POPULATION SYNTHESIS TECHNIQUES: CREATING INPUT DATA FOR MICROSIMULATION MODELS

POPULATION SYNTHESIS TECHNIQUES: CREATING INPUT DATA FOR MICROSIMULATION MODELS

By

JUSTIN D. RYAN, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Arts

McMaster University

MASTER OF ARTS (2008) McMaster University

(School of Geography and Earth Sciences) Hamilton, Ontario

TITLE: Population Synthesis Techniques: Creating Input Data for Microsimulation Models

AUTHOR: Justin D. Ryan, B.Sc. (McMaster University)

SUPERVISORS: Dr. Pavlos S. Kanaroglou and Dr. Hanna Maoh

NUMBER OF PAGES: ix, 140

ABSTRACT

Population synthesis techniques are used to create lists of population members, where each member is endowed with attributes of interest. Aggregating these attributes across the synthetic members yields distributions which conform to known aggregate tabulations. Population synthesis is used when disaggregate population information is desired, and only aggregate and sample data is available. In this work, population synthesis techniques are discussed and compared, using a small, complete test population of firms. Given the results of these comparisons, populations of individuals and households are synthesized for the City of Hamilton, Ontario. These populations are then linked together to form a hierarchically ordered 'Comprehensive' population, where individuals belong to households, which in turn occupy dwellings over space. The synthesized comprehensive population is created specifically to meet the data input needs of URM-Microsim, a state of the art residential mobility microsimulation model. Originally calibrated for use in Europe, URM-Microsim is adapted for use in the Canadian context via the aforementioned comprehensive population. Some background on residential mobility modelling, as well as the URM-Microsim model is also presented.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Pavlos Kanaroglou. Thank you for introducing me to the world of human geography, and for your encouragement and support throughout my studies. Your dedication to your work, as well as your ability to see the 'big picture', is an inspiration.

Second, I would like to thank my co-supervisor, Hanna Maoh. Thank you for your tireless mentoring efforts, as well as your enthusiastic involvement with my work. I truly appreciate your attention to detail, and your commitment.

Third, I would like to thank my colleague Mari Svinterikou, for her consultation and assistance throughout my studies, as well as for creating the URM-Microsim model, which plays a key role in this thesis.

Fourth, I would like to thank James Goruk, for his consultation and assistance with computer programming, and particularly for his work on the Synthpop program.

My thanks are due to the students, staff and associated faculty members at the Centre for Spatial Analysis (CSpA), who have collectively enriched my experience as a student, and made my day-to-day routine a pleasure.

To my family and friends: while many believe that competition is the key to success; you have taught me that it is better achieved through cooperation and love. Thank you!

Finally, I would like to acknowledge the following authors, for their insight and timely advice: Don Miguel Ruiz; Neil Fiore; Andre Kukla; Leo Tolstoy; Benjamin Hoff; Sameer Grover; Bernadette Rule.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter One: Introduction

Simulation models are detailed representations of a system, often incorporating the elements of space and time. In social science, geo-spatial simulation models are used extensively to analyze and forecast patterns resulting from human behaviour (see: Terna, 1998; Gilbert & Troitzsch, 1999). These may include models of traffic flow, demography, land-use, migration, residential mobility and firm mobility, to name a few. As urbanization and human activity continue to intensify world-wide, high quality models of these phenomena become increasingly valuable. Here, models serve as an important input to the planning process; forecasting the results of hypothetical actions or inactions, and identifying their potential pitfalls, prior to implementation.

A major defining characteristic of simulation models is the scale, or resolution, at which they operate. Traditionally, study areas are divided into a set of mutually-exclusive and exhaustive zones; and these become the base units on which simulation models operate. For instance, suppose that a study area is divided into Census Tracts (CTs). A traditional firmographic simulation model would forecast the number of firms per CT, as well as the distribution of various firm characteristics for each CT, at some future time. The required input data for this model would be at the CT level. These zone based models are often referred to as 'aggregate models'. Recently, a new breed of 'disaggregate' or 'micro-simulation' models have gained popularity among researchers (see: Beckman et al., 1996; Huang and Williamson, 2002; Williams, 2003; Frick and Axhousen, 2004; Ballas et al., 2005; Simpson and Tranmer, 2005). Here, simulations occur at the level of the agents who are directly involved in the process being studied. In the firmographic case, micro-simulation models simulate the behaviour of each firm in the study area. For more on firmographic models, see: Van Wissen, 2002; Van Oort et al, 2003; Maoh & Kanaroglou, 2007.

Increases in computing power have facilitated the development of micro-simulation models; however their implementation is limited by the availability of input data. In particular, while zone based models require zonal input data, agent based models

require data on the entire set of agents being studied. For example, a geo-spatial micro-simulation model may require detailed data on firms or individuals, including spatial attributes. In general, population micro-data is either suppressed to maintain confidentiality, or incomplete due to the high cost of its acquisition (Moeckel et al., 2003). Consequently, population synthesis techniques are devised as a viable alternative to the collection of micro-data, for use in disaggregate geo-spatial models (Beckman et al., 1996; Harding et al., 2004).

