MICROSIMULATING THE SPATIAL DISTRIBUTION OF COMMERCIAL VEHICLES BY THE LOCATION OF THEIR OWNED ESTABLISHMENTS

Aya Hagag and, University of Windsor
Hanna Maoh, Cross-Border Institute, University of Windsor

Introduction

Freight movement is a major process contributing to economic growth and prosperity. It is a fairly dynamic process due to the rapid changes arising from complex supply chain structures, logistics and technological advancements. The rapid changes and growth in commercial vehicle movements in recent years have led to a surge of studies that focus on understanding the processes giving rise to these movements. To date, little has been done to study and model vehicle ownership of business establishments. However, the number of owned commercial vehicles is a significant factor that explains the number of generated commercial trips (Madar, 2014). The lack of studies on the topic is believed to be due to the absence of detailed commercial vehicle data. According to the literature, most of the existing efforts to collect detailed commercial travel data resulted in a low response rate (Samimi et al., 2012). Private establishments usually vacillate to share information related to their business freight and/or transportation activities.

This paper addresses the problem of data scarcity by employing synthetic population techniques to microsimulate the number of commercial vehicles owned by all individual business establishments that engage in delivering goods or services in the Windsor Census Metropolitan Area (CMA). The case presented here uses the combinatorial optimization technique (CO) to synthesize the number of commercial vehicles owned by business establishments that engage in commercial travel activities (i.e. delivering goods or services). The CO uses a micro-sample extracted from a Business Establishment Commercial Travel Survey (BECTS) that was conducted by the University of Windsor in 2013. The micro-sample, which contains data on 171 individual business establishments, provides information that characterize the surveyed establishments (e.g. location, industry type, employment size, and number of owned vehicles) and generated commercial trips.

The procedure we followed consists of three consecutive steps: 1) identify all the establishments that engage in shipping goods or services in the CMA, 2) apply the Simulated Annealing Procedure to solve the Combinatorial Optimization (CO) problem of assigning commercial vehicles to business establishments, and 3) validate the synthesized records using an external dataset that was not initially used in the population synthesis procedure. The offered procedure to microsimulate the spatial distribution of commercial vehicles within an urban area is novel and has not been attempted in the past. Also, the analysis provide a basis for evaluating commercial data sources such as the R. L. Polk and Co. vehicle registry data and its potential in analyzing commercial vehicles in other areas where micro-datasets are not available.

The remainder of this paper is organized as follows. The next section provides an overview on synthetic population techniques. Then the following section highlights method of analysis and the data used. This will be followed by a section to discuss the results with a brief conclusion.

Overview

Population synthesis has been widely used in transportation research to develop activity-based microsimulation models and disaggregate land use models to address several policy relevant issues. It has been also incorporated as part of comprehensive socioeconomic and demographic model system and econometric microsimulator for urban systems (Eluru et al., 2008; Pendyala et al., 2012). In particular, population synthesis is used as a preliminary step to construct the microdata set that represent the characteristics of the agents used in a microsimulation. For these models the decision agents to be microsimulated may include individual, households, dwelling or establishment populations (Ryan et al. 2009). In general, to develop such models, substantial amount of disaggregate data is required. However, almost all the available data are anonymized, geographically diluted, or generalized to specific spatial areas to protect the privacy of individuals (Voas & Williamson, 2000). In fact, when detailed information about individual demographics exists, their spatial location is diluted to maintain confidentiality (Frick et al., 2004). Hence, to overcome the tedious effort and long-time associated with data collection and the availability of data, synthetic populations for transportation research has received more attention over the past two decades (Ryan et al., 2009; Arentze et al., 2007).

