The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 873–892

A joint model of place of residence (POR) and place of work (POW): Making use of Gibbs sampling technique to overcome arbitrary assumptions in contexts of data limitation

Hengyang Zhang
University of Toronto
emmahengyang.zhang@mail.utoronto.ca

Jason Hawkins (corresponding author)
University of Toronto
jason.hawkins@mail.utoronto.ca

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

Abstract: Place or residence (POR) and place of work (POW) are two spatial pivots defining patterns of travel behavior. These choices are considered part of long-term choice influencing short-term daily travel choices. Hence, POR-POW distributions are input into almost all daily travel demand models. However, in many cases, POW-POR is modelled in an ad-hoc way considering the gravity-based or entropy is maximizing aggregate modelling approach. Lack of data on the sequence of choices related to POR and POW is often blamed for avoiding using disaggregate choice model. Recognizing such data limitation, this paper presents an alternative methodology of modelling joint distribution of POW-POW that uses disaggregate choice models without necessarily knowing the sequence of POR and POW choices. It uses the conditional probability break downs of joint POR-POW choice probabilities as depicted in the Gibbs sampling approach. This allows capturing effects of household socioeconomic characteristics, zonal land-use characteristics, and modal accessibility factors in the POR-POW models. The model is applied for a case study in the city of Ottawa. Results reveal that the proposed methodology can replicate observed patterns of POR-POW with a high degree of accuracy.

Keywords: Location choice, mode choice, discrete choice model, Gibbs sampling, multinomial logit model, accessibility

Article history:
Received: June 19, 2019
Received in revised form: October 29, 2019
Accepted: October 31, 2019
Available online: December 4, 2019

1 Introduction

Residential and work location choices are central to transportation and land development planning. In the aggregate, travel between the place of residence (POR) and place of work (POW) determines periods of congestion on the transportation network. At the household level, these choices affect the viability of public transit and active modes (i.e., walking and cycling), the number of household automobiles, and the time available for leisure. Both choices are made according to a long-time horizon (i.e., years) and heavily influenced by prevailing patterns of land use. As such, it is essential that models used in planning accurately capture the mechanisms driving these decisions, as well as their interactions (Jiao, Liu, & Guo, 2015).

http://dx.doi.org/10.5198/jtlu.2019.1624
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

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.

Methods for modelling POR and POW decisions can be broadly classified as either aggregate or disaggregate in nature. Aggregate methods are dominated by gravity, or entropy, techniques, which segment the study region into a series of zones and distribute households and jobs among these zones. The method of distribution is generally based on a measure of spatial proximity and accessibility. That is, the probability of a person choosing a particular POW will diminish with distance from their POR. The second class of methods is disaggregate and dominated by discrete choice models. These models are typically one-dimensional, meaning the POR is determined conditionally on a known POW or vice versa. Improvements in computing power have made disaggregate models increasingly viable, and this class of models is the state-of-the-art in transportation demand analysis. However, independently modelling the POR and POW does not capture the strong dependency between these decisions observed in the literature (Waddell, Bhat, Eluru, Wang, & Pendyala, 2007). The number of models that jointly capture the POR and POW decisions is sparse, and many are not readily interpretable by model users. If one employs a nested logit structure, as proposed by Waddell et al. (2007), the outcome is a measure of the relative propensity to change residential or work location. However, it is often erroneously given an interpretation of choice timing, which cannot be determined without the inclusion of data as to the timing (and sequence) of each decision.

This paper seeks to address this gap in the literature and proposes a simple method for jointly modelling these two decisions by exploiting the concept used in the Gibbs sampling technique. To develop the joint model, two conditional location choice models are estimated for each of POR and POW. Gibbs sampling approach is used to draw a joint model based on the two conditional location choice models. The empirical model estimation uses the 2011 National Capital Region (NCR) Household Origin-Destination Survey in Ottawa, Canada (Trans Committee, 2013). This approach allows capturing modal opportunity between origin-destination pairs of the study area in a meaningful way. The remainder of the paper is structured as follows: section 2 provides an expanded review of the literature, section 3 describes the model methods, section 4 outlines the data sources, section 5 summarizes the empirical model results, and section 6 concludes and gives recommendations for future research.

2 Literature review

2.1 Location choice models

According to Schirmer, Van Eggermond, and Axhausen (2014), the development of the residential location model can be separated into three stages. Initially, the bid-rent concept was first proposed in the 19th century by von Thünen in 1826 (Wartenberg & Hall, 1966) to model the residential location choice and further developed by Alonso (1964). Bid-rent assumes that bidders maximize a utility function depending on various characteristics of the dwelling, and the dwelling is sold to the highest bidder. Following this, Lowry (1964) utilized a gravity model and allocated households to residential zones based on employment levels, residential attractiveness, and a defined deterrence function. Finally, the residential location choice was among the first applications of the multinomial logit model developed by McFadden (1978).

As for the work location, most studies analyze the interdependence between residential location and workplace, rather than explicitly modeling the work location choice. In early applications, both residential location and work location were jointly modeled in aggregate through the trip generation and trip distribution components of the 4-step model introduced by Manheim (1979). Trip generation determines the total number of commuting trips generated from each origin zone and the total number of commuting trips attracted to each destination zone based on several explanatory variables (e.g., zonal employment and area). In trip distribution, the most likely distribution of the commuting trips can be calculated and summarized in an origin-destination (OD) matrix by using the gravity model. The distribution of commuting trips encapsulates the residential and work location choice, in aggregate. However, this method cannot capture individual location choice preferences.

There have been several efforts to overcome the limitations of aggregate gravity models. A standard method is to jointly model residential and work location choice using discrete choice models. Siegel (1975) and Simpson (1987) are among the first efforts in this area. Rivera and Tiglao (2005) estimate a 2-level nested logit (NL) model to jointly model residential location choice, work location choice, and mode choice of two-worker households in Metro Manila. They suggest that choice makers are more likely to switch the lower level choice (i.e., modes) before switching the choice of the upper level (i.e., a combination of residential and work location choice). Ibeas, Cordera, Dell'olio, and Coppola (2013) indicate that a cross-nested logit (CNL) model can better simulate residential and work location choice than either the conventional MNL model or NL models. The CNL model effectively captures more spatial interactions between the two location choices. Jiao et al. (2015) develop a joint residence-workplace location choice model for Beijing using a mixed logit model, which shows that individuals tend to choose workplaces with higher GDP values and lower commuting distances.

Clark and Withers (1999) utilize panel data to model the choice timing process explicitly. They define proportional hazard models for each of residential and work location choice. The triggering event of a change in work location is included in the hazard model for termination of the current residential location, and vice versa. This study provides insights as to the interaction of the two decisions but does not represent a choice model for matching the joint distribution of residential and work location in the context of a regional model.