Population synthesis techniques are algorithms that make use of typical categorical or ordinal socio-economic datasets, usually released in tabular form by statistical agencies, to produce a complete list of a population's members, each with associated attribute data, including the geographic location of each member. It is the latter attribute that endows these techniques with a geographic character. Accordingly, these algorithms take aggregate categorical or ordinal population data that are usually available in tabular format at the zonal level (e.g. CT or ward data), as well as an aspatial micro sample from a population (e.g. 2.5% public use micro data files from Statistics Canada) as inputs when estimating the complete list of a population's members. It is important to note that the synthetic individuals constructed from the micro-sample may have continuous variables that are represented as categorical/ordinal variables in the aggregate tabular data. A variety of synthetic populations can be created to suit different needs including; individual, household, dwelling, and firm populations. Two population synthesis techniques that have emerged as dominant in the literature are the Synthetic Reconstruction (IPFSR) technique and the Combinatorial Optimization (CO) technique (Huang and Williamson, 2002).

In Chapter 3, both of these techniques are implemented and analyzed. In particular, we test each method's ability to recreate a small, complete population of firms for the City of Hamilton, Ontario, in the year 1990. From the complete firm population (11,499 in total), different levels of input data are extracted. The techniques are implemented with these different levels of input data, and outputs are compared to the entire population, in order to explicitly test their quality. Chapter 3 realizes four main

goals: to implement the IPFSR and CO techniques for general use; to compare the two techniques, by measuring each one's ability to recreate the known population; to ensure that for both techniques, higher quality input data yields higher quality synthesized populations; to gain an idea of the minimum input data requirements for each technique to produce synthetic populations of reasonable quality. These goals are realized, as mentioned above, through a series of comparisons of the outputs from both techniques, using various levels of input data, to the known population data-set.

Given the basic knowledge of population synthesis techniques provided to us in Chapter 3, we proceed in the subsequent chapter to create a synthetic population for the City of Hamilton, Ontario, to be used as input to a micro-simulation model of residential mobility.

Residential mobility refers to the movement of individuals and households; between dwellings, and across space (see: Muth, 1969; Kendig, 1984; Strassmann, 1991). The process is not only spatial, but temporal, and is best observed in periods of between one and ten years. Residential mobility modelling has been a continued area of interest for researchers, especially at the municipal or city-wide scale. These Intra-Urban Models of Residential Mobility (URM models) forecast the changing patterns created by the residential decisions of individuals and households within an urban area, over time. Residential mobility patterns are related to demographics, land-use and land-pricing changes. These factors are incorporated into URM models to varying degrees, depending on the model specifics.

URM models have numerous applications in city planning; analysis of housing markets; and socio-demographic research, among others. Notably, URM models are vital components of Integrated Urban Land-Use Models (IULM), which simulate or forecast changes in all types of land-use over an urban area (see: Kanaroglou & Scott, 2002; Timmermans, 2003). These IULMs are often combined with traffic models, which forecast land-use and traffic patterns, taking into consideration the effects of land-use on traffic patterns, and vice-versa (see: Chang, 2006). URM models act along with other land-use models, notably firm-location (firmographic) models, to forecast the location

and intensity of origins and destinations for travel in an urban area. Models of modal-choice and route-choice are implemented in IULMs to forecast the characteristics of traffic (types of vehicles), as well as its spatial distribution over the network. The temporal frequencies of traffic and land-use changes are different; where traffic changes occur over periods ranging from seconds to days, land-use changes occur over years. For this reason it is often the "average" traffic conditions, for a certain day or time, that are forecast by integrated transport and land-use models. As an example, the average pattern of peak-hour, weekday traffic on the network can be modeled (see: Anderson et al, 1996). In keeping with the trend towards disaggregate modelling, a new generation of geo-spatial integrated transport and land use microsimulation models have emerged (see: Mackett, 1990; Moeckel et al, 2002; Hunt et al, 2005). In this context, URM micro-simulation models fill the place of traditional URM models, and hence require detailed micro-data as input.

In Chapter 4, a very specific synthetic population is created for a disaggregate URM model named "URM-Microsim" (see Svinterikou & Kanaroglou, 2007; Svinterikou, 2007). The model was originally used in the European context, for the City of Mytilene, Greece. Here, the model is to be adapted to the Canadian context, through the input of a synthetic population for the City of Hamilton, Ontario, in the year 1996. For this purpose, the input synthetic population must consist of four hierarchically related elemental populations, namely: individuals; households; dwelling units; and buildings. We refer to such a multi-level, hierarchically linked set of populations as a 'Comprehensive population'. Making use of the knowledge gained from Chapter 3, the elemental populations making up the comprehensive population are synthesized, keeping in mind the specific data requirements of the URM-Microsim model. Following this, the elemental populations are linked, using a series of rules which attempt to minimize unrealistic interactions between them. For instance, when assigning individuals to a household, the ages of parents and children must conform to certain limits. Once the Comprehensive population is created, measures of its validity are presented and discussed.

