The Journal of Transport and Land Use http://jtlu.org
Vol. 14 No. 1 [2021] pp. 1227–1247

The evolution of choice set formation in dwelling and location with rising prices: A decadal panel analysis in the Greater Toronto Area

Jason Hawkins
University of Toronto
jason.hawkins@mail.utoronto.ca

Khandker Nurul Habib
University of Toronto
khandker.nurulhabib@mail.utoronto.ca

Abstract: Home location choice is based on both the characteristics of the dwelling (e.g., size, style, number of bedrooms) and the location (e.g., proximity to work, quality of schools, accessibility). Recent years have seen a steep increase in the price of housing in many major cities. In this research, we examine how these price increases are affecting the types of dwelling and locations considered by households. A large sample of real estate listings from 2006 and 2016 from the Greater Toronto Area is used to develop the empirical models. Two recently developed discrete choice models are used in the study: a nested logit model with latent class feedback (LCF) and a semi-compensatory independent availability logit (SCIAL) model. A method of alternative aggregation is proposed to overcome the computational hurdle that often impedes the estimation of choice set models. We find a significant increase in the probability of larger households considering townhouses and apartments over detached single-family dwellings between 2006 and 2016.

Article history:
Received: July 22, 2020
Received in revised form: June 18, 2021
Accepted: September 20, 2021
Available online: November 18, 2021

1 Introduction

Throughout the 20th century, urban literature found a pattern of migration by middle-class households from the city to the suburbs (Boustan & Shertzer, 2013; Hortas-Rico, 2015). However, recent trends suggest that many of these households are moving back into the city (Hyra, 2015; Smith, 2019). There is a range of explanations provided for these shifting patterns. One plausible narrative is that as the industrialization of cities decreased urban land values, increasingly affluent Western households purchase detached (single-family) dwellings in low-density communities away from the dirty streets of the urban core, called downtown (Platt Boustan et al., 2013). These changes were facilitated by decreasing travel costs associated with the introduction of the private automobile. Like many Western cities transition towards a service-centered economy, urban land becomes more valuable and encourages upmarket development that entices affluent households back into the city to take advantage of urban amenities - see the work of Florida (2012), Glaeser (2012), and others. In addition to these sociological drivers, the modern city is faced with increasing commuting costs - no improvement in transportation speed and continued expansion of metropolitan regions (Banister, 2005). In addition, the effects of climate change are leading many governments and individuals to reconsider urban design to reduce vehicle kilometers traveled (VKT) and the energy footprint of our homes (Moore et al., 2010) - for example,

http://dx.doi.org/10.5198/jtlu.2021.1854
ISSN: 1938-7849

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

California SB375 (CARB, 2020). We note that the COVID-19 pandemic may influence these patterns, but it is too early to draw definite conclusions about its long-term impact on location choice.

In this study, we focus on how dwelling choice has changed as a function of both the type of dwelling (i.e., detached, semi-detached, townhouse, or apartment) and location within a metropolitan region. We apply discrete choice modeling to examine the trade-offs made by households between dwelling and location attributes. The Greater Toronto Area (GTA) is used as representative of a large Western city for the empirical investigation.

The City of Toronto ranks as the most expensive city in Canada, according to Mercer's Cost of Living Survey (115th globally) (Mercer Canada, 2019). Average commute times in the GTA are comparable at 34.0 minutes with major American cities (New York 34.7 minutes and Washington D.C. 32.8 minutes) (Statistics Canada, 2016). As commuting times rise, the cost of travel begins to outweigh the allure of a larger dwelling. In conjunction with increasing commuting times, the average price of a single-family detached dwelling in the GTA rose from $819,319 in 2019 to $900,000 in 2020 (Kalinowski, 2020). These trends are leading many families to consider townhouses or condominium apartments and developers to respond with family-sized units (Caton, 2016; Hume, 2018; LeBlanc, 2018). Many in the industry predict that the current trend is driven by affordability but that it will continue as households become accustomed to proximity to transit and amenities (Hume, 2018).

We collect a large dataset of residential property sales for the GTA, containing a wide range of its attributes and detailed geolocation to associate dwellings with detailed sociodemographic information and features of the local built form. We contribute to the literature by making two novel applications of discrete choice models. The first model is a latent class with feedbacks (LCF) discrete choice model. This model overcomes the standard nested logit method of estimating the joint decision of dwelling type and location choice by including both potential nesting structures as a function of latent features of the decision-maker. The second model is the semi-compensatory independent availability logit (SCIAL) model. We apply the models in the years 2006 and 2016 to examine how the joint choice of dwelling type and location changed in the GTA over this decade.

The choices of both dwelling type and location have essential implications for transportation and urban policy. The preference for a dwelling type will affect the density of development in a region; the preference for detached dwellings over apartments will lead to a lower population density and affect the viability of some travel modes (e.g., walking, cycling, and transit). Likewise, the choice of location will affect the distribution of travel and infrastructure investments in a region. By applying the models to a panel of data, we can quantify the change in preferences within the GTA within specific socioeconomic groups. We first provide an overview of the existing literature to provide context for our work. We then outline the data and models used in the empirical application. Marginal effect plots are used to show the variation in preferences across the population and over the decade between 2006 and 2016. We then provide a discussion of how our results relate to planning policy.

2 Literature review

There is extensive literature relating to the question of residential location choice. In our review, we focus on models that jointly consider location and dwelling type choice and the methods of analysis.

The first models to consider the dwelling choice as a bundle of individual decisions estimated them as sequential conditional logit models (Boehm, 1982; Brownstone & Englund, 1991; Quigley, 1985). Boehm (1982) considers tenure choice (own or rent), size of dwelling (small or large), and neighborhood quality (low or high). This initial effort at separating the components of the choice bundle suffers from a lack of structural variables to test the hypothesized decision ordering. Quigley (1985) addresses this shortcoming by formulating the decision as a nested logit model, considering tenure, neighborhood, and individual dwelling choice. Brownstone and Englund (1991) represent the choice of tenure and dwelling type as a joint decision and use a multinomial logit model, after testing nested logit structures. They estimate the equilibrium demand for housing in Sweden and do not explicitly consider location choice. In all the above cases, model estimation was constrained by computation power such that only sequential estimation for a small number of alternatives was possible. Tu and Goldfinch (1996) estimate a model with nine dwelling types and seven regions but remain constrained by computing power. They decompose the choice into two stages: the choice of a region based on key dwelling attributes and the choice of a dwelling type within each region based on non-key dwelling attributes.

Bhat et al. (Bhat, 2015; Fu et al., 2015) develop a choice set formation model based on the Generalized Heterogeneous Data Model (GHDM). They begin from the principle that households cannot consider the full set of alternatives in the case of residential choice because the universal choice set may contain tens of thousands of alternatives. The literature suggests that, when faced with a large universe of potential alternatives, decision-makers narrow the choice set by employing heuristic non-compensatory rules. Bhat (2015)estimates a structural model of dwelling size, type, tenure, number of bedrooms, and number of bathrooms as a mechanism for representing this narrowing of the choice set (formation of the consideration set). He then proposes the use of a standard residential choice model to determine the specific land parcel among those matching the first stage criteria. However, this approach would be difficult to apply to the question of the propensity to switch location versus dwelling type because it does not capture such directionalities among choice dimensions.