Most of the discrete choice models found in the literature have an implicit assumption of the choice timing. Many assume that residential and work locations are chosen simultaneously, while others assume that one choice uniformly precedes the other. Also, the disaggregate choice approach tends to obscure the aggregate link between POR and POW. Since the work of Wilson (1970), the gravity model has provided a theoretical basis for aggregate trip distribution patterns through an extension to entropy. This assumes that the distribution of trips satisfies a most probable, lowest information, condition.

To add to this body of literature on travel demand modelling, the paper proposes an alternative methodology that uses two disaggregate discrete choice model in an intuitive conditional relationship fashion. However, it does not require the assumption of timing or sequence of the choice of home or work locations. We believe, the proposed methodology will benefit travel demand modelling system applications where POR-POW need to be used as pre-requisite.

2.2 Location choice attributes

Appropriate attributes are necessary to construct reliable POR and POW choice models. Schirmer et al. (2014) conduct a review of POR choice models and categorize their major attributes into two groups: 1) characteristics of the decision maker and 2) characteristics of the location. The characteristics of the decision maker are typically household attributes. According to Heldt, Gade, and Heinrichs (2016), typical household attributes used in empirical studies include household size, presence of children, age of homeowner, household resources (including income level, education level, and employment status). The location attributes are further classified into three levels, including socioeconomic environment, land-use attributes, and accessibility. The attributes that affect POR are also closely associated with POW.

Socioeconomic environment captures the socioeconomic characteristics of each potential location, such as population density, educational level, and dwelling price. For example, Abraham and Hunt (1997) find that a higher population density is usually associated with more potential opportunities and better accessibility; thus, it has a positive utility in the household residential location choice. Bhat and Guo (2007) indicate that school quality is also an important factor in POR choice, especially for households with higher educational attainments. Dwelling price and tax are also major factors in POR choice. Kryvobokov (2015) constructs a separate hedonic price model as an input to the residential location choice model.

Heldt et al. (2016) find that POR choice is also significantly related to the land-use patterns. Typically, "points of interest" are utilized to represent the land-use pattern by capturing specific functional points relevant to the public, such as education facilities, hospitals, and subway stations. Moreover, the mix of land uses in an area influences the POR choice (Tran, Zhang, Chikaraishi, & Fujiwara, 2016).

Chen, Chen, and Timmermans (2008) classify accessibility as being to either work, retail, or open spaces and parks. Accessibility to work can be captured by commuting time and an employment index (i.e., number and type of jobs in each zone), where commuting time has a negative effect on accessibility and employment a positive effect. However, the effect of accessibility to retail is mixed depending on the empirical data (Timmermans, Van Noortwijk, Oppewal, & Van Der Waerden, 1996). Farley (1995) finds that accessibility to open areas has a variable effect on POR utility, depending upon the household income. In summary, the attributes discussed above should be treated as a preliminary starting point, which should be evaluated through an iterative testing procedure as supported by the empirical dataset.

3 Methodology

In general, the model construction can be separated into three parts:

1. Evaluating the modal accessibility by estimating a mode choice model
2. Developing two condition location choice models, including a residential location choice model given a work location choice and a work location choice model given a residential location choice
3. Adapting Gibbs sampling to develop a joint model of residential and work location choice

3.1 Modal accessibility

According to the literature review, modal accessibility is a vital attribute with regards to POR and POW choice. As such, this paper develops a MNL commuting mode choice model to evaluate the modal accessibility based on the log sum measure.

Before constructing the model, it is necessary to define the feasibility of specific choice sets. The mode choice is categorized into six alternatives: Auto Drive (AD), Auto Passenger (AP), Carpool & Ride (CR), Active (A), Park & Ride (PR), and Transit (T). The following rules are utilized to identify the feasible choice set:

1. AD is feasible if the respondent has a valid driver's license, and the number of vehicles in the household is greater than or equal to one.
2. AP is generally a feasible mode unless the respondent indicates he/she will not consider AP during the commuting trip. As the survey does not provide this information, AP is considered as a feasible mode for all trip records.
3. CR is feasible when transit service is available to the work location.
4. The feasibility of mode A (cycle and walk) is defined by a threshold value of *Euclidean Travel Distance* between the origin and destination. The threshold is set at 6.5 kilometers, which is the 85th percentile travel distance for survey cycling trips.
5. PR is a feasible choice if the individual has a valid driver's license and at least one vehicle. The availability of a transit stop at the work location is also a requirement.
6. Transit is feasible if the transit service is available for both the origin and the destination zone.
7. Finally, the chosen mode is always included in the feasible choice set.

After identifying the feasible choice sets and potential attributes, the commuting mode choice model can be developed to evaluate modal accessibility. We use a MNL mode choice model and define the accessibilities as log sum measures. Thus, the *home zone accessibility* is given by

$$A_n = ln\left(\sum_{m \in M^n} e^{V_{nm}}\right) \tag{1}$$

$$A_i = \frac{\sum_N A_n * F_n}{\sum_N F_n} \tag{2}$$

Where A_n is the accessibility for individual n, V_{ij} is the systematic utility for individual n and mode choice m (from a feasible choice set M^n for individual n), A_i represents the average accessibility of home zone i, F_n is an expansion factor for individual n based on statistical accounts, and N is the set of all individuals residing in home zone i. A similar procedure is utilized to develop the *work zone accessibility*. Thus, the resultant *home zone accessibility* and *work zone accessibility* of each zone in the model region can be applied to the two location choice models.

3.2 Conditional location choice models

The two conditional location choice models are a residential location choice model, given the work location, and a work location choice model, given the residential location. Typically, a household only has one residential location but may have multiple work locations if there is more than one person in the household. These intrahousehold interactions further complicate the model and are not considered in this initial work. The two conditional location choice models are developed for single-person households so that each household has one residential location and one work location. An extension could be made to multiple-worker households through a nested logit structure, similar to Abraham and Hunt (1997), wherein the work location for the second worker is nested below the work location decision of the first worker. However, there are issues of defining the household role of each worker – should members be defined as husband/wife? What happens in the case of non-traditional household configurations and multi-worker households? An alternative formulation would be to include additional workers as a new step in the Gibbs sampling procedure outlined in Figure 1 below. Each iteration of the algorithm would include an additional POW estimation for the second, third, etc. worker in each household until convergence of the full POR/POW system.

The model system is further simplified by using traffic analysis zones (TAZ) as the level of location choice, to reduce the number of potential alternatives and match the data available for empirical analysis. To construct the model choice sets, random sampling is performed, and a subset of zones are randomly drawn from the full TAZ choice set (10 in our case study). A subset of zones is used to reduce computation time, which tends to increase with the size of the choice set. Given that we must repeat the estimation process many times to reach convergence, we use a fairly small choice set. However, the number could be increased to improve parameter efficiency in future applications of the approach. These 10 zones and the chosen zone together form the consideration set for each household observation.