In Chapter 2, motivation for this research is elaborated upon. This includes background information on residential mobility modelling, population synthesis techniques, and the URM-Microsim model. This discussion helps to place the work carried out in Chapters 3 and 4 in context, as well as being of interest in its own right. Chapter 5 provides conclusions resulting from the study, as well as a brief discussion of future research avenues.

An Appendix is included in this work, containing some of the code used to create linkages between the elemental populations making up the Comprehensive population (from Chapter 4), as well as a detailed description of how particular variables in the Comprehensive population were adapted for use in the URM-Microsim model.

Chapter Two: Research Motivations and Literature Review

2.1 Introduction

The research presented in this thesis is primarily concerned with the creation of synthetic micro-data, for use in microsimulation models; and in particular, microsimulation models of residential mobility. The main motivations for this research are:

- To explore population synthesis techniques
- To establish a method for the creation of synthetic 'Comprehensive' populations, for use in microsimulation models
- To create a Comprehensive population for the City of Hamilton, Ontario, for use in the URM-Microsim model

This chapter will highlight concepts related to residential mobility, and will provide an overview of the URM-Microsim model, for which we will be creating a synthetic Comprehensive population in Chapter 4. Common methods for creating synthetic populations will also be discussed, drawing from key literature such as: Beckman et al., 1996; Williamson et al., 1998; Voas and Williamson, 2000. Finally, we will discuss existing efforts to create comprehensive populations, a notable example being: Guo and Bhat, 2007.

2.2 Background on Residential Mobility Modelling

Residential mobility refers to the movement of households and their associated individuals; between dwellings, and across space (see: Muth, 1969; Kendig, 1984; Strassmann, 1991; Henley, 1998). The process is not only spatial, but temporal, and is best observed in periods of between one and ten years. In Canada, between the years 2000 and 2001, approximately 14% of individuals moved to another place of residence. Of these, approximately 57% moved within their municipality. This latter type of

residential mobility is referred to as "intra-urban". The intra-urban residential mobility process has far reaching effects on the social and physical urban landscape, and hence is a topic of interest for researchers in geography, urban planning and sociology, to name a few. In the City of Hamilton, between the years 2000 and 2001, 13% of the population changed residences, with 60% of these moves being intra-urban.

In geography, study of the intra-urban residential mobility process began to blossom in the 1960s. This was due in large part to a paradigm shift in the field, where emphasis was placed on explaining social phenomena, and not only describing and mapping it, as was done prior (see: Short, 1978; Golledge, 1980). The resulting methods are referred to as 'behaviouralism' and acknowledge the collective role of individual and household actions in determining aggregate patterns. Here geographers began to incorporate ideas and research from the fields of economics, sociology, psychology and demography, among others (see Quigley & Weinberg, 1977; Strauss, 2008). A key conclusion reached by researchers is that: "Households [choose to] move when they perceive their existing dwelling to be un-desirable, or at least less desirable than an alternative" (Harris, 1991). It is the specifics of household satisfaction and dissatisfaction with their dwelling that is central to understanding residential mobility and creating generalized theories pertaining to the process. To this end, numerous empirical studies of residential mobility have been carried out since the 1950s. An early and oft cited work by Rossi (1955) helped to conceptualize the residential mobility process in a concise manner, and develop many fundamental ideas pertaining to it. He states that: "the major function of mobility [is] the process by which families adjust their housing to the housing needs that are generated by the shifts in family composition that accompany life cycle changes (p. 9)." Rossi also emphasizes the important distinction between voluntary and involuntary moves, where involuntary moves are: "necessitated by events totally beyond the control of the household" (Clark and Onaka, 1983, p. 49). In general, most theories of residential mobility apply primarily to voluntary moves, since these involve choices made at the household level. The theory provided by Rossi is elaborated upon by Speare et al. (1975), who saw the residential mobility decision as: "the result of an ongoing

decision-making process for which three stages can be distinguished: (1) the development of a desire to consider moving, (2) the selection of an alternate location, and (3) the decision to move or stay."