The data used to create synthetic populations are usually; aggregate zonal population data and disaggregate sample data. The former are available in terms of summary cross-tabulations of demographics represented as one-way, two-way, or multiway cross-tabulations that describe the joint aggregate distribution of relevant demographic and socioeconomic variables at the zonal level (Arentze et al., 2007; Ryan et al., 2009). For example, the Summary Files (SFs) used in the United States and the Small Area Statistics (SAS) file used in the United Kingdom (Beckman, Baggerly, & McKay, 1996; Voas & Williamson, 2000). On the other hand, the disaggregate data represents a sample of individuals with information about the characteristics of each individual in it, excluding addresses and unique identifier. Examples are the Public-Use Microdata Samples (PUMS) in U.S. and the Sample of Anonymized Records (SAR) in UK (Beckman et al., 1996; Voas & Williamson, 2000).

A wide variety of techniques exist in the literature to estimate detailed microdata such as stratified sampling, geodemographic profiling, data fusion, data merging, reweighting, iterative proportional fitting synthetic reconstruction (IPFSR) and combinatorial optimization (CO) (Huang & Williamson, 2001). However, both IPFSR and CO have been identified as the most dominant techniques in the recreation of synthetic population microdata, although the IPFSR have been used more widely (Ryan et al., 2009). The synthetic reconstruction techniques presented by Wilson and Pownall (1976), current state-of-the-art, make use of the iterative proportional fitting (IPF) technique to create multiway tables of proportions that are consistent with the aggregate data totals. Then synthetic population of households is drawn from the microdata to match the proportions in the estimated multiway table (Beckman et al., 1996). In fact, with the variations in the types of input and how certain synthesis routines are carried out, a wide variety of the current population synthesizers involve the use of the IPF technique. Examples are the work done by (Arentze et al., 2007; Auld et al., 2009; Guo & Bhat, 2007; Martin Frick et al., 2004; Simpson & Tranmer, 2005).

Some of the more recent studies listed above attempted to address some of the shortcomings in the method presented in Beckman et al. (1996). For instance, the method in Beckman et al. (1996) does not address the zero-cell-value problem and the inability to control for statistical distributions of both household- and individual-level attributes. Guo & Bhat (2007) introduces a new population synthesis procedure that allows the user to adjust the choice of control variables and the class definition of these variables at run time to avoid initial incorrect value of zero in the contingency tables. That is when a specific demographic group in the population is represented in the aggregate data but not represented in the sample of the disaggregate data. Auld et al. (2009) also addressed this problem and developed a routine that allows for the aggregation of control variable categories during execution at the sub-regional

level based on a user-controlled aggregation threshold parameter. The inability for controlling attributes on multiple analysis levels in a population synthesis program was also tackled by (Auld et al., 2010). Their methodology is implemented within a population synthesizer to allow multiple-level synthetic populations such as household- and person-level, establishment and employee or household and vehicle estimation.

The combinatorial optimization (CO) technique has been used as an alternative to the IPFSR method. This iterative technique is simpler as it synthesizes the population on a zone by zone basis. It starts by choosing a random set of households from a sample microdata with a replacement, then assessing the effects of replacing one of the selected households. Accordingly, the replacement will be made only if the swap improves the fit. Williamson et al. (1998) presented different combinatorial optimization techniques to produce enhanced small-area microdata estimates. These included hill climbing approach, simulated annealing approach and genetic algorithm approach. It was found that the simulated annealing is the best performing approach. Later, Voas & Williamson (2000) assessed the implementation of the CO technique and proposed a sequential fitting procedure to further improve the quality of the synthetic microdata. Given their popularity, the CO and IPFSR techniques have been compared to find which would achieve better results. According to Ryan et al. (2009), both techniques are capable of producing synthetic microdata that fit constraining tables extremely well. However, the CO technique is deemed superior for its ability to produce more accurate results with the variation in tabulation details and input sample size.