3.3 Gibbs sampling as an alternative to gravity models

Anas (1983) outlines the aggregate approach to the joint determination of POR and POW. He proves the equivalence of a joint model of POR and POW and a doubly constrained gravity model, where the probability of an individual choosing to reside at location *i* and work at location *j* is given as follows.

$$P_{ij} = \frac{e^{\lambda_i + \lambda_j + \sum_k \lambda_k \overline{X}_{ijk}}}{\sum_I \sum_J e^{\lambda_i + \lambda_j + \sum_k \lambda_k \overline{X}_{ijk}}} \quad (3)$$

where λ_i, λ_j, and λ_k^A are Lagrange multipliers in the maximum entropy optimization system. The vector of $\overline{X}_{ijk}$ are *k* aggregate attributes of the impedance between POR location *i* and POW location *j*. The model relies on the following restrictive assumptions, which we aim to address:

1. All households have the same utility function.
2. All parameters are constant over the population.
3. The joint choice of POR and POW is the outcome of a single maximization by all households.

Gibbs sampling provides a statistical theoretical basis for the disaggregate choice approach. To our knowledge, this approach has not been applied in the literature to location choice. However, Farooq, Bierlaire, Hurtubia, & Floötteröd (2013) apply the technique in the related case of population synthesis. In this context, the objective is the synthesis of a model population from a sample representing a subset thereof. The probability of a sample household being allocated to a particular zone is then conditional on the attributes of the household and the known aggregate totals (i.e., market shares) for each zone. The disaggregate Gibbs approach has the following advantages over the gravity approach:

1. It is the outcome of the maximization of individual household utility functions.
2. Flexibility in the model formulation (e.g., could specify each location model as a nested logit, mixed logit, or other forms of the model).
3. There is no assumption of joint choice. Instead, the joint distribution evolves from the iterative estimation of conditional choice models. This follows the Bayesian paradigm of updating a prior distribution based on new information about the location likelihood function to develop a new posterior distribution.
4. Although we do not condition the present joint model on household attributes, the method is not restricted in its ability to condition the probability of POR and POW choices on household attributes.

Both the models described above are conditional choice models. That is, the residential location choice is made given a known work location, and the work location choice is made given a known residential location choice. Each choice is estimated for individual households, with the standard practice being to estimate a joint model using a nested logit structure. The aggregate matching of the joint distribution of residential and work location choice within a discrete choice context is an issue that has mostly confounded research efforts.

Gibbs sampling is a Bayesian statistical method by which the joint distribution can be approximated through a sampling of conditional distributions. Assuming we have two variables, x, and y, their conditional distributions are given by

$$f(x|y) = \frac{f(x,y)}{f(y)} \quad (4)$$

$$f(y|x) = \frac{f(x,y)}{f(x)} \quad (5)$$

where $f(x)$ is the marginal probability for variable x, $f(y)$ is the marginal probability for variable y, $f(x|y)$ is the conditional probability of y given y, $f(y|x)$ is the conditional probability of y given x, and $f(x,y)$ is the joint probability. The law of conditional expectations gives

$$E[E[X|Y]]=E(X), \quad (6)$$

also, therefore

$$E[f(y|X)]= \int f(y|x)f(x)dx=f(y). \quad (7)$$

Gibbs sampling can be used to obtain the marginals, f(x) and f(y), according to

$$f(x) \approx \frac{1}{R}\sum_{r=1}^{R} f\left(x|y_j^{r_j-1}\right). \quad (8)$$

where R is the number of random draws and r_j-1 is the value of the conditional distribution after the last iteration. In the case of joint location choice modelling where x is the POR and y is the POW, we would have that the marginal probability of $POR= f(x)=P_{i|j}$ is conditional upon the previous POW probability distribution $f\left(y_j^{r_j-1}\right) = P_{j|i@r-2}$.

The standard method of maximum likelihood estimation, used to determine each conditional model, provides point estimates of the joint likelihood function. The application of Gibbs sampling to these estimates provides the posterior joint distribution for both choices. Geman and Geman (1984) show that this process converges to the maximum of the joint distribution as a Markov Chain, with transitions approaching infinity. The joint distribution represents the equilibrium condition of the conditionals. It can be shown that this equilibrium is independent of the initial conditions.