2.2.1 Variables Characterizing Residential Mobility

A household's 'desire to consider moving' usually results from considerations of, or changes in: Housing Characteristics; Neighbourhood Characteristics; Household Characteristics. Several important studies of residential mobility and housing characteristics were carried out by: Rossi, (1955); Moore, (1972); Pickvance, (1974); Goodman, (1976); Michelson, (1977); Thorns, (1981); Clark et al, (1986); Beer, (1999); Adair et al, (2000); Margulis, (2002); Kim et al, (2005); Clark et al, (2006); Erickson et al, (2006). The housing characteristics exerting the largest influence over a household's decision to move were found to be tenure and space. In the case of tenure, renting a dwelling is found to increase the probability that a household will decide to move, while the household sentiment of "not having enough space" has the same effect. The cost of housing is not found to have a large impact on a household's decision to move, however it becomes an important factor in choosing a new dwelling, once a decision to move has been made.

It is generally agreed upon that neighbourhood characteristics influence the household mobility decision process; however no consensus has been reached on the exact means by which this takes place. Some of the important work done on this topic includes: Rossi, (1955); Tiebout, (1956); Simmons, (1968); Goldscheider, (1971); Speare, (1974); Brown, (1975); Leven et al, (1976); Varady, (1983); Boehm and Ihlanfeldt, (1986); Connerly, (1986); Lewis, (1991); Lee et al, (1994); St. John et al, (1995); Henneberry, (1998); Adair et al, (2000); Greenberg and Lewis, (2000); Colwell et al, (2002); Parkes et al, (2002); Kim et al, (2005-1); Kim et al, (2005-2); Clark et al, (2006). Neighbourhood characteristics which are commonly researched in connection with the residential mobility process include: the social and demographic composition of neighbourhoods; social ties between members of a household and members of their

neighbourhood; the level of public services and amenities available in the neighbourhood; accessibility to goods and services. A further composite characteristic of 'Neighbourhood Quality' is frequently mentioned in the literature; however no consensus exits on its exact definition, or its influence on residential mobility.

Household characteristics are known to greatly influence the residential mobility process. Important work done on this topic includes: Rossi, (1955); Fredland, (1974); Speare, (1974); Pickvance, (1974); Doling, (1976); Duncan and Newman, (1976); McCarthy, (1976); Michelson, (1977); Michelson, (1980); Stapleton, (1980); Coupe and Morgan, (1981); Kendig, (1981); McLeod and Ellis, (1982); Clark and Onaka, (1983); Kendig, (1984); Meyer and Speare, (1985); Clark et al, (1986); Clark and Van Lierop, (1987); Speare and Goldscheider, (1987); Nijkamp et al, (1992); Baccaini, (1997); Ewert and Prskawetz, (2002); Li, (2004); Ostrovsky, (2004); Kim et al, (2005-2); Morrow-Jones and Wenning, (2005). The two most important household characteristics in this context are income and life-cycle changes. Of these, household life-cycle change is considered to be the most influential factor in the residential mobility process. Here the basic theory is that changes in the course of a household's "life", such as marriage, divorce, or childbirth, alter housing needs and create household dissatisfaction with their current dwelling. As explained by Clark and Onaka (1983): "Changes in household life cycle generate mobility either by altering specific housing needs (too little space, need for private yard, etc.) or by creating or eliminating a demand for an independent housing unit (household formation or dissolution)."

2.2.2 Residential Mobility Modelling Efforts

Given the theoretical and empirical advancements made in understanding residential mobility, researchers began to formulate models of the process. Two early seminal modelling efforts were made by Alonso (1960, 1964) and Wolpert (1965, 1966). Although both of these models, and the vast majority of models since then, are 'behavioural' in nature, Alonso's work falls into the sub-category of 'economic modelling', while Wolpert's work is considered 'sociological modelling.' Alonso builds

on agricultural land use theory (see: Von Thünen, (1826)) to determine land purchasing patterns made by households, given a set of simplifying assumptions. Wolpert's model, on the other hand, provides a framework for the residential mobility process comprised of three main elements: place-utility; action space; and environmental stress. Here, households assign a utility to locations in space, based on their level of satisfaction with the attribute set associated with that location. A household may only assign place-utilities to locations for which it has adequate information, and the set of these locations is known as its action space. Over time, households may be subject to 'environmental stresses', due to life-cycle changes, neighbourhood changes and income changes, among others. Once these stresses accumulate beyond some threshold, households will react by relocating to another dwelling, located somewhere in their action space.

Traditionally, models of the residential mobility process were implemented at the zonal level. That is to say, for a study area divided into a set of zones, the zones themselves would be modelled; updating attributes of each zone based on the particular theory incorporated into the model, as well as the initial state of each zone. Note that this type of 'aggregate' modelling requires input information at the zonal level, and in turn produces output at the zonal level. An alternative to zone based, aggregate modelling is microsimulation, also known as 'agent-based' modelling. In the case of residential mobility microsimulation models, the agents being modelled are the individuals or households in the study area (who are directly involved in the decision making process which leads to residential mobility patterns). Microsimulation modelling was introduced by Orcutt (1957), and is intimately connected to behaviouralism. In particular, microsimulation models allow for the easy incorporation of behavioural theories, since these theories apply directly to the agents being modelled. As stated by Clarke and Holm, (1987), the microsimulation technique: "... allows for the incorporation of theories that have been developed at that level as opposed to those relating to aggregates of micro units." For further justification of the use of microsimulation in modelling, see: Mackett, 1990; Goulias and Kitamura, 1992.