Both Kaplan et al. (2013; 2012) and Zolfaghari et al. (2013) examine the residential choice set formation problem. Kaplan et al. develop a three-step procedure to develop the choice set. In the first step, an aggregation heuristic is applied to cluster similar location alternatives. These aggregate alternatives are then examined in a compensatory multinomial logit model (forming the second step). The final step is a conjunctive heuristic, composed of naturally ordered criteria and multinomial criteria. Naturally ordered criteria include time and cost constraints imposed on the household by their income and location of work. Multinomial criteria include such factors as apartment sharing and neighborhood criteria. By restricting the choice space to alternatives that satisfy a set of observed thresholds, Kaplan et al. do not consider the universal choice set (a deviation from the general Manski model discussed below). Zolfaghari et al. propose a similar model but make changes such that they maintain the universal choice set. They accomplish this objective by grouping sets of alternatives according to different combinations of thresholds. The number of choice sets then grows exponentially with the number of thresholds and threshold levels rather than the number of alternatives.

Van Eggermond et al. (2018) provide an alternative approach based on explicit inclusion of choice set formation variables elicited in a survey, whereas the above research (and our work) can be seen as using probabilistic inference - i.e., the choices of socio-demographically similar households indicate the choice set. Based on their survey results, they introduce temporal, locational, affordability, and market preferences to their location choice model. They find that proximity to parents and social networks play a significant role in the choice set process in Singapore. It would be interesting to apply such an approach in other cities and examine whether this is a universal pattern or culturally specific. The psychology literature likely has some interesting insights in this regard (Stieger & Lewetz, 2016).

We contribute to this literature by applying two novels discrete choice models, which overcome several limitations of previous work. The first model (LCF) provides a structure for incorporating the influence of sociodemographic factors on the nesting structure of dwelling and location choice. The second model (SCIAL) allows us to focus on the choice set generation component and include the universal choice set. We overcome the computational challenges of the choice set generation stage through the application of a recently proposed alternative aggregation method. Finally, we provide the first panel application of a dwelling choice set generation model. This extension allows us to consider changes in preferences that will be of interest to transportation and land-use forecasters, as well as policymakers.

3 Methods of analysis

We consider the trade-offs made between dwelling choice and location choice through the estimation of two classes of discrete choice model: a latent class with feedbacks (LCF) model and a semi-compensatory independent availability logit (SCIAL) model (Habib, 2019; Hossain et al., 2020). In the first instance (LCF), the joint decision of dwelling and location is formulated as a pair of nested logit models. The nested logit models capture the relative propensity to switch dwelling type and location choice. Rather than estimate a single nesting structure, the LCF formulation allows us to specify both structures, with the probability of each structure describing the decision-making process being a function of latent sociodemographic attributes of the household. In the second instance (SCIAL), the choice of dwelling and location is assumed to be a joint decision by the household. However, the probability of a household considering a combination of dwelling type and location is captured through the estimation of a choice set formation model. In the following sub-sections, we outline the econometric details of the two model structures. Table 1 provides a summary of the nomenclature used to describe the models.

Table 1. Summary of nomenclature

Variable	**Description**
Latent Class with Feedback (LCF) Model	
L_i	Joint likelihood of latent class, dwelling, and location choice for observation i
U_l	Utility for choice of location (V_l = systematic component and ϵ_l = random component)
U_d	Utility for choice of dwelling (V_d = systematic component and ϵ_d = random component)
V_{C1}	Systematic utility for membership in Class1
$Pr(C1)$	Probability of inclusion in latent class 1 (similar for class 2)
$Pr(dl)$	Probability of joint choice of location alternative l and dwelling alternative d
$Pr(l)$	Probability of choosing location alternative l (similar for dwelling alternative d)
$Pr(d\|l)$	Probability of choosing dwelling alternative d given choice of location alternative l (similar for choice of location alternative l given choice of dwelling alternative d)
δ_l	Indicator that location alternative l is chosen (similar for dwelling alternative d)
μ_d	Scale parameter for dwelling alternative d (*similar for location alternative* l)
μ_c	Scale parameter for class membership
$EMU(C1)$	Expected maximum utility for feedback to Class 1 (similar for Class 2)
Semi-Compensatory Independent Availability Logit (SCIAL) Model	
$Pr(j)$	Generic probability of choosing alternative j
$Pr(j \mid C_i)$	Generic probability of choosing alternative j given choice set C_i
$Pr(C_i)$	Generic probability of choice set i
V_j	Generic systematic utility for alternative j (similar for alternatives k and T)

Variable	Description
A_j	Probability alternative *j* is included in the choice set (similar for alternatives *k* and *T*)
A'_j	Probability alternative *j* is not included in the choice set (similar for alternatives *k* and *T*)
Z	Attributes of the decision-maker
γ	Parameters associated with each of the decision-maker attributes included in *Z*
$Pr(l \mid C)$	Probability of choosing location alternative *l* given choice set *C*
$L(i)$	Likelihood for choice of aggregate alternative *i* from aggregate choice set *A* (where *A* is a partition of the disaggregate choice alternatives *C*)
$\bar{V}_i$	Mean utility of alternatives contained in aggregate alternative *i* (similar for alternative *j*)
m_i	Number of disaggregate alternatives in aggregate alternative *i*

3.1 Latent class with feedbacks (LCF) model

The utility functions of the choice components are

$$U_l = V_l + \epsilon_l \tag{1}$$

$$U_d = V_d + \epsilon_d \tag{2}$$

where U_l and U_d are the total random utility for the choice of location and dwelling type, respectively. Similarly, V and ϵ (with subscripts corresponding to location or dwelling type) are the systematic and random components of utility, respectively. We consider two latent classes: households who are more likely to switch dwelling type than location (Class1) and households who are more likely to switch location than dwelling type (Class2). We next present the full likelihood function for an observation, *i*, before outlining its components below.

$$L_i = Pr(C1) \prod_{l=1}^{L} \Big(\prod_{d=1}^{D} P\, r(l) Pr(d|l)^{\delta d} \Big)^{\delta l} + Pr(C2) \prod_{d=1}^{D} \Big(\prod_{l=1}^{L} P\, r(d) Pr(l|d)^{\delta l} \Big)^{\delta d} \tag{3}$$

where δ_l and δ_d are indicators of whether location l and dwelling type d, respectively, are chosen by the household. In the GEV structure of the model, the choice dimensions of dwelling type and location choice have scale parameters of μ_d and μ_l, respectively, while the class membership model has class-specific scale parameters of μ_c. For Class1, the probability of dwelling choice d conditional upon location choice *l* is given by the following multinomial logit expression

$$Pr(d \mid l) = \frac{\exp\,(\mu_d V_{d|l})}{\sum_{j \in D} \exp\,(\mu_d V_{j|l})} \text{ ; D is the set of feasible dwelling type} \tag{4}$$

$\frac{1}{ud} \, ln(\sum_{j \in D} \exp(\mu_d V_{j|d}))$ is the expected maximum utility (EMU) of dwelling choices for the location choice *l*. With dwelling type choice EMU feedback, the probability of location choice becomes

$$Pr(l) = \frac{\exp\left(\mu_l V_l + \frac{\mu_l}{\mu_d} \; ln(\sum_{j \in D} \left(\exp\,(\mu_d V_{j|d})\right)\right)}{\sum_{k \in L} \exp\left(\mu_l V_k + \frac{\mu_l}{\mu_d} \; ln\left(\sum_{j \in \mathrm{D}} \exp\; (\mu_d V_{j|d}\,)\right)\right)} \tag{5}$$