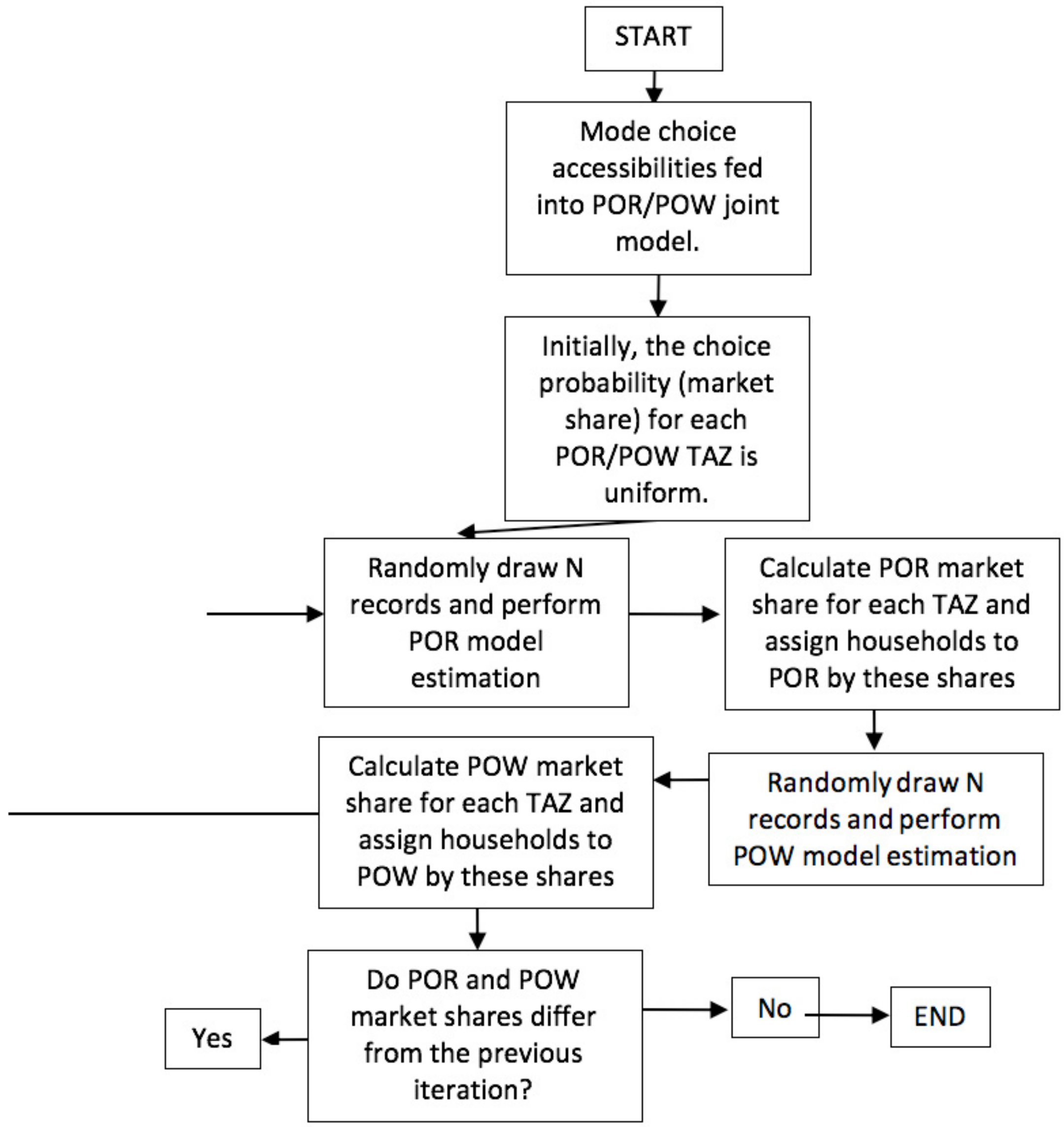

Figure 1. Gibbs sampling procedure

Note, in the current context; only generic alternatives are used in each of the POR and POW models. This removes the complication of the joint probability also being conditional upon the attributes of each household. It is possible to extend the Gibbs sampling method to this more complex context, but the current work considers the more straightforward problem.

Joint model estimation follows an iterative procedure, as outlined in Figure 1. The POR model is first run, and a random sample drawn of the estimated location probabilities for the chosen zone. The normalized probability for each zone provides a measure of the market share. That is, the proportion of the population choosing to reside in each zone. These market shares form the intermediate, posterior distributions that are used in the subsequent determination of the conditional POW model. The reverse process is then conducted, whereby the market shares from the POW model are used to allocate households to work zones for use in the next iteration of the POR model. This process is continued until the market shares for each zone (in both POR and POW models) do not change significantly between iterations (e.g., to 1 decimal places of precision). The outlined procedure constitutes a form of Bayesian updating, wherein the distributions from past iterations are used as conditional posteriors in the updating of the joint posterior distribution.

We know that the observed location choices are the outcome of an equilibrium procedure of balancing labor and dwelling supply and demand. In the absence of additional data and models, Gibbs sampling approximates this process through the equilibrium of conditional POR and POW distributions.

Gibbs simulation starts with the conditional residential location simulation via random draws. This results in the modeled market share of each residential zone. The next step is to conduct the work location choice model, which is conditional on the modeled residential zone market shares. Then, we model the residential location based on the modeled workplace market share. After sufficient iterations, the market share of residential locations and work locations do not change, which indicates converges to the joint distribution of residential location and work location.

4 Data for empirical application

Four datasets are used in this study: the 2011 National Capital Region (NCR) Household Origin-Destination Survey, an Enhanced Points of Interest (EPOI) dataset, a CanMap RouteLogistics Ontario dataset, and the Ottawa Neighborhood Study (ONS) - Neighborhood Boundaries.

4.1 2011 NCR household origin-destination survey

The NCR Household Origin-Destination Survey is compiled by the Ottawa Trans Committee approximately every five years (since the late 1970s). The latest survey conducted in 2011 is the primary database to estimate the proposed location choice models and mode choice model. It documents 159,311 trips made by 62,897 individuals from 25,374 households (Trans Committee, 2013). Of these trip records, 36,216 are commuting trips with both the origin and destination located within the City of Ottawa. The origin, destination, and home location of each trip are recorded according to the 2011 TRANS TAZ system, which comprises 17 districts (422 zones). The average sampling rate in Ottawa is 5.0% and the total population of the NCR is 1,233,800 in 510,000 households (Trans Committee, 2013).

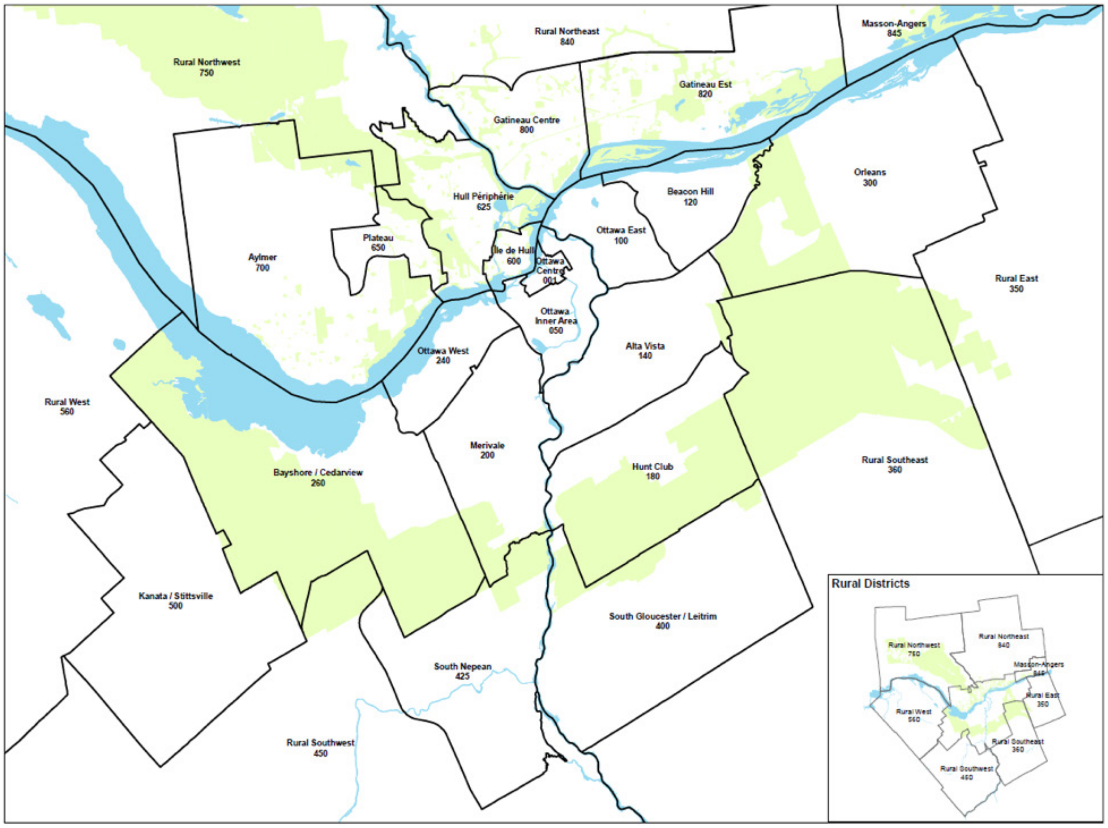

Figure 2. 2011 NCR survey area (Trans Committee, 2013)