Microsimulation modelling has gained popularity among researchers over the last few decades, primarily due to the advantages it has over traditional aggregate modelling. Microsimulation allows for highly detailed representations of model agents (such as individuals, households and buildings in the residential mobility case); as well as an incorporation of this detailed representation into the framework of the model. Agents may also be endowed with a detailed spatial attribute, which serves to further enrich their attribute set. The detailed attributes of each agent are updated over time during microsimulation modelling, and so longitudinal analysis can be performed on agents once they have been treated by a given model. Furthermore, due to the disaggregate nature of the agents, model results can be aggregated for analysis in a number of ways, according to specific needs. For instance, if individuals in an urban area are being modelled, the results can be presented by sex, by age, by zone or indeed by any combination of attributes common to all of the individuals. It should be noted here that although microsimulation models produce a database containing each agent, and their attributes over time, it is some form of aggregate result that is often required for analysis and policy making decisions. A major advantage of microsimulation modelling is the ease with which different aggregate results can be formulated from the output.

Although they are essentially outweighed by the advantages, microsimulation modelling does have several disadvantages. First, microsimulation models require a detailed data set comprised of a record for each agent being modelled, along with associated attribute data. These datasets are for the most part difficult to come by, due to privacy concerns, as well as the high cost of data acquisition and collection. One solution to this dilemma is offered by population synthesis techniques (see Section 2.3), which provide a set of synthetic agents that mimic known aggregate tabulations of attributes, and can be used as input to microsimulation models. Another disadvantage of microsimulation models is the large amount of computing time and power they require. Generally, in every microsimulation model, two main steps are involved: entering a database of agents, along with associated attribute data, into the model; and updating these agents over some time period, based on the rules and theory incorporated into the

model. Since many of the processes which are studied using microsimulation models involve large numbers of agents, large amounts of computing resources are often needed to treat these. As an example, in a residential mobility model of a medium sized urban area, there may be 250,000 households, each containing between one and ten individuals. In this case both the households themselves as well as the individuals might have tens of attributes which are considered by the model.

A number of efforts have been made to model the residential mobility process using the microsimulation modelling framework. Here, the agents involved are usually individuals, households, and some representation of the housing stock. Simulation of these agents over time reveals patterns of residential mobility, demographics and land-use in general. Some of the more prominent residential mobility microsimulation efforts are: HUDS (Kain and Apgar, 1985); ILUTE residential mobility model (Miller et al, 1987; Salvini and Miller, 2005); HOUSIM (Clarke et al, 1989); UPDATE (Clarke, 1995); Fransson model (Fransson, 1994; Fransson and Makila, 1994); LOCSIM (Oskamp, 1995; Hooimeijer and Oskamp, 1996; Oskamp, 1997); SAMS (Kitamura et al, 1996); SVERIGE (Vencatasawmy et al, 1999); UrbanSim household mobility and location module (Waddell, 2000); ILUMASS land-use model (Moeckel et al, 2002; Wagner and Wegener, 2007); MAS/RE (Chong and Jianquan, 2007); SwarmCity (Devisch et al, 2004); TLUMIP (Weidner et al, 2006); ALBATROSS (Arentze and Timmermans, 2000; 2004); PUMA (Ettema and Timmermans, 2006); IRPUD (Wegener, 1998; Spiekermann and Wegener, 2007). Note that many of the previous models are not devoted entirely to residential mobility simulations, and instead include these as a sub-module, or component of the overall modelling effort. Joining this list of residential mobility microsimulation models is URM-Microsim (Svinterikou, 2007; Svinterikou & Kanaroglou, 2007), which will be elaborated upon in the following section.

2.2.3 The URM-Microsim Model

The Urban Residential Mobility Microsimulation model (URM-Microsim) is presented by Svinterikou (2007). In reference to the discussion in sections 2.2.1 and

2.2.2 of this chapter, URM-Microsim is a sociological, life-cycle type residential mobility microsimulation model. Chapter 4 of Svinterikou (2007) details the basic form and functions of the model. Here, it is noted that URM-Microsim is: "a dynamic, spatio-temporal microsimulation model that captures the developments in demographic structures and housing markets, as well as the interrelation between them." To this end, the model consists of three sub-modules: Housing Demand; Housing Supply; and Residential Search and Migration. Figure 2.1 depicts the sub-modules, as well as relationships between them.