The joint probability of dwelling type and location choice for *Class*1 is then given by

$$Pr\,(dl) = Pr(d \mid l)Pr(l) = \frac{\exp\,(\mu_l\,V_l + \mu_d\,V_{d|l})\left(\sum_{j\in D}\exp(\mu_d\,V_{j|d})\right)^{\frac{\mu_l}{\mu_d}\;-1}}{\sum_{k\in L}\exp(\mu_l\,V_k\,)\left(\sum_{j\in D}\exp\,(\mu_d\,V_{j|d})\,\right)^{\frac{\mu_l}{\mu_d}}} \tag{6}$$

The expected maximum utility for Class1 that has dwelling type nested under location alternatives is given by

$$EMU(C1)=ln\left(\sum_{k\in L}\exp\left(\mu_l\,V_l+\frac{\mu_m}{\mu_l}\;ln\left(\sum_{j\in M}\exp\,(\mu_m\,V_{j|l})\right)\right)\right)^{\frac{1}{\mu_1}} \tag{7}$$

Similar expressions can be derived for the components of Class2. *Pr(C*1) is then given by

$$Pr(C1)= \frac{\exp\left(\mu_c\,V_{C1}+\mu_c\,EMU(C1)\right)}{\exp\left(\mu_c\,V_{C1}+\mu_c EMU(C1)\right)+\,\exp\left(\mu_c\,V_{C2}+\mu_c\,EMU(C2)\right)} \tag{8}$$

3.2 Semi-compensatory independent availability logit (SCIAL) model

The SCIAL model is a variant of the integrated discrete choice model with choice set formation proposed by Manski (1977) as follows

$$Pr(j) = Pr(j \mid C_i\,)Pr(C_i\,) \tag{9}$$

Swait and Ben-Akiva (1987) proposed an operational version of the model, which considers the probability of including alternative *j* in the choice set C_i as independent of considering other alternatives in the choice set. The model requires the estimation of 2^K-1 possible choice sets for *K* feasible alternatives. They name the model as the independent availability logit (IAL) model. In the IAL model, the consideration of an alternative is determined through a binary logit model wherein each alternative is fully considered as belonging to the choice set or being excluded from it.

Along a similar vein, Martinez et al. (2009) propose the constrained multinomial logit (CML) model. This model includes a penalty function, capturing the partial inclusion of an alternative in the choice set. However, the CML model does not capture the attributes of the decision-maker that contribute to their consideration of each alternative.

Habib (2019) derives the SCIAL model as a merging of the IAL and CML models within the original framing by Manski (1977). The SCIAL model includes both the probabilistic choice set formation of IAL and the semi-compensatory choice of CML. It takes the following form

$$Pr(j \mid C_i\,) = \frac{\exp\left(\mu V_j + \frac{1}{\mu}\,ln(A_j\,)\right)}{\sum_{k\in C_i} exp\left(\mu V_k + \frac{1}{\mu}\,ln(A_k)\right)} \tag{10}$$

where A_j is the probability of an alternative being included in the choice set.

$$Pr(C_i) = \frac{\prod_{k\in C_i} A_k \prod_{k\in(C-C_i)} A'_k}{1-\prod_{T\in C} A_T} \tag{11}$$

where A_k is the probability of an alternative being included in the choice set, A'_k is its complement, and $\prod_{T\in C} A_T$ are the set of all choice set alternatives. The choice set consideration function from equation 10 A_j is modeled as a binary logit model.

$$A_j = \left(1+ \exp\left(- \sum \gamma Z\right)\right)^{-1} \tag{12}$$

where Z is a vector of attributes of the decision-maker and γ is the corresponding parameter vector. However, the need to estimate 2^K possible choice sets introduces computational challenges as it requires large matrices to be stored in memory. It is proposed to use an aggregated choice set and the method used by Habibi et al. (2019) for car type choice. An aggregate choice set is defined as comprising aggregated locations and dwelling types, with disaggregate choice alternatives represented through a set of statistics.

Defining the disaggregate choice l as the combination of location choice among planning districts (PD) and dwelling types, the following expression provides a means of estimating a model of the aggregate choice i as the combination of location choice among regions and dwelling types

$$L_i = \sum_{l\in L_i} Pr(l \mid C) = \sum_{l\in L_i} \frac{\exp(V_l)}{\sum_{k\in C} \exp(V_k)} \tag{13}$$

where $P(l|C)$ is the probability of selecting disaggregate alternative l from the choice set of disaggregate alternatives $C$and $L(i)$ is the likelihood of selecting aggregate alternative $i \in A$ where the elements of A represent a partitioning of C.

The utility of the aggregate choice i is then expressed as a function of the attributes of the disaggregate alternatives contained within its partition as follows

$$L_i = \frac{\exp\left(\bar{V}_i + ln\,(m_i) + ln\left(\frac{1}{m_i} \sum_{l\in L_i} \exp(V_l - \bar{V}_i)\right)\right)}{\sum_{j\in A} \exp\left(\bar{V}_j + ln(m_j) + ln\left(\frac{1}{m_j} \sum_{k\in L_j} \exp\,(V_k - \bar{V}_j)\right)\right)} \tag{14}$$

where m_i is the number of alternatives in L_i and $\bar{V}_i$ is the mean utility of alternatives over $l \in L_i$. The above likelihood function represents the utility of an aggregate alternative as a combination of the mean utility, a size term, and a logsum term representing the heterogeneity in the underlying disaggregate alternatives.

4 Data

4.1 Primary data sources

The primary source of data for this research is a database of real estate listings collected for the GTA by the Toronto Real Estate Board (TREB). The TREB data is a near population sample of all residential sales made in the GTA dating from 2003 to the present. It includes the following attributes for each dwelling (among others): listed price, sale price, number of bedrooms, dwelling type, year built, days on the market. We combine these real estate data with census data. Statistic Canada conducted a census in both model years (2006 and 2016) from which we can obtain a range of sociodemographic and land-use information for each dissemination area (DA) in the GTA.

4.2 Data preparation

A significant shortcoming of using retrospective real estate listings is that they generally lack any information on the seller or purchaser. We generate sociodemographic attributes for the decision-makers (i.e., purchasers) by geolocating each real estate listing and matching it to a dissemination area (DA), the smallest census geography provided by Statistics Canada. Figure Figure 1 provides urban and rural examples of the association between real estate listings (shown in blue), DA (black boundaries), and PD (red boundaries). The DA show a close matching with real estate listings in the urban locations, and the larger DA areas in rural areas are deemed acceptable as there are lower populations in these DA, and sociodemographic attributes tend to be less variable.