4.2 Enhanced points of interest (EPOI)

A description of land uses an essential component in the modelling of POR choice. Hence, this paper also adapts the EPOI database developed by DMTI Spatial Inc. (2010), which comprises "over one million Canadian business and recreation points of interest" nationally. The dataset was edited in 2010, which is suitably close to the 2011 analysis year.

The Ontario EPOI database contains 513,680 points of interest. In general, these points are categorized as (DMTI Spatial Inc., 2010): 1) agriculture, forestry, and fishing, 2) mining, 3) construction, 4) manufacturing, 5) transportation, communications, electric, gas, and sanitary service, 6) wholesale trade, 7) retail trade, 8) finance, insurance and real estate, 9) services, 10) public administration, and 11) unclassified.

4.3 CanMap RouteLogistics Ontario

The EPOI database provides a count of business points but fails to capture the magnitude of business activity and other additional characteristics of land use. Hence, the CanMap RouteLogistics Ontario database is also used to provide additional land-use information.

This database is also developed by DMTI Spatial Inc. (2014) and uses 2014 as its analysis year. It provides a detailed roadway network and additional land-use attributes. The CanMap study provides the spatial distribution of land use as 1) commercial area, 2) government and institutional area, 3) open area, 4) park and recreational area, 5) residential area, 6) resource and industrial area, and 7) waterbody. These areas can be determined for each TAZ to give a better sense of land-use composition. Additional details about educational and health care facilities can also be drawn from this dataset.

4.4 Ottawa Neighborhood Study (ONS) - Neighborhood Boundaries

In addition to land-use attributes, individual household decisions are strongly influenced by population statistics. ONS Neighborhood Boundaries data were obtained from Open Data Ottawa (2017). This dataset is reviewed and updated every five years, and we use the 2011 version. Population statistics are provided for each neighborhood (district), defined by census tracts, physical and demographic similarities, and physical barriers.

4.5 Level-of-service (LOS) data

The LOS data is primarily drawn from the regional travel model (RTM) developed in the travel demand software EMME using the 2011 NCR Household Origin-Destination Survey. All EMME output files are summarized into a LOS lookup table, which lists all LOS attributes between each OD pair. Five commuting modes are defined in the RTM: auto, bus, rail mode, drive & bus, and drive & rail.

Both bus and rail are recorded as public transit in the 2011 NCR Household Origin-Destination Survey. We assume that individuals know the public transit service, and they will prefer the mode with the shorter total travel time. If both modes have the same total travel time, individuals are assumed to choose the rail mode, which is more comfortable and reliable.

The LOS attributes of the park & ride mode can be achieved from LOS attributes of either of the drive & bus or drive & rail modes, depending on total travel time. For simplification reason, we assume park & ride and carpool & ride modes share the same LOS attributes.

4.6 Trip data for mode choice model

A total of 36,216 commuting trips from the 2011 NCR Household Origin-Destination Survey are used to construct the mode choice model and obtain accessibility measures. Table 1 outlines the correspondence made between survey modes and model modes. There are 264 trip records (<1% of total trip records) with mode classified as "other." These trips are removed from the dataset for model estimation. Also, 8,693 trip records are excluded due to the absence of model variables and 248 trip records excluded because they were identified as "Poor Trips" (meaning missing critical information or having incorrectly coded locations) by the 2011 NCR Household Origin-Destination Survey. Thus, 27,011 trip records performed by 16,143 individuals from 9,423 households are utilized in model estimation. Such aggregation of modes is routine in applied transportation demand models (City of Edmonton, 2004; TMG, 2015). It is partially a function of network assignment being largely insensitive to auto passengers compared with taxi trips as both are assigned to the network as passenger trips beginning at the passenger origin and ending at his/her destination. The aggregate 4-step travel demand model, which remain the most common model used in practice, do not assign empty vehicle trips to the network – that is, taxis or buses that are travelling to pick up passengers. As such, we deem the aggregation of modes outlined in Table 1 as reasonable for the current application.

Table 1. Choice set classification for commuting modes

2011 NCR Household Origin-Destination Survey Record	*Corresponding Mode Choice in the Model (% of trips)*
Car driver	Auto Driving mode (55%)
Car passenger	Auto Passenger mode (15%)
Taxi	
School bus (yellow bus)	
Car passenger-Urban Transit (OC Transpo, STO, O-Train)	Carpool & Ride mode (1%)
Urban Transit (OC Transpo, STO, O-Train)-Car passenger	
Bicycle	Active mode (12%)
Walk (entire trip)	
Car driver-Urban Transit (OC Transpo, STO, O-Train)	Park & Ride mode (3%)
Urban Transit (OC Transpo, STO, O-Train)-Car driver	
Urban Transit (OC Transpo, STO, O-Train)	Transit mode (5%)
Other bus and minibus	
Other	Ignore (5%)

4.7 Household data for location choice models

The next part of the study is to develop two conditional location choice models for single-person households in Ottawa, including a residential location choice model given work location and a work location choice model given the residential location. Since only the single-person households are considered, the household records are reduced from 9423 to 979. For each household records, its Home Zone can be determined based on the corresponding Household ID, while the Work Zone is either the origin or destination zone of the commuting trip. After conducting the simple random sampling process, 11 sampled residential zones and 11 sampled work zones are given for each observation, so that the potential zonal attributes are identified accordingly.

All zonal variables are generic and have a fixed coefficient in the utility functions. The only sociodemographic attribute utilized in this study is population density in persons per km2. The accessibility attributes include home zone accessibility, work zone accessibility, and auto commuting distance. The auto commuting distance is a critical attribute linking the two location choice models. By knowing the POW zone for each household, the auto commuting distance can be calculated to capture the attractiveness of potential POR zones. Similarly, knowing the POR, commuting distances to alternative POW locations can be calculated and included in the POW model. A similar link exists between the measures of accessibility in each location choice model, derived from the mode choice model estimation. The land-use attributes include land-use mixing entropy, area proportion of various land-use types, and several service points. Service points can be any combination of public transit services, police/fire stations, educational facilities, and health care centers.

