	-0.2293	-0.2232
	D_V			-1.0250	0.9258			-0.1507	0.8262	-0.6887	0.5760					-0.8079	0.7770	-0.2893	0.5402	-0.2352	0.6062
Suitability	e_c					-1.4465	0.2973					-0.1872	0.3477	-1.9167	0.5578	-1.0855	0.2769	-0.4168	0.5616	-0.1458	-0.3315
	e_p					-1.2820	0.9548					-0.0837	0.3544	-1.0668	0.7112	-0.9593	0.5415	-0.2976	0.5772	-0.1197	-0.0537

Table C-5 Mathematical models

Model	Solution
MM-N	$Sales = 59949021 + 3575175.74850705 \times ETP^2 \times cos(7569808.40234965 \times ETP) - 16006445.6002844 \times ETP - 2003466882757.61 \times CC_{std} - 3575175.52110255 \times cos(7676585.41034939 \times ETP)$
MM-D	$Sales = 22834008.3445785 + 27.8594262541033 \times D_V + 3463461.71205782 \times cos(2.05360787074268 \times D_V) + 6406019.97129763 \times cos(cos(4.56225343867134 - 2.05360793654371 \times D_V) - 2.25983877592123 \times D_V) - 5.94576071728043 \times Imm$
MM-S	$Sales = 47042164 + 1.34651074261852 \times 10^{-7} \times {e_p}^2 + 0.399358756840768 \times e_p \times sin(4.67160825160463 + 6.24910901352305 \times 10^{-12} \times {e_c}^2) - 2.68720110076568 \times e_p - 4.1233865982514e-12 \times {e_c}^2 - 12260269.8078034 \times sin(4.67160825160463 + 6.24910901352305 \times 10^{-12} \times {e_c}^2)$
MM-ND	$Sales = 98810157 + 18.4712932990908 \times D_V \times ETP^3 - \frac{491820}{sin\left(sin(cos(0.273486737758484 - 18.1794559208219 \times ETP^2))\right)} - 48771421.1786275 \times ETP - 5343228102286.45 \times CC_{std} - 1.09731557757784 \times 10^{-5} \times Imm^2$
MM-NS	$Sales = \frac{41593617}{ETP} + 3.95138206990593 \times e_p \times ETP + \frac{571637654447161 + 325352 \times e_c}{e_p \times ETP} - 87125638.0497329 - 5765625660388.86 \times CC_{std} - 8.18332187565871 \times 10^{-10} \times e_p \times e_c \times ETP - 3.95138206990593 \times ETP \times e^{3.95138206990593 \times ETP^2}$
MM-DS	$Sales = 11520505 + 19.1512596716705 \times D_V + 19441405.8498844 \times sin(e^{sin(sin(0.255963307494653 + 0.255963307494653 \times Imm))}) - 3.19091193801937 \times Imm - 5617430.05848272 \times sin(6.06400079268102 + cos(D_V) - 0.249993748074288 \times Imm)$
MM-NDS	$Sales = 56417606 + e_p + 3879411089253.75 \times ETP \times CC_{std} + 9.18861035767154 \times ETP \times \frac{cos(0.0916227305193614 \times Imm)}{CC2} - 0.00557917257642006 \times e_c - 22626353.0373718 \times ETP - 6525672363595.11 \times CC_{std}$

Appendix D – Exploration of variable reduction methods

In Chapter Three, models were selected via cross-validation based on MSE values. However, there turned out to be high correlations among the selected variables. Variance inflation factor (VIF) is an indicator of collinearity among predictors in a model. Therefore, model selection based on VIF may help to reduce collinearity problem. The following sections present three model selection approaches used the unselected variables presented in Chapter Three section 2.2. These approaches were not adapted in this thesis for the resulting models had poor performances.

1. VIF-based backward selection

This model selection approach uses a backward selection of predictors by greedily removes predictor with the highest VIF in each step until only two predictors are left in an OLS model. Whereas the resulting models (Table D-1 to D-3) had lower adjusted R-squared values compared to the OLS models used in Chapter Three. Especially the models based on demographic and suitability variables had almost zero explanatory power in sales modeling. Therefore, this approach was not adapted in this thesis.

Table D-1 Network variable VIF in a backward variable selection approach
*Variables selected in this thesis are in **bold***

Predictor	I	II	III	IV	V	VI	VII	VIII
ETP	1.48	1.32	1.26	1.13	1.12	1.12	1.02	1
CC_{avg}	128.97	4.38	4.03					
$\boldsymbol{CC_{std}}$	1.27	1.27	1.27	1.27	1.19	1.15		
NCC_{sum}	2.66	2.6	2.6	1.71	1.34			
NCC_{avg}	133.15							
BC	5.46	5.22						
NLC	3.55	3.37	1.59	1.51	1.14	1.05	1.02	1
DC_1	3.66	3.66	3.18	2.26	1.26	1.04	1.03	
DC_2	4.75	4.28	2.79	2.76				
R-Sq	0.94	0.87	0.84	0.46	0.41	0.41	0.30	0.26
R-Sq(adj)	0.90	0.81	0.78	0.29	0.27	0.30	0.20	0.19

Table D-2 Demographic variable VIF in a backward variable selection approach
*Variables selected in this thesis are in **bold***

Predictor	I	II	III	IV	V
Imm	68.92	39.56	30		
$\boldsymbol{D_V}$	6.37	5.53	3.78	3.23	1
D_O	228.53				
S	1.1	1.08	1.07	1.06	1
D_C	67.04	67.04	26.04	3.32	
Inc	222.93	112.65			
R-Sq	0.39	0.30	0.25	0.12	0.01
R-Sq(adj)	0.20	0.13	0.11	0.01	0.00