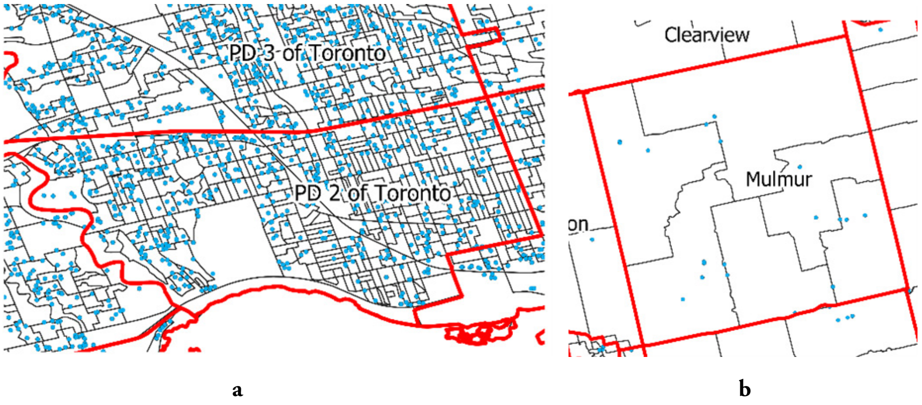

Figure 1. Example a) urban (PD 2 of Toronto), and b) rural (Mulmur PD) showing DA and residential sales density

For each listing, we impute household attributes from DA statistics. The purchaser income and household size are assumed equal to the median income and mean household size, respectively. Given the small geographic extent of DA, it is deemed a reasonable assumption that these attributes will be relatively similar - similar dwelling prices and sizes will encourage a homogenous group of households. A Canadian DA contains between 500 to 1,000 residents and is similar in its attributes to the lower population range of a block group in the United States. We use sample enumeration to assign a family composition, education level, and mover status to each household. In the case that no dwelling area is provided for a listing, we impute its area based on the sales price, the dwelling type, and the number of bedrooms. We estimate mean and standard deviation statistics for PD from the DA statistics obtained from Statistics Canada. The imputation method is limited by the fact that we are unable to distinguish between households that rent or own their dwelling. Finally, we estimate a measure of distance to the CBD for each dwelling (a point in downtown Toronto). Such geo-imputation is common in the health sciences literature when developing models in the absence of data describing individuals (Curriero et al., 2010; Henry & Boscoe, 2008; Klassen & Platz, 2006).

4.3 Definition of choice sets

The standard practice in location choice modeling is to define a set of generic alternatives. However, in the present case, we are interested in the propensity to switch between specific regions of the GTA. As such, we begin by defining the choice set as the combination of four dwelling types (single-family detached, semi-detached, townhouse, and apartment) and 40 planning districts. As some combinations are not present in the dataset, the choice set contains 155 alternatives in 2006 and 159 alternatives in 2016.

For both the LCF and SCIAL models, the large number of alternatives present challenges to estimation. In the case of the LCF model, defining a separate nest for each location would mean a model with 40 nests and require the estimation of upwards of 40 scale parameters. Such a model is intractable, and we estimated an initial set of models with five location nests representing the five regional authorities within the GTA: Durham Region, Halton Region, Peel Region, York Region, and the City of Toronto. As stated above, the SCIAL model requires the comparison of 2^K-1 possible choice sets. For a model with 159 alternatives, this means there will 7.3×10^{47}possible choice sets for each observation. By using the set of 20 aggregate alternatives (four dwelling types × five regions), we can reduce this number to 1.05 million choice sets. However, the model requires the storage of multiple matrices for the full datasets with dimensions of approximately 1 million × 90,000 and several operations that are not easily vectorizable. With some effort, it is possible to store the model in memory, but computation times are quite high. Therefore, we sought to further reduce the aggregate choice set. The initial application of the SCIAL model by Habib uses a choice set of seven alternatives. In our application, such a drastic reduction would affect the ability of the model to capture trade-offs between dwelling type and location. We settled upon a choice set of twelve alternatives (four regions and three dwelling types).

The four regions are defined using an aspatial k-means clustering of PDs rather than grouping along politically and spatially defined boundaries. This approach has the advantage that it allows for spatially discontinuous regions with similarities along a controlled dimension of analysis. The GTA is a polycentric region containing several large cities. By using an aspatial approach, we can cluster PDs by similarity in population density (as a proxy for urban form) and capture changes in the residential patterns of sociodemographic groups (e.g., moving from a suburban to an urban PD). We can also define multiple clusterings and test how the choice of regions affects model results. K-means clusters are defined by each of population density, proportion of persons commuting by auto mode, proportion of households living in detached dwellings, and median income.

The results of these clusterings are presented in Figure 2. Table 1 summarizes the mean clustering variable value within each region. There is a significant variation between regions for each of the considered variables. In addition, we reduce the number of dwelling types from four to three to further reduce computation time. Correlation analysis suggests that the semi-detached and townhouse types are the most similar, mainly according to average price (Pearson's r = 0.91), median household income (Pearson's r = 0.94), and frequency of colocation by PD (Pearson's r = 0.87).

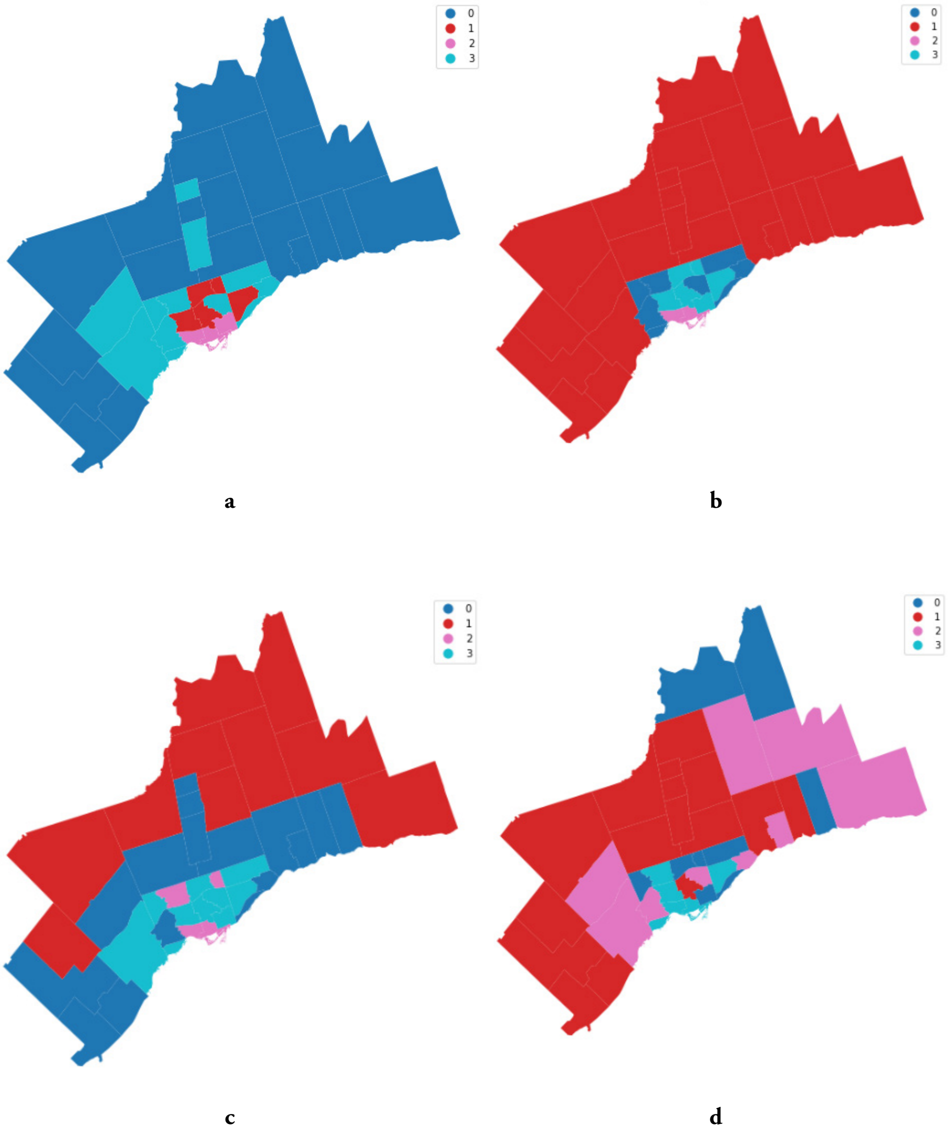

Figure 2. K-means clustering of PDs by a) population density, b) proportion of persons commuting by auto mode, c) proportion of households living in detached dwellings, and d) median income