5 Results and discussion

5.1 Commuting mode choice model

Initially, all potential attributes are included in the commuting mode choice model. By adopting a backward elimination method, the variables are eliminated or adjusted if they are not significantly related to individual's commuting mode choice (assuming a 0.10 significance level). Model convergence is achieved after 18 iterations using a gradient tolerance of 1×10^(-7). The final log likelihood is -10863.13, and the adjusted rho-square is 0.624. The final mode choice model is presented in Table 2.

As shown in Table 2, all parameters have t-stat higher than the critical value, which indicates a significant relationship with mode choice. Also, all generic parameters (including IVTT, walk time, fare, and Euclidean travel distance) have the expected negative sign. Each commuting mode contains an alternative specific constant (ASC) in its utility function to capture any randomness and error, except for the Active mode, which is set as the reference mode. Overall, the values of all ρ2 indicate a good model fit, with the mode choice model exhibiting an excellent fit (Hensher, Rose, & Greene 2005). This implies that the proposed mode choice model can effectively capture the decision-making process while individuals select commuting modes.

In order to validate the proposed commuting mode choice model, bootstrapping is performed to compare the observed market share and the predicted market share. There are 27,011 observations in the trip database, so that bootstrapping is conducted 1000 times with a sample size of 27,011. The samples are drawn with replacement to ensure unbiased randomness. The final results of the bootstrap are presented in Table 3. Based on the results of bootstrapping, we can conclude that the proposed mode choice model fits the sample market shares. The maximum prediction error is roughly within ±1%, and the average prediction error is within ±0.5%. Thus, the mode choice model is used as a reliable source of zonal accessibilities in subsequent model analysis.

Table 2. Empirical commuting choice model parameter

Parameter name	Description	Parameter		
		Value	**Std. err.**	**t-stat**
ASC AD		-1.73	0.135	-12.78
ASC AP		-1.36	0.129	-10.59
ASC CR		-2.64	0.517	-5.10
ASC A		0.00	fixed	
ASC PR		-4.59	0.471	-9.74
ASC T		1.92	0.375	5.11
In-vehicle travel time (min)		-0.0101	0.00245	-4.12
Walk time (min)		-0.0446	0.0203	-2.19
Distance (km)		-0.318	0.017	-18.33
Transit fare ($)		-0.141	0.085	-1.65
Age greater than 55 (for AP)		0.332	0.082	4.05
Age less than 25 (for AD)		-0.601	0.091	-6.64
Age less than 25 (for AP)		1.47	0.088	16.73
Age less than 25 (for PR)		-1.70	0.20	-8.46
Dwelling type 1 (for T)	Dwelling Type	1.13	0.23	4.93
Dwelling type 2 (for CR)	1: Single-detached	-1.37	0.61	-2.23
Dwelling type 2 (for T)	2: Semi-detached	1.59	0.31	5.09
Dwelling type 3 (for AD)	3: Row/Townhouse	0.38	0.079	4.81
Dwelling type 3 (for AP)	4: Apartment/Condo - tenant	0.367	0.074	4.96
Dwelling type 3 (for T)	5: Apartment/Condo – owner	2.06	0.29	7.18
Dwelling type 4 (for AD)		0.187	0.095	1.97
Dwelling type 4 (for CR)		-1.5	0.55	-2.71
Dwelling type 4 (for T)		0.791	0.26	3.05
Household income 1 (for AD)	Income Level	0.6	0.17	3.62
Household income 1 (for AP)	1: $0 - $29,999	0.491	0.12	4.23
Household income 2 (for AD)	2: $30,000 - $59,999	0.368	0.096	3.83
Household income 3 (for AD)	3: $60,000 - $89,999	-0.186	0.086	-2.16
Household income_3 (for AP)	4: $90,000 - $119,999	-0.26	0.073	-3.54
Household income 4 (for AD)	5: $120,000 - $149,999	-0.23	0.082	-2.83
Household income 4 (for AP)	6: $150,000 - $179,999	-0.21	0.069	-3.03
Household income 5 (for AD)	7: $180,000 - $209,999	-0.23	0.083	-2.82
Household income 6 (for AD)	8: $210,000 and above	-0.70	0.11	-6.54
Household income 6 (for AP)	9: Decline	-0.43	0.10	-4.28
Household income 7 (for AD)		-0.52	0.11	-4.82
Household income 8 (for AD)		-0.73	0.12	-6.27
Household income 8 (for AP)		-0.40	0.11	-3.57
Household income 8 (for PR)		-0.89	0.34	-2.58
Household income 8 (for T)		-0.58	0.26	-2.20

Intrazonal dummy (for AD)		4.57	0.42	10.89
Intrazonal dummy (for AP)		4.25	0.41	10.45
Intrazonal dummy (for A)		5.70	0.41	14.00
Peak hour departure (for AD)		-0.16	0.05	-3.11
Gender (for AP)		-0.58	0.046	-12.56
Gender (for CR)		-0.90	0.21	-4.26
Gender (for PR)		-0.30	0.15	-2.01
Gender (for T)		-0.34	0.12	-2.75
Transit pass (for AD)		-1.26	0.13	-10.04
Transit pass (for AP)		-0.55	0.12	-4.68
Transit pass (for CR)		2.76	0.26	10.7
Transit pass (for PR)		3.67	0.23	16.02
Transit pass (for T)		3.27	0.19	17.09
Driver license (for AP)		-1.33	0.09	-15.24
Driver license (for CR)		-1.11	0.26	-4.31
Driver license (for T)		-1.23	0.17	-7.20
# of persons in HH (for AD)		-0.13	0.023	-5.74
# of persons in HH (for AP)		0.047	0.023	2.10
# of persons in HH (for PR)		-0.22	0.070	-3.17
# of vehicles in HH (for AD)		1.43	0.050	28.79
# of vehicles in HH (for AP)		0.51	0.043	11.92
# of vehicles in HH (for CR)		0.29	0.144	2.01
# of vehicles in HH (for PR)		1.43	0.1	14.26
# of vehicles in HH (for T)		0.46	0.081	5.62
# of daily trips (for AD)		0.16	0.024	6.49
# of daily trips (for CR)		-0.25	0.12	-2.00
Number of individuals		16143		
Number of observations		27011		
Null log-likelihood		-29030		
Initial log-likelihood		-29030		
Final log-likelihood at converge		-10863		
ρ^2 value		0.63		
Adjusted ρ^2 value		0.62		