Study Area and Data Description

The analysis in this paper is focused on the Windsor Census Metropolitan Area (CMA) located on the south shore of the Detroit River and Lake St. Clair. Windsor occupies approximately 1022.31 square kilometers of Canada's land (Government of Canada, 2012). According to the 2011 National Household Survey, the CMA housed 323,342 people, 126,845 and 123,305 jobs in the year 2011. The data used for this analysis was obtained from three sources. The first dataset was acquired from InfoCanada and consists of all 10,771 establishments registered in the Windsor CMA in 2013. The attributes provided for each establishment includes; the InfoUSA (IUSA) code designation for each establishment, contact information (i.e. telephone number and full address), employment size and industry classification according to the North American Industry Classification System (NAICS) six-digit code and the Standard Industry Classification (SIC) code. Also, all the establishments were geocoded, to get their zonal location within the Windsor CMA using the street addresses and postal codes. This procedure performed in ArcMap GIS software. Of the 10,771 establishments, 9,939 were geocoded, representing (92%) of the addresses. Figure 1 illustrates the location of the geocoded establishments within the Windsor CMA, where establishments mainly operate from 63 zones rather than all the 73 zones comprising the study area.

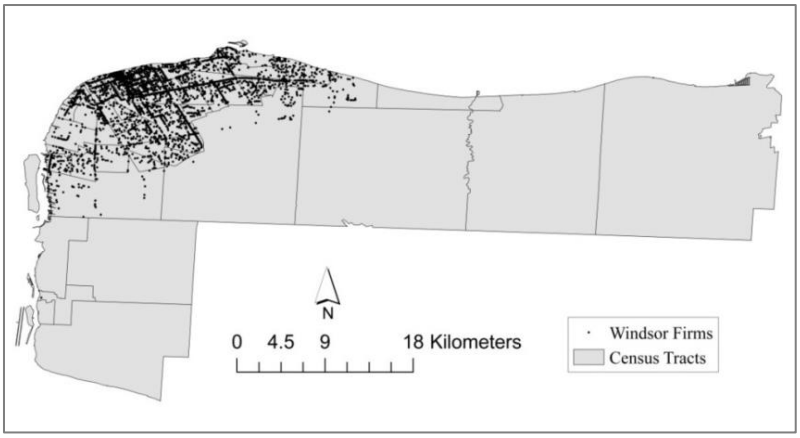


Figure 1. Establishment location in the Windsor census metropolitan area, 2013

The second data source is the Business Establishment Commercial Travel Survey that was conducted by the University of Windsor in 2013. This survey was undertaken in two stages; telephone based survey and web based survey. During stage one, businesses were contacted through a short telephone survey to know whether they engage in shipping and/or receiving goods and services and if yes, whether the establishment would partake in the online survey. During stage two, a link to the web survey was provided for those establishments that agreed to participate in the online survey. Recruited establishments were given 40 questions about the general establishment characteristics and inbound and outbound commercial activities on the day the survey. The data collection process is described in details in (Madar, 2014). Out the establishments who provided contact information to receive the survey, 171 completed the survey. However, only 161 establishments completed the questions that pertain to the attributes required to create the microsample that will be used for the synthesizing process as will be discussed in the following section.

The third dataset is the Polk data for the year 2013 acquired from R. L. Polk and Co. This dataset consists of all registered CVs at the census tract level in the CMA, where commercial vehicles are classified into eight different classes according to their Gross Vehicle Weight (GVW) and geo-referenced to the census tract level. Also, this dataset provide information with regard the model, make and year of each registered vehicle. Given this information, the counts of registered vehicles in each zone by their year, model and make were considered. The exploration indicated the existence of large counts of the same vehicle spanning from 23 to 220 vehicles per unique year, make and model in zone 5590033.0. For instance, this is the only zone that has a count of 220 vehicles of 2013 Chrysler 200. The majority of these vehicles are produced by Chrysler and is in the zone where the headquarters of Chrysler Canada is located. Accordingly, this zone was dropped from the analysis since it is most likely that these vehicles are commercial fleets registered to a particular headquarter and as such do not operate from within the zone. By comparison, the count of vehicles pertaining to a unique year, make and model class in any of the other zones was fairly low with an average of 5 vehicles. In fact, the average among 59 zones was 3 vehicles and only 6 zones had counts of 30, 29, 24, 22, 18 and 13 vehicles per unique year, make and model class, respectively.