Table 2. Mean clustering variable values among PDs in each region

Cluster	Median Income ($10,000)	Detached Houses (Proportion of All Dwellings in PD)	Auto Commuters (Proportion of All Persons in PD)	Population Density (Persons per km^2)
0	63.1	0.60	0.67	4.78
1	88.8	0.85	0.88	46.97
2	77.4	0.13	0.33	67.61
3	51.9	0.32	0.55	25.11

5 Results

As a first approximation to the problem, we plot the proportion of residential sales by dwelling type in each of the two study years in Figure 3. This analysis provides two forms of validation of our methods. First, it suggests our method of allocating incomes to purchasing households according to the median income of the DA where the property is located seems reasonable. As expected, the probability of purchasing a single-family detached dwelling tends to increase among high-income households. Second, the proportion of smaller dwellings (i.e., apartments) is higher across all income quintiles in 2016 compared with 2006. This result provides preliminary evidence for a transition towards a willingness to purchase a smaller dwelling to reduce household commuting time. If this change is systematic rather than a short-run effect, there are significant policy implications for residential zoning. In many cities, detached single-family zoning accounts for most of the available residential land. Many jurisdictions (Minneapolis, MN; California; Oregon; Vancouver, BC) have, or are considering, eliminating detached single-family zoning. In the City of Toronto, this zoning type accounts for 57% of residential land - for comparison, in Minneapolis, it is 70%, and in San Jose, it is 94% (Badger & Bui, 2019). If households are considering, if not choosing, other dwelling types, then policies must be adapted to facilitate the necessary shifts in supply.

There is certainly an argument to be made that higher proportions of apartment sales in 2016 are a function of the market supplying more of these dwellings. However, this pattern is an outcome of the complicated semi-dynamic relationship between supply and demand for housing. That is, more apartments are constructed in response to increased demand for this type of dwelling, which is itself a response to demand for detached dwellings driving up their price in a market with constraints on the available supply of land.

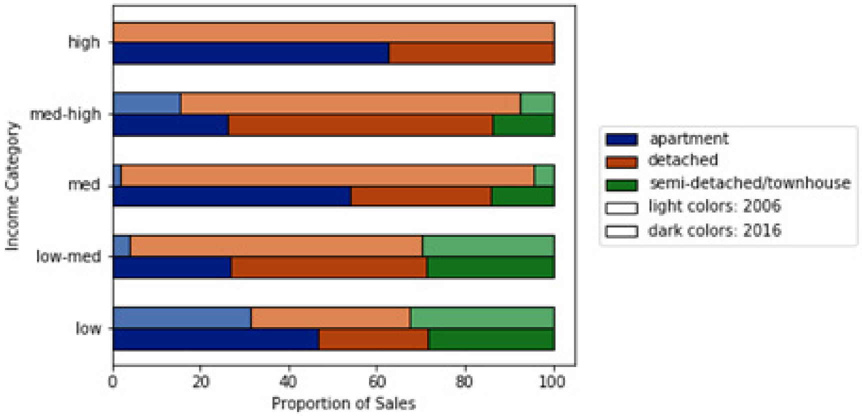

Figure 3. Proportion of residential sales in each study year by household income and dwelling type

5.1 Base MNL model results

We first estimate a simple multinomial logit (MNL) model. For each of the MNL models in Table 3, nested logit models were estimated, assuming dwellings nested within regions and regions nested within dwellings in turn and using the four regions and three dwelling type definitions of choice sets. We were unable to obtain significant scale parameters in any of these model specifications. We also note that the positive parameter for sale price violates economic theory stating that, all else being equal, people associate a disutility with the price. This parameter remained positive in most model specifications. We correct the sign on the price parameters in our SCIAL models (presented in Table 4) by introducing a Box-Cox transformation, which is reported by Gaudry (2018) to provide an appropriate correction in other models.

Table 3. Summary of baseline MNL model results for 2006 - 3 specifications

Variable name	**Model 1**		**Model 2**		**Model 3**	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
Sale price ($100,000)	0.036	6.05				
Price variation ($100,000)					0.237	11.82
Distance to CBD (km)	-0.012	-10.31				
Detached* - Toronto	0.305	4.37				
Semi-detached* - Toronto	-0.907	-7.98				
Townhouse* - Toronto	-1.061	-9.90				
Detached* - Area (100s m^2)			0.0278	6.91		
Semi-detached* - Area (100s m^2)			0.0168	3.77		
Townhouse* - Area (100s m^2)			0.00620	1.46		
Bedrooms (per HH member)	-0.820	-10.08				
Dwelling age			0.500	1.10		
Commute time (proportion <30 minutes)					0.501	1.77
Detached - Married**			0.501	4.34		
Detached - Separated**			0.501	4.72		
Diploma - Toronto					0.472	5.98
Bachelor's degree - Toronto					0.493	5.10
Graduate degree - Toronto					0.494	3.92
Detached* - Mover (last 5 years)					0.487	5.76
Semi-detached* - Mover (last 5 years)					0.503	5.20
Townhouse* - Mover (last 5 years)					0.497	4.91

* Reference: apartment
** Reference: single household

5.2 LCF model results

For the LCF model, the components of the model are as follows:

1. Variables describing location choice
2. Scale parameters for each of the dwelling over region nests
3. Scale parameters for each of the regions over dwellings nests
4. Scale parameters for each of the latent segments (fix one for identification)
5. Variables describing membership in a latent segment

We explored several sets of variables describing location choice, interacted with various scale parameter specifications. The specifications explored include fixing all the scale parameters to one except for one of the latent segments, estimating only the scale parameters for each of the regions or dwellings, estimating a single scale parameter for each of the regions, or dwellings, and a general specification where we estimate all the scale parameters (except one). In addition, we explored an exhaustive set of sociodemographic variables describing latent membership. In no case were we able to obtain significant scale parameters. Finally, we estimated several of these models with the additional aggregation of dwellings and regions employed in the SCIAL model below (four regions and three dwelling types). However, no significant scale parameters were obtained with this modification, either. These results suggest that the hypothesized nesting structures are not borne out by the data. However, our results should not be interpreted as definitive, and the relationship may exist using a different dataset or combination of variables.

5.3 SCIAL model results

The SCIAL model provides more robust results. Due to computer memory constraints and the large number of alternatives, the model was run with a random 10% sample of listings. However, given that the data represents the population of sales, this sample still includes 4,081 observations for 2006 and 9,324 for 2016. The aggregate choice set used in the choice set formation component of the model is summarized in Table 3.