Table 3. Bootstrapping for mode choice model

Mode	**Sample Share**	**Mean Prediction Error**	**Max Over Prediction**	**Max Under Prediction**
Auto driving	40.98%	-0.26%	0.30%	-0.95%
Auto passenger	18.83%	-0.14%	0.44%	-0.68%
Carpool & ride	0.50%	-0.03%	0.10%	-0.16%
Active	12.77%	0.36%	0.88%	-0.20%
Park & ride	1.23%	0.06%	0.23%	-0.11%
Transit	25.69%	0.01%	0.26%	-0.22%

5.2 Joint POR and POW model

Upon validating the mode choice model, the modal accessibility can be calculated based on the logsum formula given in equation 2. The resultant accessibility is then utilized to support the development of the two conditional location choice models. Overall, the individual models were found to capture the conditional choice behavior of single-person households. The proposed joint model showed an excellent level of fit and converged after only 18 iterations of the joint Gibbs sampling procedure. The procedure took approximately 25 minutes to converge. Higher sampling rates, beyond 10 alternatives, are increasingly common in the location choice literature to reduce standard errors. However, a larger choice set would increase computation time, which is a particularly pertinent consideration given that the method requires multiple iterations of the estimation procedure. The model could also be improved using additional data sources and more advanced model specifications.

The residential location choice model is a conditional MNL model that estimates POR probabilities by TAZ, conditional on a known work location choice. The final residential location choice model is presented in Table 4.

Table 4. POR location choice model

	Parameter		
Name	**Value**	**Std. err.**	**t-stat**
Zonal population density (per/km^2)	0.000106	0.000021	5.10
# of service points	0.039	0.013	3.06
Land-use mixing entropy	0.86	0.30	2.89
Recreational area (% of TAZ)	1.57	0.50	3.17
Residential area (% of TAZ)	1.68	0.16	10.32
Water body area (% of TAZ)	2.02	0.33	6.11
Home accessibility	0.42	0.040	10.51
Auto commuting distance (km)	-0.076	0.0056	-13.59
# of observations	979		
Null log-likelihood	-2348		
Initial log-likelihood	-2348		
Final log-likelihood at converge	-1884		
ρ^2 value	0.20		
Adjusted ρ^2 value	0.19		

According to Table 4, all attributes are generic variables. As expected, a higher population density is correlated with a higher probability of choice. Several service points are defined in the model as the summation of the number of public transit services, police/fire stations, educational facilities, and health care centers. The positive coefficient for the number of service points indicates that zones with more service points have a higher probability of being selected by a household. Moreover, zones characterized by higher percentages of the residential area, recreational area (like parks), and waterbody typically have a better living condition; thus, they are preferred as home locations. A higher level of land-use mixing provides many service opportunities and has a significant impact on POR choice. Lastly, the model suggests a preference for zones with higher accessibilities and shorter auto commuting distances.

Although all attributes have reasonable coefficients, several unobserved factors are not explained in the proposed model. Further research should be conducted to enhance the model accuracy by including additional attributes or adjusting the model structure.

Similar to the POR model, the POW model is also a conditional MNL model, which estimates work location probabilities, conditional on a known residential location choice. The final POW model is presented in Table 5.

Table 5. POW location choice model

Name	Parameter		
	Value	**Std. err.**	**t-stat**
Zonal population density	0.00011	0.000018	6.22
# of service points	0.056	0.010	5.39
Recreational area (% of TAZ)	-3.44	0.59	-5.82
Residential area (% of TAZ)	-1.88	0.13	-14.53
Industrial area (% of TAZ)	-0.82	0.21	-3.95
Open area (% of TAZ)	-1.05	0.19	-5.41
Work accessibility	0.087	0.040	2.16
Auto commuting distance (km)	-0.094	0.0058	-16.25
# of observations	979		
Null log-likelihood	-2348		
Initial log-likelihood	-2348		
Final log-likelihood at converge	-1896		
ρ^2 value	0.19		
Adjusted ρ^2 value	0.19		

Similar to the residential location choice model, work zones are generally preferred that have higher population densities and more significant numbers of service points. In this model, only the public transit services and police/fire stations are considered as the service point. Zones characterized by higher percentages of the recreation area, residential area, industrial area, and open area are less preferred as work locations. Modal accessibility also plays an essential role in work location choice. Typically, people tend to prefer a work location with higher work accessibility and/or a shorter auto commuting distance. Further research is required to capture the decision-making process of POW location choice.

Based on the two conditional location choice models, the Gibbs sampling method (following the steps presented in Figure 1) is performed to achieve a joint distribution of residential location choice and work location choice for single-person households in Ottawa. The final modeled POR & POW market shares for each zone are compared with the actual market shares to evaluate the model accuracy.

Table 6. Location choice models validation

Location	*Absolute Mean Prediction Error*	*Max Prediction Error*	*Min Prediction Error*
Residential location	0.19%	1.59%	-0.83%
Work Location	0.22%	2.25%	-1.58%

Figure 2 summarize the critical error values. Errors are estimated by taking a series of random, bootstrapped, samples, and making a comparison with the original sample. This provides a measure of the prediction error in the case that the population distribution of locations (residential and work) differing from the sample distribution. The residential location choice has a higher level of accuracy with regards to work location choice in the joint model. The average prediction errors of both location choices are relatively low. This indicates that the proposed joint model generally captures location choice behavior. Maximum and minimum errors are well within the bounds of acceptable values and support the finding that the method provides a strong representation of the joint distribution of POR and POW location choices.

6 Conclusions and future work

This paper presented a new method for the joint modelling of residential and work location choice and a case study using single-person households in Ottawa. Two conditional location choice models were first developed by following the MNL specification and considering socioeconomic data, zonal land-use information, and modal accessibility derived from a mode choice model. Gibbs sampling was tested as a method for approximating the joint distribution of POR and POW location choice. The method was based on the combination of disaggregate choice models and could be readily applied to a wide range of choice model specifications.

The Gibbs sampling approach presents a theoretical basis for the development of a joint POR and POW distribution that is founded in disaggregate behavioral models. We argue that this approach is superior to the traditional gravity approach, which is based on aggregate statistics and has a weak basis in microeconomic choice theory. The proposed approach has the flexibility to be extended in several directions, as outlined below.

Comparison of the Gibbs sampling approach with a traditional approach gravity model would be a useful extension. This would require the estimation of additional models and the development of a metric for comparison between these quite different models. We leave this for future work.

All the choice models used in the study follow the MNL concept, which assumes all alternatives are independent and irrelevant. This simplifies model development but fails to capture the potential spatial correlation between alternatives. Also, the two conditional location choice models only include zonal attributes. This makes the application of Gibbs sampling straightforward but ignores heterogeneity in household preferences. The population could be segmented by socioeconomic category, which would introduce an additional dimension to the conditioning process. This would mean that each iteration of the Gibbs sampling process would include the additional step of conditioning against the market shares for each segment.