Method of Analysis

CO technique is used to create a disaggregate list of establishments with attributes, when aggregated conform to a predefined zonal totals. For the case presented here, the simulated annealing approach was used to reach to the optimal solution where swaps are accepted even if they lead to a moderate degradation in performance, in order to allow the algorithm to backtrack from suboptimal solutions. The annealing parameters used to synthesize the population are as follows: Initial Temperature = 400, alpha (cooling parameters) = 0.999, final temperature = 0.001 and number of swaps = 1. The goodness of fit is assessed using the Relative Sum of Squared Z-Scores (RSSZ) documents in Ryan et al. (2009). For more information on the simulated annealing approach in the context of CO method see Williamson et al. (1998). This process was repeated 10 times to confirm the consistency of these populations. The program to execute this process was written in C# and is called the Combinatorial Optimizer program.

To generate the aggregate cross-tabulations, as an input to synthesizing process, first the list of individual business establishments that engaged in shipping and receiving had to be determined. This is performed using a participation quotient (PQ) approach. The latter is based on the number of establishments that reported to engage in shipping and/or receiving goods. The PQ index is calculated as the ratio of (F_n^s) the number of businesses that reported to engage in shipping or receiving within a specific industry category *n*, to the total number of establishments who reported engaging in shipping or receiving, divided by the ratio of (F_n) the number of businesses in the respective industry to the number of establishments in the entire population of establishments as illustrated in Equation 1.

$$PQ_S = \frac{F_n^S / \sum_n F_n^S}{F_n / \sum_n F_n} \quad (1)$$

This method is inspired by the Location Quotient technique, a well-established method that has been used in economics and economic geography (Miller and Blair, 2009). Accordingly, all establishments that belong to industries with participation quotient greater than 0.7 was kept for the analysis, while establishments that belong to participation quotient less than 0.7 were dropped except for those establishments who were originally surveyed. Out of 9,939 establishments 3,478 were found to engage in shipping and/ or receiving goods and/or services (35%). The threshold value of 0.7 was chosen after considering different values, where 0.7 achieved the most reliable results as will be discussed in the results section.

To create representative aggregate cross-tabulations for the total establishment population, the attributes considered for each establishment are: census tract ID, number of employees and a two-digit SIC classification. The list of the 10 (2-digit) SIC categories to which the establishments belonged includes: Agriculture, Forestry, Fishing; Mining; Construction; Manufacturing; Transportation & Public Utilities; Wholesale Trade; Retail Trade; Finance, Insurance & Real Estate; Services; and Public Administration. On the other hand, the number of employees attribute was reclassified into ten categories representing the following discrete employment ranges: 1 – 10; 11 – 20; 21 – 30; 31 – 40; 41 – 50; 51 – 60; 61 – 70; 71 – 80; 81 – 90; and greater than 90 employees. Additionally, establishments with no employment values were dropped (i.e. 200 establishments). Cross-tabulations for the CO were derived based on the above industry and employment size categories. Next, a micro-sample of 162 establishments were extracted from the survey responses were information about the 2-digit SIC industry classification and employment category and the number of commercial vehicles owned per vehicle type are stated. The CO method was then used to create a list of 3478 establishments, where each establishment in the list has attributes that represent the employment size class and corresponding 2digit- SIC category. Each synthesized establishment was linked directly to an establishment from the micro-sample, and as such the values for the number of vehicles owned per vehicle type and establishment were assigned.

Results and Discussion

For each synthetic population generated, the total number of vehicles at each zone is calculated by summing the number of vehicles owned by each establishment in the zone. Then, the results are validated by comparing the zonal aggregates to the Polk Data. For this study, the R-squared value was used to assess the correlation between the synthesized and the actual number of vehicles per zone. Consequently, for the different PQ cut-off values tested, a set of 10 populations were generated. After calculating the zonal aggregates for each synthesized population, an average was estimated from the 10 different populations and compared to the Polk data. Table 1 presents the R-squared achieved, where, PQ greater than 0.7, is associated with the highest fit with R-squared value of 0.88.