Table 4. Choice set alternatives used in SCIAL model estimation

Choice set Alternative	Dwelling Type	Income Class	Dwelling Class (Proportion Single Detached)	Mode Class (Proportion auto-drive commutes)	Population Class
1	Single-family	Low-Med	Low	Low-Med	Low
2	Single-family	High	High	Low	High
3	Single-family	Low	Med-High	High	Low-Med
4	Single-family	Med-High	Low-Med	Med-High	Med-High
5	Townhouse	Low-Med	Low	Low-Med	Low
6	Townhouse	High	High	Low	High
7	Townhouse	Low	Med-High	High	Low-Med
8	Townhouse	Med-High	Low-Med	Med-High	Med-High
9	Apartment	Low-Med	Low	Low-Med	Low
10	Apartment	High	High	Low	High
11	Apartment	Low	Med-High	High	Low-Med
12	Apartment	Med-High	Low-Med	Med-High	Med-High

We began model estimation by testing a variety of variables in the choice model with the 2006 data: proportion of commuters with travel times less than 30 minutes, average days dwellings were on the market for each combination of dwelling type and PD, price variation in the PD, age of the dwelling and age of the individual, dwelling type and mover status of the individual, and family type and detached dwelling. In most specifications, these variables were not significant at the 0.10 level. We decided to maintain a consistent set of choice model variables (distance to Toronto CBD, sale price, and dwelling-specific area) to reduce the combinatorial complexity of the model testing. This choice was necessitated by the long estimation times for each model (~1-2 days), the need to test four choice set definitions for both 2006 and 2016, and our focus being choice set generation variables. The final models are presented in Table 5. We did not obtain a consistent model with the choice set defined by the proportion of commuting trips made by the auto-drive mode. In addition to the variables included in the final models, we also explored household income, mover status (in the last 5 years), level of education, and various transformations of all the variables. Alternative-specific constants (ASC) in the model can be considered as fixed effects that control for unobserved variation across regions and dwelling types.

The models have consistently negative signs for sale price. The positive sign for distance is counter to the assumption of bid-rent theory, which states that land prices tend to decrease with distance from the CBD (Alonso, 1960). We attribute this sign to a lack of normalization of prices by dwelling area - larger dwellings will tend to be more expensive. The dwelling area variables included in the model only account for variation in the effect of the area between dwelling types, with the detached single-family set as the reference type. We find that single individuals and those under the age of 35 tend to consider different dwellings than families and older individuals. They are generally less likely to consider detached single-family dwellings. Household size also tends to affect the types of dwellings and locations considered by households. Each model was compared against its MNL version (i.e., same explanatory variables but neglecting the choice set generation model). In all cases, AIC statistics show an improvement in fit with the extension to a SCIAL model.

Table 5. Summary of SCIAL model results

	Variable name	**Model 1 (by dwelling type)**		**Model 2 (by population density)**	
		2006	2016	2006	2016
Choice Variables	ASC - Townhouse	-1.297	-1.332	1.068	1.070
	ASC - Apartment	-0.565	-0.431	0.329	0.304
	ASC - Region 2	0.139	0.284	0.383	0.617
	ASC - Region 3	6.605	6.604*	1.333*	1.338*
	ASC - Region 4	-0.772*	-0.726	1.268*	1.247*
	Transformed sale price	-0.804**	-0.918**	-0.654**	-0.631**
	Sale price - BC transformation	0.205	0.152	0.505*	0.505**
	Area (100s m²) - Townhouse	0.311**	1.976**	1.333**	1.338**
	Area (100s m²) - Apartment	-0.034	0.088**	-4.557*	-4.535**
Choice Set Variables	ASC - Townhouse	-3.063**	-2.947**	-3.151**	-2.924**
	ASC - Apartment	-3.031*	-2.635**	-0.919**	-0.459**
	ASC - Region 2	-1.480**	-0.748**	-2.725**	-2.682**
	ASC - Region 3	-0.761**	-0.543**	-1.184**	-1.323**
	ASC - Region 4	-0.279	0.388*	-1.771**	-1.890**
	Age < 35 - ALT2			-0.097	-0.146
	Age < 35 - ALT3			0.339**	0.289**
	Age < 35 - ALT4			0.001	-0.081
	Age < 35 - ALT5			0.568**	0.583**
	Age < 35 - ALT6			0.290	0.348**
	Age < 35 - ALT7			0.904**	0.972**
	Age < 35 - ALT8			0.494*	0.473**
	Age < 35 - ALT9			0.201	-0.171
	Age < 35 - ALT10			0.906**	1.145**
	Age < 35 - ALT11			0.960**	0.849**
	Age < 35 - ALT12			0.853**	0.792**
	HH size - ALT 2	0.433	0.818	0.584**	0.095*
	HH size - ALT 3	0.081	0.281*	0.244**	0.282**
	HH size - ALT 4	0.815*	0.537	0.198**	-0.222**
	HH size - ALT 5	0.055	-0.400**	0.278**	0.299**
	HH size - ALT 6	0.500**	0.312	0.586**	0.390**
	HH size - ALT 7	0.594**	1.342**	0.682**	0.798**
	HH size - ALT 8	0.771**	0.545**	0.391**	0.322**
	HH size - ALT 9	0.356**	0.270	-0.694**	-1.042**
	HH size - ALT 10	-0.336	-0.592	0.771**	1.135**
	HH size - ALT 11	-0.080	-0.830**	0.859**	0.649**
	HH size - ALT 12	1.300**	1.354**	1.016**	0.956**
ρ_c^2		0.287	0.263	0.174	0.186
AIC		17,024	37,437	18,327	45,728
AIC (MNL)		24,219	56,001	24,219	56,001

* Significant at 0.10 level

** Significant at 0.05 level

5.4 Marginal effects analysis

In this section, the focus is marginal effects representing the effect of explanatory variables on the choice set inclusion probability. Only variables with significant parameters at the 0.10 level are included in the plots. 2006 marginal effects are excluded unless a significant difference was found between years according to paired t-tests. Marginal effect plots for binary variables show the probability with and without the variable in the model. The difference in choice set probability can be interpreted as a measure of the marginal effect of the variable.

Figure 4 suggests that younger individuals are more likely to consider apartments in medium to high-density areas and townhouses in low or medium density areas. There are no significant differences between 2006 and 2016 results for this variable, suggesting that the effect of age on location and dwelling type preferences did not change over the decade of the study period.

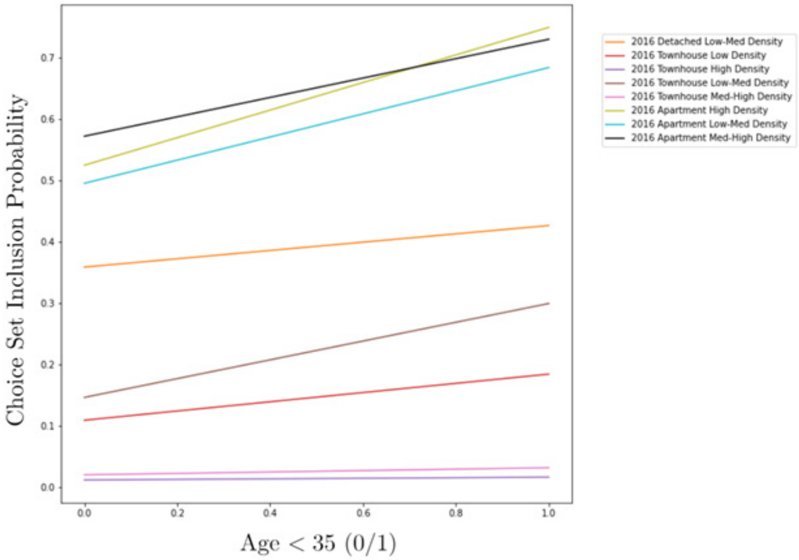

Figure 4. Marginal effect of individual younger than 35 on choice set inclusion — Clustered by PD population density (reference = detached in low-density region)

When clustering PD by population density, a significant difference is found between 2006 and 2016 for household size effects (see Figure 5). Larger households became less likely to choose a detached house in a medium to high-density region over the intervening decade. These households also became more likely to choose an apartment in a high-density region. These results support the hypothesis that larger households are increasingly considering high-density urban areas and smaller dwellings. Concerning detached low-density regions, larger households are far less likely to consider apartments than detached houses. I was unable to obtain a consistent model with the inclusion of an income variable. However, this result suggests an income effect through differences in the price of apartments in low-density and high-density areas and the additional household income associated with larger households that can pool their income. Larger households will generally prefer detached dwellings but, given a choice between apartments, they are more likely to consider one in a high-density area.