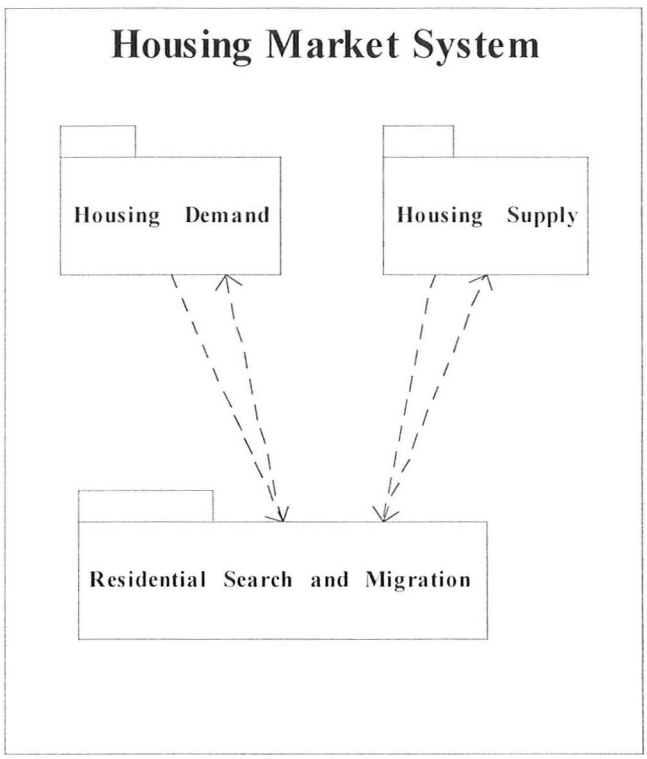

Figure 2.1: The sub-modules of URM-Microsim. Source: Svinterikou, 2007.

In brief, the Housing Demand sub-module simulates the lives of individuals and households in the study area. This includes simulating demographic events such as births, marriage, household formations and household dissolutions, as well as ageing the population over time. The Housing Supply sub-module maintains the database of dwellings and buildings, and simulates events such as building constructions, demolitions and structural conversions. The Residential Search and Migration sub-module simulates

the search that households wishing to move residences undertake, as well as the outcome of these searches, in terms of updated household distributions across the set of dwellings, newly vacant dwellings, and household out-migrations. Each of the sub-modules updates the population database on a yearly basis, and the outcomes are used to inform simulations of subsequent time periods. In order for URM-Microsim to operate, a detailed base-year population database is required, comprised of individuals, households, dwelling units, and buildings, along with spatial identifier attributes. Furthermore, a set of probability distributions is required, which governs the processes simulated by the model. A more elaborate description of the sub-modules, as well as their data requirements, follows.

The Housing Demand sub-module ultimately aims to identify households from the population that wish to change residences. It does so by simulating the lives of individuals and their associated households, in the study area. As stated by Oskamp (1997): "As individuals and households pass through different life stages (fertility, marriage, occupation, education), their housing preferences change, triggering mobility." In particular, individuals and households may be subjected to three event types, according to a set of conditional probability distributions. These event types are: Demographic change; Income change; and Demand change. Demographic changes include: death; fertility; union formation; union dissolution; flat-mate formation; flat-mate leaving; nest leaving; and out-migration. Individuals eighteen years of age and older may be subjected to income change events, which arise from changes in occupation, or wage changes in an individuals' current occupation. For all households, there is a small probability of a Demand change event, which occurs when a household that has not experienced a demographic or income change nonetheless seeks a new residence. Once a household has been identified as 'demanding a new dwelling', a search intensity is assigned to it, based on the event that the household has most recently experienced. Search intensity is a measure of how much effort a household will expend searching the market for a new dwelling. Here, for example, a newly divorced individual will have very high search intensity, while a married couple who have just had their first child will search the market

with a low intensity. When considering the Housing Demand sub-module, it is important to keep in mind the specific characteristics of individuals and households required by the model to simulate life events. Table 2.1a lists the attributes of individuals required by URM-Microsim, while Table 2.1b does the same for households. The probabilities and probability distributions required by URM-Microsim for use in the Housing Demand sub-module are listed in Table 2.2.

Table 2.1a: Attributes of individuals as required by URM-Microsim

Attribute	Description
Individual ID	
Household ID	
Zone	CT or Ward, for example
Sex	
Date of Birth	
Marital Status	
Position in household	Household head, child, couple, flat-mate
Education	Highest level attained
Employment Status	
Income	
Occupation	If employed
Industry	If employed

Table 2.1b: Attributes of households as required by URM-Microsim

Attribute	Description
Household ID	
Dwelling ID	
Zone	
Size	Number of household members
Income	Sum of members' incomes
Tenure	Does household rent or own dwelling?