Hence, a total of 3,478 establishments are determined to engage in commercial vehicle activities, representing 35% of all establishments in the Windsor CMA. Although a slightly larger share was reported for the Greater Toronto Area (43%) (MITL, 2010) and Edmonton (49%), a 35% of establishments engaging in commercial activities for Windsor is an acceptable share given its overall size when compared to mega regions like Toronto and Edmonton. The variations between the total numbers of vehicles estimated from each run were insignificant as discerned by the descriptive statistics shown in Table 2. The total number of vehicles can also be found in graphical form in Figure 2. The estimated standard deviation, variance and range indicate the results achieved from the 10 runs are consistent with each other.

These establishments are then classified into 5 major industrial groups, following the categorization used in (Hunt & Stefan, 2007). That is, retail, basic industry, services, wholesale and transportation accounting for 55%, 19%, 12%, 8% and 6% of establishments that engage in commercial activities, respectively. Intuitively, not all industries will be dependent on the same types of vehicles for their business transportation activities. For instance, the basic industry is more likely to own heavy duty trucks while establishments from the services sector are more prone to own and use small cars and light commercial trucks. Therefore, the diversity in the clustering pattern of these industries over space would result in varied spatial distribution of CVs in the various zones across the city.

Table 1. Results of Comparisons between Synthetic and the Actual Number of vehicles, using the R^2

	Number of Establishments	Percent from Total Population	R^2
PQ > 0.2	5906	59%	0.83
PQ > 0.3	5048	51%	0.86
PQ > 0.4	4452	45%	0.86
PQ > 0.5	4148	42%	0.86
PQ > 0.6	3744	38%	0.87
PQ > 0.7	**3478**	**35%**	**0.88**
PQ > 0.8	3022	30%	0.87
PQ > 1.0	2795	28%	0.87

Table 2. Total Number of Vehicles Summary Statistics

Summary statistics	CO
Mean	10005
Median	9,951
Standard Deviation	256.412
Sample Variance	65749.56
Kurtosis	-0.3569
Skewness	0.231
Range	865
Minimum	9591
Maximum	10456
Count	10

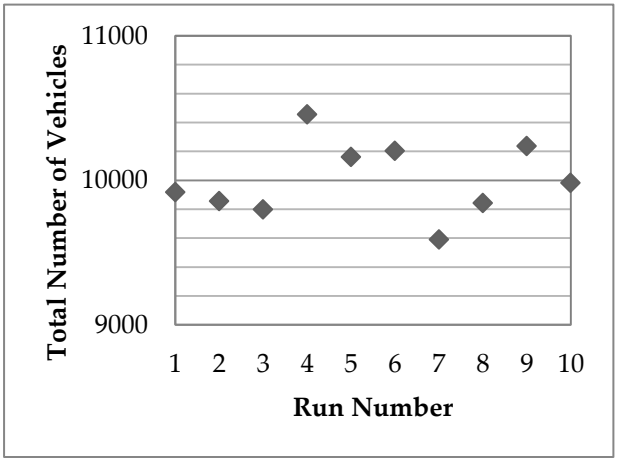


Figure 2. Total Number of Vehicles Comparison for 10 Establishment Population Synthesized

Figure 3 provides a comparison between the synthesized zonal aggregates versus the Polk data, while Figure 4 highlights the spatial distribution of these vehicles. The synthesized and the actual number of vehicles per zone have similar spatial distributions. As the trend in Figure 3 suggests, the Polk totals are very close to the synthesized totals per zone. Accordingly, for the zones with Polk totals less than the synthesized totals, the Polk totals can be used if one is to make use of the Polk data in future analysis. However, if the Polk zonal totals were greater than the synthesized totals, the synthesized totals are used. With this adjustment, the use of Polk data to examine the commercial vehicle ownership location to study the spatial prevalence of the various types of commercial vehicles in a given zone becomes doable. This is the case since the synthesized population is a true representation of the vehicles registered and operated from the same zone. This eliminates any concerns of vehicles registered to establishments in a census tract but operating from another location, an inherited problem in any acquired Polk dataset.