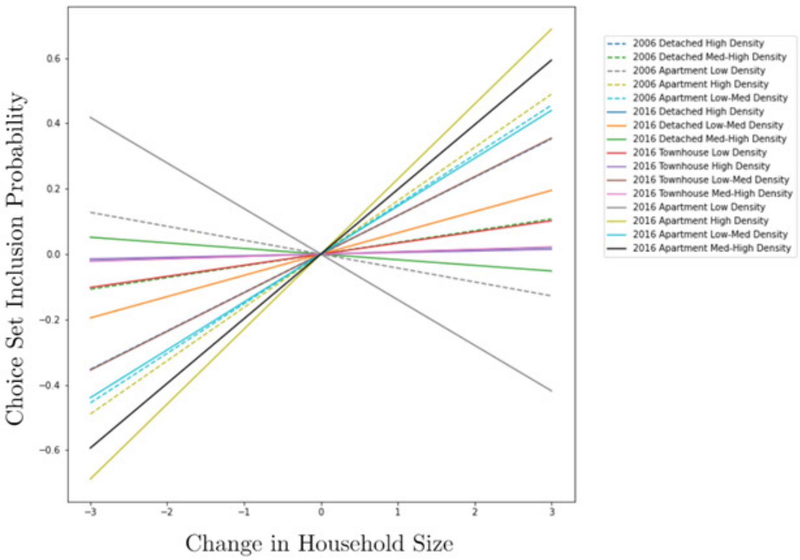

Figure 5. Marginal effect of household size on choice set inclusion — Clustered by PD population density (reference = detached in low-density region)

When clustering PD by the proportion of detached dwellings, it is found that larger households exhibit changes in their preferences for townhouses (see Figure 6). Larger households in 2016 were less likely to consider townhouses in low-detached dwelling areas and more likely to consider them in medium-high detached dwelling areas. Household size is most strongly associated with considering townhouses in areas with a low to medium proportion of detached dwellings. These locations are characteristic of inner suburban neighborhoods, having many of the amenities of urban neighborhoods without the high density and traffic.

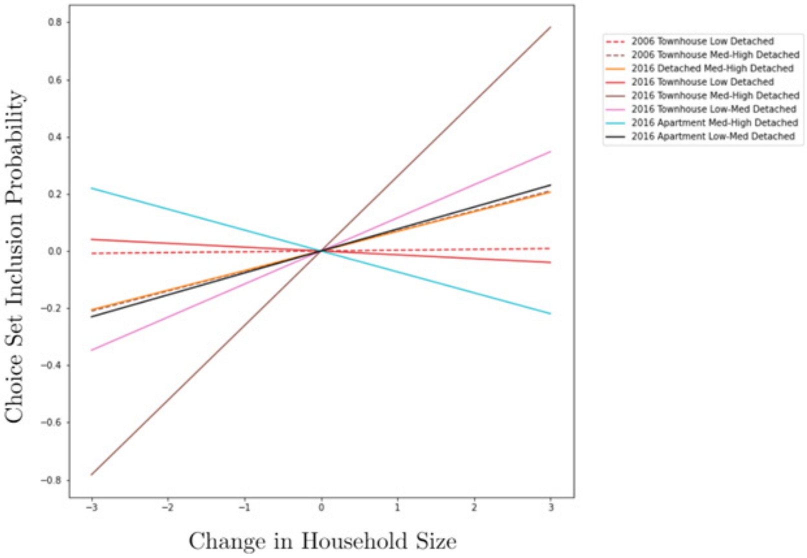

Figure 6. Marginal effect of household size on choice set inclusion — Clustered by PD proportion detached dwellings (reference = detached in low detached region)

6 Conclusions and future work

In this study, we used a panel of real estate transactions to examine the evolution of dwelling and location choice preference in a large metropolitan area over the decade between 2006 and 2016. The research began from a need to bring a quantitative perspective to the discussion around the re-emergence of a preference for urban living in Western cities. Two advanced discrete choice model formulations, LCF, and SCIAL were explored for their ability to capture the sociodemographic factors that influence the preference for dwelling type and location typologies. These characteristics were imputed for each real estate transaction using spatially detailed census and land-use data. The computational challenges associated with choice set generation models, such as SCIAL, were addressed through the combination of clustering of alternatives by location according to common features and the use of a recently proposed choice aggregation method. Model estimation was performed on a large sample of the data in each of 2006 and 2016. Finally, marginal effects were estimated for choice set inclusion.

A significant increase was found over the analysis period in large households considering apartments in dense areas and townhouses in inner suburban areas. These results support policies aimed at encouraging the construction of townhouses and semi-detached dwellings through modification of zoning. Other results, suggesting no significant difference in marginal effects between years, provide support for the persistence of many patterns of dwelling and location preference over time. Many operational land-use and transportation models make simplifying assumptions about choice set formation that ignore sociodemographic differences. The above results indicate that such simplifications may lead to incorrect policy conclusions due to differences in choice set inclusion probability across individual and household attributes that are persistent through time. Finally, the sociodemographic imputation used in this research reduces household heterogeneity. The LCF model may provide stronger results with the direct collection of these variables. An alternative data generation approach would be to use household demographics from the TTS travel survey and fuse these data with the real estate data using detailed coordinates. Finally, a more complete study of changes in preferences might include consideration of changes in housing stock in the region, which will influence the supply side of the location choice problem.

References

Alonso, W. (1960). A theory of the urban land market. *Papers and Proceedings of the Regional Science Association, 6*(3), 49–57. https://doi.org/10.1111/j.1435-5597.1960.tb01710.x

Badger, E., & Bui, Q. (2019, June 18). Cities start to question an American ideal: A house with a yard on every lot. *New York Times.* https://www.nytimes.com/interactive/2019/06/18/upshot/cities-across-america-question-single-family-zoning.html

Banister, D. (2005). Unsustainable transport: City transport in the new century. In *Unsustainable transport: City transport in the new century.* London: Routledge. https://doi.org/10.4324/9780203003886

Bhat, C. R. (2015). A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. *Transportation Research Part B: Methodological, 79,* 161–188. https://doi.org/10.1016/j.trb.2015.05.021

Boehm, T. P. (1982). A hierarchical model of housing choice. *Urban Studies, 19*(1), 17–31. https://doi.org/10.1080/00420988220080021

Boustan, L., & Shertzer, A. (2013). Population trends as a counterweight to central city decline, 1950-2000. *Demography, 50*(1), 125–147. https://doi.org/10.1007/s

Brownstone, D., & Englund, P. (1991). The demand for housing in Sweden: Equilibrium choice of tenure and type of dwelling. *Journal of Urban Economics, 29*(3), 267–281. https://doi.org/10.1016/0094-1190(91)90001-N

CARB. (2020). *Sustainable communities.* Sacramento, CA: California Air Resource Board.