Lastly, the choice models are only developed for single-person households, which represents only a portion of the total households in Ottawa. A household may have one residential location, but multiple work locations, if there is more than one person in the household. The model could be extended to include these households by conditioning on both work location choices. This extension would require the estimation of separate POW models for each worker in the household, such that POR and POW for worker 1 would become conditional upon the work location of worker 2, 3, etc.

References

Abraham, J., & Hunt, J. D. (1997). Specification and estimation of nested logit model of home, workplaces, and commuter mode choices by multiple-worker households. *Transportation Research Record, 1606,* 17–24. doi.org/10.3141/1606-03

Alonso, W. (1964). *Location and land use.* Harvard University Press.

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological, 17*(1), 13–23.

Bhat, C. R., & Guo, J. Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B, 41*, 506–526. doi.org/10.1016/j.trb.2005.12.005

Chen, J., Chen, C., & Timmermans, H. J. P. (2008). Accessibility trade-offs in household residential location decisions. *Transportation Research Record, 2077*(1), 71–79.

City of Edmonton. (2004). Segment 16: Adult worker home to work trips (internal government report). Edmonton, Alberta: City of Edmonton.

Clark, W. A. V. & Withers, S. D. (1999). Changing jobs and changing houses: Mobility outcomes of employment transitions. *Journal of Regional Science, 39*(4), 653–673.

DMTI Spatial Inc. (2010). Enhanced points of interest. Retrieved from https://mdl.library.utoronto.ca/collections/geospatial-data/enhanced-points-interest-9

DMTI Spatial Inc. (2014). CanMap RouteLogistics Ontario. Retrieved from https://mdl.library.utoronto.ca/collections/geospatial-data/canmap-routelogistics-%0Aontario-5

Farley, J. E. (1995). Race still matters: The minimal role of income and housing cost as causes of housing segregation in St. Louis, 1990. *Urban Affairs Reviews, 31*(2), 244–254.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B, 58*, 243–263. Retrieved from https://ac.els-cdn.com/S0191261513001720/1-s2.0-S0191261513001720-main.pdf?_tid=d7c4c0be-557e-4222-b613-b7be0c21bbd9&acdnat=1536849457_1982f341aacaaf3cbfd9b126c13eeb67

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721–741. Retrieved from http://image.diku.dk/imagecanon/material/GemanPAMI84.pdf

Heldt, B., Gade, K., & Heinrichs, D. (2016). Determination of attributes reflecting household preferences in location choice models, *Transportation Research Procedia, 19*, 119–134. https://doi.org/10.1016/j.trpro.2016.12.073

Hensher, D., Rose, J., & Greene, W. (2005). *Applied choice analysis: A primer.* Cambridge, UK: Cambridge University Press.

Ibeas, Á., Cordera, R., Dell'olio, L., & Coppola, P. (2013). Modelling the spatial interactions between workplace and residential location. *Transportation Research Part A, 49*, 110–122. doi.org/10.1016/j.tra.2013.01.008

Jiao, P., Liu, M., & Guo, J. (2015). Integrated model of joint residence-workplace location choice and commute behavior using latent class and mixed logit methods. *Discrete Dynamics in Nature and Society*. doi.org/10.1155/2015/251781

Kryvobokov, M. (2015). A two-level regional approach to residential location choice model. *Letters in Spatial and Resource Sciences, 8*(2), 181–196.doi.org/10.1007/s12076-014-0126-y

Lowry, I. S. (1964). *A model of metropolis.* Retrieved from https://www.rand.org/content/dam/rand/pubs/research_memoranda/2006/RM4035.pdf

Manheim, M. (1979). Fundamentals of Transportation Systems Analysis. Cambridge, MA: MIT Press.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record, 673*(673), 72–77. Retrieved from http://onlinepubs.trb.org/Onlinepubs/trr/1978/673/673-012.pdf

Open Data Ottawa. (2017). ONS Neighborhood Boundaries. Retrieved from http://data.ottawa.ca/dataset/neighbourhoodboundaries

Rivera, M. A. I., & Tiglao, N. C. C. (2005). Modeling residential location choice: Workplace location choice and mode choice of two-worker households in metro Manila. *Proceedings of the Eastern Asia Society for Transportation Studies, 5*, 1167–1178. Retrieved from https://s3.amazonaws.com/academia.edu.documents/34860336/1167.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1523630977&Signature=6iiplkOiGJCIpYZ5n9NoT8CBBQY%3D&response-content-disposition=inline%3B filename%3DMODELING_RESIDENTIAL_LOCATION_CHOICE_WOR.pd

Schirmer, P. M., Van Eggermond, M. A. B., & Axhausen, K. W. (2014). The role of location in residential location choice models a review of literature. *Journal of Transport and Land Use, 7*(2), 3–21. doi.org/10.5198/jtlu.v7i2.740

Siegel, J. (1975). Intra-metropolitan migration: A simultaneous model of employment and residential location of white and black households. *Journal of Urban Economics, 2*(1), 29–47. doi.org/10.1016/0094-1190(75)90038-8

Simpson, W. (1987). Workplace location, residential location, and urban commuting. *Urban Studies, 24*(2), 119–128. Retrieved from http://journals.sagepub.com/doi/abs/10.1080/713703872

Timmermans, H., Van Noortwijk, L., Oppewal, H., & Van Der Waerden, P. (1996). Modeling constrained choice behavior in regulated housing markets by means of discrete choice experiments and universal logit models: An application to the residential choice behavior of divorcees. *Environment and Planning A, 28*(6), 1095–1112. doi.org/10.1068/a281095

TMG. (2015). GTAModel 4.0: Design, background, and calibration (technical report). Toronto: University of Toronto Travel Modelling Group.

Tran, M. T., Zhang, J., Chikaraishi, M., & Fujiwara, A. (2016). A joint analysis of residential location, work location and commuting mode choices in Hanoi, Vietnam. *Journal of Transport Geography, 54*, 181–193. doi.org/10.1016/j.jtrangeo.2016.06.003

Trans Committee. (2013). *2011 NCR Household Origin-Destination Survey* (technical report). Ottawa: Trans Committee.

Waddell, P., Bhat, C., Eluru, N., Wang, L., & Pendyala, R. M. (2007). Modeling interdependence in household residence and workplace choices. *Transportation Research Record: Journal of the Transportation Research Board, 2003*(1), 84–92. doi.org/10.3141/2003-11

Wartenberg, C. M., (Trans.) & Hall, P. (Ed.) (1996). *Von Thünen's isolated state: An English translation of der isolierte staat*, by J. H. Von Thünen (1826). Oxford, UK: Peragamon Press Ltd.

Wilson, A. G. (1970). *Entropy in urban and regional model.* New York: Routledge.