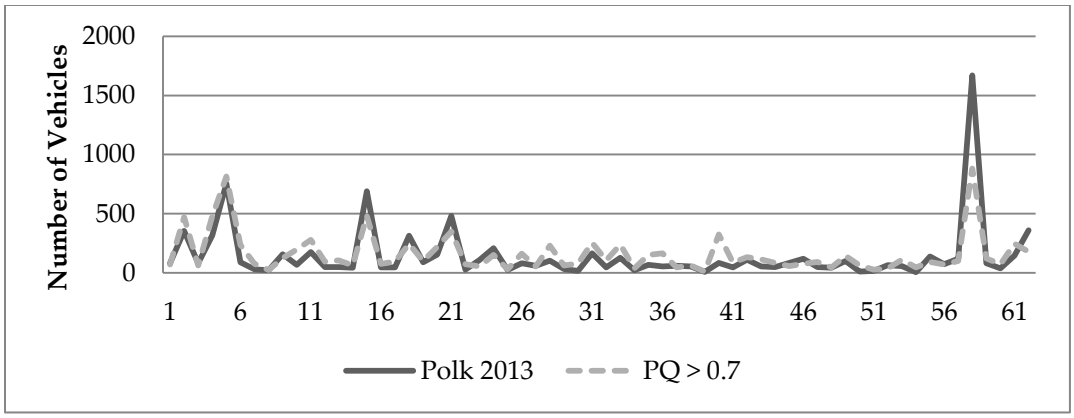
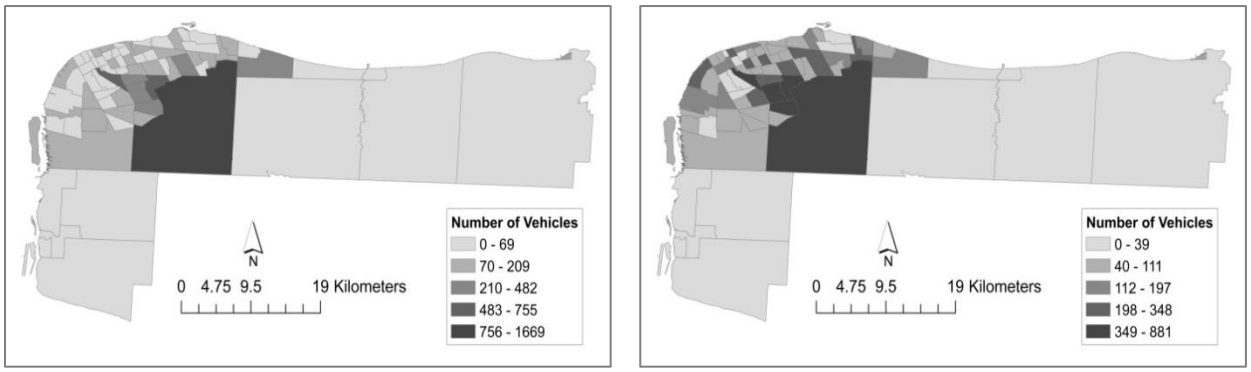


Figure 3. Comparison of the zonal aggregates of the synthesized population against the Polk Data



(a) (b)

Figure 4. Spatial distribution of (a) Polk zonal aggregates and (b) Synthesized zonal aggregates in Windsor in 2013