Caton, H. (2016, May 3). More families are choosing condo life over house in suburbs, study finds. Toronto.com. https://www.toronto.com/news-story/6525372-more-families-are-choosing-condo-life-over-house-in-suburbs-study-finds/

Curriero, F. C., Kulldorff, M., Boscoe, F. P., & Klassen, A. C. (2010). Using imputation to provide location information for nongeocoded addresses. *PLoS ONE, 5*(2), e8998. https://doi.org/10.1371/journal.pone.0008998

Florida, R. (2012). *The rise of the creative class* (2nd ed.). New York: Basic Books.

Fu, X., Bhat, C. R., Pendyala, R. M., Vadlamani, S., & Garikapati, V. M. (2015). *Understanding the multiple dimensions of residential choice.* Paper presented at the Transportation Research Board 94th Annual Meeting, Washington, DC.

Gaudry, M. J. I. (2018). *The replacement of regressions of convenience by Box-Cox Likelihoods: Transport model evidence and link to Hume's constant conjunction* (AJD report no. 174). Montreal: University of Montreal. https://doi.org/10.13140/RG.2.2.27635.63527

Glaeser, E. (2012). *Triumph of the city* (1st ed.). London: Penguin Books.

Habib, K. N. (2019). Mode choice modelling for hailable rides: An investigation of the competition of Uber with other modes by using an integrated non-compensatory choice model with probabilistic choice set formation. *Transportation Research Part A: Policy and Practice, 129,* 205–216. https://doi.org/10.1016/j.tra.2019.08.014

Habibi, S., Frejinger, E., & Sundberg, M. (2019). An empirical study on aggregation of alternatives and its influence on prediction in car type choice models. *Transportation, 46*(3), 563–582. https://doi.org/10.1007/s11116-017-9828-5

Henry, K. A., & Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *International Journal of Health Geographics, 7.* https://doi.org/10.1186/1476-072X-7-3

Hortas-Rico, M. (2015). Sprawl, blight, and the role of urban containment policies: Evidence from U.S. Cities. *Journal of Regional Science, 55*(2), 298–323. https://doi.org/10.1111/jors.12145

Hossain, S., Hasnine, M. S., & Habib, K. N. (2020). A latent class joint mode and departure time choice model for the Greater Toronto and Hamilton area. *Transportation, 48*, 1217–1239. https://doi.org/10.1007/s11116-020-10092-1

Hume, C. (2018, December 19). Family-sized condo units are changing Toronto's real estate market. *Storeys.* https://storeys.com/family-sized-condo-units-are-changing-torontos-real-estate-market/

Hyra, D. (2015). The back-to-the-city movement: Neighbourhood redevelopment and processes of political and cultural displacement. *Urban Studies Journal Limited, 52*(10), 1753–1773. https://doi.org/10.1177/0042098014539403

Kalinowski, T. (2020, February 6). Toronto house prices are expected to soar by nearly 10 percent in 2020 — and rents will jump up too. *Toronto Star.* https://www.thestar.com/business/2020/02/06/toronto-house-prices-expected-to-climb-nearly-10-per-cent-in-2020.html#:~:text=too%20%7C%20The%20Star-,Toronto%20house%20prices%20are%20expected%20to%20soar%20by%20nearly%2010,rents%20will%20jump%20up%20too&text=Any%20lingering%20concerns%20that%20the,Real%20Estate%20Board%20(TRREB)

Kaplan, S., Bekhor, S., & Shiftan, Y. (2013). Hybrid compensatory–noncompensatory choice sets in semicompensatory models. *Transportation Research Record: Journal of the Transportation Research Board, 2322*(1), 10–19. https://doi.org/10.3141/2322-02

Kaplan, S., Shiftan, Y., & Bekhor, S. (2012). Development and estimation of a semi-compensatory model with a flexible error structure. *Transportation Research Part B: Methodological, 46*(2), 291–304. https://doi.org/10.1016/j.trb.2011.10.004

Klassen, A. C., & Platz, E. A. (2006). What can geography tell us about prostate cancer? *American Journal of Preventive Medicine, 30*(2S), S7–S15 https://doi.org/10.1016/j.amepre.2005.09.004

LeBlanc, D. (2018, September 20). Moving from a house to a condo proved to be the right choice for this Toronto couple. *The Globe and Mail.* https://www.theglobeandmail.com/real-estate/toronto/article-moving-from-a-house-to-a-condo-proved-to-be-the-right-choice-for-this/

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision, 8*(3), 229–254. https://doi.org/10.1007/BF00133443

Martínez, F., Aguila, F., & Hurtubia, R. (2009). The constrained multinomial logit: A semi-compensatory choice model. *Transportation Research Part B: Methodological, 43*(3), 365–377.

Mercer Canada. (2019, June 26). Mercer Canada: 2019 cost of living survey. https://www.mercer.ca/en/newsroom/mercers-25th-annual-cost-of-living-survey.html

Moore, A. T., Staley, S. R., & Poole, R. W. (2010). The role of VMT reduction in meeting climate change policy goals. Transportation Research Part A: Policy and Practice, 44(8), 565–574. https://doi.org/10.1016/j.tra.2010.03.012

Platt Boustan, L., Bunten, D., & Hearey, O. (2013). Urbanization in the United States, 1800-2000. In L. Cain, P. Fishback, & P. Rhode (Eds.), T*he Oxford handbook of American economic history*. Oxford, England: Oxford University Press. https://doi.org/10.1007/BF02800653

Quigley, J. M. (1985). Consumer choice of dwelling, neighborhood and public services. *Regional Science and Urban Economics, 15*(1), 41–63. https://doi.org/10.1016/0166-0462(85)90031-6

Smith, N. (2019). Toward a theory of gentrification. A back to the city movement by capital, not people. *Suburban, 7*(3), 65–86. https://doi.org/10.36900/suburban.v7i3.546

Statistics Canada. (2016). 2*016 census highlights: Factsheet 14.* https://www.fin.gov.on.ca/en/economy/demographics/census/cenhi16-14.html

Stieger, S., & Lewetz, D. (2016). Parent-child proximity and personality: Basic human values and moving distance. *BMC Psychology, 4*(1), 1–12. https://doi.org/10.1186/s40359-016-0132-5

Swait, J., & Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B, 21*(2), 91–102. https://doi.org/10.1016/0191-

2615(87)90009-9

Tu, Y., & Goldfinch, J. (1996). A two-stage housing choice forecasting model. *Urban Studies, 33*(3), 517–537. https://doi.org/10.1080/00420989650011898

Van Eggermond, M. A. B., Erath, A., & Axhausen, K. W. (2018). Residential search and location choice in Singapore. *2018 Transportation Research Board Annual Meeting Online.* Washington, DC: TRB. https://doi.org/10.3929/ethz-b-000193146

Zolfaghari, A., Sivakumar, A., & Polak, J. (2013). Simplified probabilistic choice set formation models in a residential location choice context. *Journal of Choice Modelling, 9*, 3–13. https://doi.org/10.1016/j.jocm.2013.12.004