Table 2.2: Probabilities required by Housing Demand sub-module

Reference #	Description
1	Probability of death, by age and sex
2	Women's fertility probability by age group
3	Women's fertility probability by marital status
4	Probability of being born female
5	First year mortality probability by sex
6	Probability of being born a twin
7	Marriage probability by age and sex
8	Probability of cohabitating by age and sex
9	Probability of female of age x marrying male of age y
10	Probability of male of age x marrying female of age y
11	Probability that a cohabitating couple will marry
12	Divorce rates by years of marriage
13	Divorce rates by years of marriage and age of female
14	Rate of "de-cohabitation"
15	Probability that female keeps children after a divorce
16	Probability that female keeps dwelling after a divorce
17	Probability of out-migration after divorce, by sex
18	Probability of children older than 17 leaving home
19	Probability of children older than 17 out-migrating (child lives at home)
20	Probability of getting a new room-mate
21	Probability of leaving a group of room-mates
22	Probability of a leaving room-mate out-migrating
23	Out-migration probability based on event experienced
24	Probability of students or 'foreigners' in-migrating
25	Education rates by age and sex
26	Probability of a household with no event changing housing demand
27	Ownership specific residential search probabilities
28	Event and Search Criteria specific residential search probabilities

The Housing Supply sub-module is responsible for maintaining the database of dwellings and buildings, as well as simulating changes which may occur to this stock. Note that a dwelling is defined as a set of rooms in a building which may serve as a residence for a household, while a building is a physical structure which may contain a countable number of dwellings. In the typical case of a single family home, there is one dwelling in the building; however if the family constructs a separate apartment in their basement, the building will contain two dwelling units. The Housing Supply sub-module

may subject dwellings (and their associated buildings) to four event types: New construction; Structure conversion; Demolition; and Housing expenditure change. New dwelling constructions may occur, where the number of these is determined by the overall level of household mobility, as well as the supply of vacant dwellings. The location of new dwellings is based on market housing demand trends, as well as the availability of suitable land. The type of new dwellings to be constructed is based on market housing demand trends. Structural conversions to existing dwellings may occur, subject to a probability distribution, as well as the profitability and costs of such an action. Three types of structural conversions may occur: Splitting; Combination; and Renovation. Each dwelling in poor condition is subject to the probability of demolition. In the case that an occupied dwelling is to be demolished, demolition will not take place until the household has relocated. Finally, each dwelling is subject to the probability of a change in housing expenditure (such as maintenance, heat and electrical expenditures), based on dwelling type, size, age, and construction quality. Table 2.3a lists the specific attributes required by URM-Microsim for members of the dwelling population, while Table 2.3b does the same for buildings.

Table 2.3a: Attributes of dwelling units as required by URM-Microsim

Attribute	Description
Dwelling ID	
Building ID	
Zone	
Size	Floor space
Number of Rooms	
Market value	If owned
Rent value	If rented (monthly)
Availability	For rent or sale?

The Residential Search and Migration sub-module affects households and dwelling units. Households with a non-zero search intensity are simulated to search the housing market for a vacant dwelling which meets their needs. If a suitable dwelling is found, the household moves into it. Otherwise, the household either remains in its

current dwelling, or out-migrates, according to a certain probability. Households with higher search intensities consider more dwellings in their search than those with lower intensities. Dwellings are defined by the following seven characteristics: price/rent; number of rooms; dwelling type; dwelling age; structural condition; amenities (garden, parking, etc.); and neighbourhood quality. Different households assign different levels of importance to each of these dwelling characteristics for their search of the housing market. Dwelling characteristic importance levels are based on the event which the household has most recently experienced, as well as household income, and whether the household is searching for a dwelling to rent or own. The dwelling tenure which a household searches for is based on the following considerations: previous tenure; household income; life-cycle stage; current mortgage interest rates; and a tax incentive to first time buyers. Once household moves have been simulated, dwelling rent and market values are updated based on supply and demand in the housing market.

Table 2.3b: Attributes of buildings as required by URM-Microsim

Attribute	Description
Building ID	
Block ID	Parcels are the Hamilton equivalent
Zone	
Type	Indication of structure
Floor space	In square feet
Number of floors	
Number of dwellings	
Year of construction	

URM-Microsim was originally calibrated for use in the City of Mytilene, Greece. A comprehensive synthetic population, to be described in Chapter 4, as well as the vital statistics listed in Table 2.2, are required to adapt URM-Microsim for use in the context of the City of Hamilton. Although the comprehensive synthetic population is endowed with the particular variables required by URM-Microsim (see Tables 2.1a, 2.1b, 2.3a and 2.3b), these variables were derived from Canadian sources, and occasionally require manipulation in order to match the categorization scheme required by the model. As an

example, the 'age' of individuals might be presented as an integer in the Canadian data, but might be required in five-year intervals by the model. The re-classification schemes which were used to treat attribute variables in the comprehensive synthetic population are presented in Appendices VII through X. Appendix XI contains the specific domains of all URM-Microsim database variables.