Summary and Conclusion

This paper presents a process for developing synthetic population to estimate the number of commercial vehicle owned at the business establishment level. The analysis also provides the basis to justify the use of zonal data of all the registered commercial vehicles within the study area from sources like Polk to study the spatial prevalence (or assignment) of such vehicles in a given census tract. The main concern is that some CVs that are registered to the establishment in the census tract as given in the Polk data might not be physically located or operating from that location. The hypothesis is that synthesized aggregates are a true representation of the vehicles registered and operated from the same zone. Hence achieving comparable results would provide a basis to use the of Polk data in future research. Using combinatorial optimization techniques, a synthetic population of establishments that engage in commercial activities for the Windsor CMA is created. Then the total number of vehicles owned was assigned to each establishment of the population by linking each establishment directly to an establishment of the micro-sample attained from the survey responses. Using the Polk data as a validation source, comparisons against the synthesized total number of vehicles were made. The results of comparison show a relatively good fit with a correlation of 0.88 between the synthesized and the validation data. This indicate that the Polk data can be used to examine the spatial distribution of commercial vehicles especially that it has more details on the year, make and model of the registered vehicles. Future research will focus on using the Polk data to develop new models of commercial vehicle ownership location to study the spatial prevalence of the various types of commercial vehicles, as derived from the Gross Vehicle Weight (GVW) classes, in a given zone.

References

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach: Transportation Research Record: Journal of the Transportation Research Board: Vol 2014, No. *Transportation Research Record: Journal of the Transportation Research Board*, 85–91.

Auld, J. A., Mohammadian, A. (Kouros), & Weis, K. (2009). Population Synthesis with Subregion-Level Control Variable Aggregation. *Journal of Transportation Engineering*, *135*(9), 632–639.

Auld, J. A., Rashidi, T. H., Mohammadian, A., & Weis, K. (2010). Evaluating transportation impacts of forecast demographic scenarios using population synthesis and data transferability. In *Proceedings of the 89th Annual Meeting of the Transportation Research Board (DVD), Washington, DC* (Vol. 1115). Retrieved from Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415–429.

Eluru, N., Pinjari, A., Guo, J., Sener, I., Srinivasan, S., Copperman, R., & Bhat, C. (2008). Population Updating System Structures and Models Embedded in the Comprehensive Econometric Microsimulator for Urban Systems. *Transportation Research Record: Journal of the Transportation Research Board*, *2076*, 171–182.

Government of Canada, S. C. (2012, February 8). Windsor, Ontario - Census metropolitan area - Focus on Geography Series - Census 2011. Retrieved from https://www12.statcan.gc.ca/census-recensement/2011/as-sa/fogs-spg/Facts-cma-eng.cfm?LANG=Eng&GK=CMA&GC=559

Guo, J., & Bhat, C. (2007). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *2014*, 92–101.

Huang, Z., & Williamson, P. (2001). A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. *Department of Geography, University of Liverpool*. Retrieved from http://pcwww.liv.ac.uk/~william/Microdata/Pop91/Methodology/workingpapers/hw_wp_2001_2.pdf

Hunt, J. D., & Stefan, K. J. (2007). Tour-based microsimulation of urban commercial movements. *Transportation Research Part B: Methodological*, *41*(9), 981–1013.

Madar, G. (2014). *Micro-Data Collection and Development of Trip Generation Models of Commercial Vehicles: An Application for Windsor, Ontario*. Retrieved from http://scholar.uwindsor.ca/etd/5186/

Martin Frick, I. V. T., Axhausen, K. W., & Zürich, I. (2004). Generating synthetic populations using IPF and monte carlo techniques: Some new results. Retrieved from http://matsim.org/uploads/ab225.pdf

Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidharthan, R., Christian, K. (2012). Application of Socioeconomic Model System for Activity-Based Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *2303*, 71–80.

Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, *41*(2), 181–203.

Samimi, A., Mohammadian, K., & Kawamura, K. (2012). Behavioral freight movement modeling: Methodology and data needs. *Travel Behaviour Research in an Evolving World*, 147.

Simpson, L., & Tranmer, M. (2005). Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software*. *The Professional Geographer*, *57*(2), 222–234.

Voas, D., & Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, *6*(5), 349–366.

Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, *30*(5), 785–816.