Table D-3 Suitability variable VIF in a backward variable selection approach
*Variables selected in this thesis are in **bold***

Predictor	I	II	III	IV	V	VI	VII
b	1.08	1.07	1.07	1.06	1.02	1.02	1
v	1.27	1.26	1.24	1.24	1.23	1.01	1
r	1.28	1.27	1.27	1.26	1.25	1.03	
d	1.44	1.44	1.43	1.34	1.05		
l	3.58	3.48	3.26	2.07			
d_c	11.55	10.51	3.27	2.07			
d_r	16.59	15.62					
$\boldsymbol{e_c}$	17.64						
$\boldsymbol{e_p}$	10.32	6.61	5.74				
R-Sq	0.35	0.18	0.16	0.16	0.14	0.14	0.12
R-Sq(adj)	0.00	0.00	0.00	0.00	0.00	0.03	0.05

2. The best models in cross-validation

Cross-validation provided a list of variable combinations that generated low MSE in OLS models. However, there was a trade-off between the VIF of variables and the fitness of the models. The selected variables combination might not have the lowest VIF values among the best 10 groups in each cross-validation but they all generated the highest adjusted R-squared values and the lowest SSE. In contrast, the models with lower VIF values have poor performance regarding fitness and information loss (Table D-4 to D-6). Therefore, they were not used in this thesis.

Table D-4 The top 10 groups of network variables in cross-validation, sorted by VIF
*Variables selected in this thesis are in **bold***

Predictor	VIF	R-Sq	R-Sq(adj)	SSE(sqr million)	10-fold MSE(sqr million)
ETP, NLC	1.00	0.26	0.19	832.56	-45.26
ETP, BC	1.01	0.21	0.14	885.06	-45.99
ETP, DC1	1.02	0.21	0.15	882.95	-47.37
NCC_avg, NLC	1.02	0.14	0.07	963.56	-47.06
CC_avg, NLC	1.02	0.13	0.06	976.56	-47.76
ETP, CC_std	**1.11**	**0.34**	**0.28**	**741.12**	**-42.67**
NCC_sum,NCC_avg	1.79	0.27	0.20	825.53	-43.48
CC_avg, NCC_avg	1.85	0.24	0.17	858.85	-44.94
CC_avg, DC1	2.05	0.19	0.11	915.90	-44.52
NCC_avg, DC1	2.14	0.23	0.17	860.44	-42.17

Table D-5 The top 10 groups of demographic variables in cross-validation, sorted by VIF
*Variables selected in this thesis are in **bold***

Predictor	VIF	R-Sq	R-Sq(adj)	SSE(sqr million)	10-fold MSE(sqr million)
Imm, S	1.02	0.03	0.00	1085.30	-52.90
DO, S	1.02	0.03	0.00	1094.02	-53.02
S, Inc	1.02	0.03	0.00	1094.53	-53.17
S, DC	1.03	0.02	0.00	1097.49	-53.71
DV, DC	3.14	0.12	0.05	984.42	-52.40
DV, DO	3.59	0.16	0.09	945.53	-49.04
Imm, DV	**3.69**	**0.20**	**0.14**	**894.45**	**-45.51**
DV, Inc	4.13	0.18	0.11	921.67	-47.58
DO, DC	44.11	0.03	0.00	1095.86	-53.47
DC, Inc	47.59	0.02	0.00	1098.43	-53.43

Table D-6 The top 10 groups of suitability variables in cross-validation, sorted by VIF
*Variables selected in this thesis are in **bold***

Predictor	VIF	R-Sq	R-Sq(adj)	SSE(sqr million)	10-fold MSE(sqr million)
b, d	1.02	0.02	0.00	1097.39	-53.06
r, d	1.03	0.03	0.00	1091.48	-52.59
d, dr	1.22	0.03	0.00	1093.70	-52.68
d, l	1.22	0.02	0.00	1106.94	-53.25
d, ec	1.31	0.05	0.00	1063.68	-52.24
d, dc	1.31	0.06	0.00	1051.42	-53.36
d, ep	1.37	0.02	0.00	1106.78	-52.44
l, ec	2.91	0.05	0.00	1064.11	-51.29
dr, ec	8.29	0.04	0.00	1078.99	-52.21
ec, ep	**8.40**	**0.18**	**0.11**	**916.43**	**-46.93**

3. Dimension reduction

PCA and PLS are two commonly adapted dimension reduction methods when collinearity exists. A major difference between PCA and PLS is that PCA only captures the correlation among predictive variables while PLS includes relations among both predictive and target variables (Maitra & Yan, 2008).

In this case study, the results of PCA and PLS were compared by the number of principle components and the goodness of fit (Table D-7). The objective of component reduction in PCA was set to two. The PCA method produced three models with either low goodness of fit or extremely high VIF. Among the models produced by PLS method, the network model had high goodness of fit but also a large number of components; the demographic and suitability models had comparable fitness with the models adapted in this thesis but did not show advantages regarding the number of components. Although the results of PCA and PLS were not directly comparable because of the variance in the number of components, both of them showed some disadvantages compared to the PLS models adapted in this thesis.

Table D-7 PCA and PLS based on the full list of variables

Group	PCA				PLS		
	PC numbers	VIF	R-sq	R-sq(adj)	PC numbers	R-sq	R-sq(adj)
Network	2	1.96	0.01	0.00	9	0.94	0.90
Demographic	2	2.36	0.05	0.00	4	0.28	0.14
Suitability	2	183521.03	0.18	0.11	2	0.18	0.10