In addition to vital statistics and a comprehensive population, URM-Microsim also requires a spatial representation of the study area. Here, the study area is geographically divided into zones and blocks, where blocks belong to zones, and zones are exhaustive and mutually exclusive. In the Hamilton context, zones correspond to Census Tracts, while blocks correspond to parcels. The following attributes are required for zones: name; type; quality; and population type. For blocks, the required attributes are: level of shopping access; level of school access; level of transportation access.

2.3 Methods for Creating Synthetic Populations

Population synthesis techniques are algorithms that make use of typical categorical or ordinal socio-economic datasets, usually released in tabular form by statistical agencies, to produce a complete list of a population's members, each with associated attribute data, including the geographic location of each member. It is the latter attribute that endows these techniques with a geographic character, hence making them appealing to geographers. Accordingly, these algorithms take aggregate categorical or ordinal population data that are usually available in tabular format at the zonal level (e.g. census tract or ward data), as well as an aspatial micro sample from a population (e.g. 2.5% public use micro data files from Statistics Canada) as inputs when estimating the complete list of a population's members. It is important to note that the synthetic individuals constructed from the micro sample may have continuous variables that are represented as categorical/ordinal variables in the aggregate tabular data. Various types of synthetic populations can be created to suit different needs; these may include individual, household, dwelling and firm populations, to name just a few.

The basic ideas behind population synthesis are straightforward. Given whatever aggregate information for the population is available, a list of the population members is estimated, using some algorithm, such that the characteristics of the synthetic population correspond to the aggregate information. Then it can be said that the synthetic population is among the set of 'best possible' estimates of the actual population, given the input information. Of course, different algorithms may produce synthetic populations, which all conform to the input data, but differ in their quality. This is the result of each algorithm having its own underlying theory. Two population synthesis techniques that have emerged as dominant in the literature are the Synthetic Reconstruction (IPFSR) technique and the Combinatorial Optimization (CO) technique (see: Huang and Williamson, 2002).

The Synthetic Reconstruction technique is based on the work of Wilson and Pownall (1976), and is a member of a family of methods which rely on Iterative Proportional Fitting (IPF). IPF is a method of updating cells in a table, based on known marginal totals (see: Deming and Stephan, 1940; Wong, 1992). As noted by Guo and Bhat, 2007, as well as Johnston and Pattie, 1993, the IPF procedure is equivalent to solving an entropy maximization problem. This point is of interest when devising Synthetic Reconstruction algorithms, although we do not employ it in our approach (see Section 2.3.1).

Williamson et al., 1998, discuss several Combinatorial Optimization techniques, namely: Hill Climbing; Simulated Annealing; and Genetic Algorithm approaches. In each of these, sets of synthetic population members are randomly chosen, and then slowly altered until the sets match given aggregate constraints. In Hill Climbing, members of the initial set are swapped from a larger sample of possible members at random, and only swaps which improve the fit of the set are permitted. In Simulated Annealing, a similar approach is followed, however in some circumstances swaps which do not improve the overall fit of the set are permitted, in order to avoid getting stuck on sub-optimal, local-maxima solutions. Genetic Algorithms rely on combining portions of various "parent" sets of population members, until a "child" set is found which conforms

to the required aggregate constraints. In this work (see Section 2.3.2), we make use of the Hill Climbing approach to Combinatorial Optimization.

Synthetic population techniques are inherently mathematical and can be used to create population lists that are spatial or aspatial. In the latter case, location as a variable can be treated the same as any other variable having several outcomes. For instance, while a location variable could have outcomes: zone1; zone2; and zone3, an age variable could have outcomes: young, adult and elder. There is no conceptual difference between location and age variables as they are treated in most population synthesis techniques. However, a common situation in the geographical context is to have input data consisting of a small sample from the population (generally with no spatial identifiers such as a location variable), as well as tabulations representing the distribution of population characteristics over space (i.e. across zones or census tracts that constitute the location variable). It is to this situation that the CO and IPFSR techniques are particularly well suited, and where location variables take on a special role, relative to other variables.

2.3.1 The Synthetic Reconstruction Technique

The Synthetic Reconstruction technique is presented by Wilson and Pownall (1976). Here the emphasis is on the creation of synthetic populations, given a known multiway table of conditional probabilities pertaining to population characteristics. Technical aspects such as the ordering of conditional probabilities in the selection of attributes are explored fully, and an example of a synthetic population of households is presented. Variations of the Synthetic Reconstruction technique, generally making use of the Iterative Proportional Fitting (IPF) technique for the creation of multiway tables, are widespread (see for example: Smith et al., 1995; Beckman et al, 1996; Huang and Williamson, 2002; Frick and Axhousen, 2004; Walker, 2004; Ballas et al, 2005; Simpson et al, 2005; Arentze et al, 2007; Guo et al, 2007). It is important to note that while IPFSR makes use of IPF, the two techniques are not equivalent. IPF was first proposed by Deming and Stephan (1940) and since then has been widely used in many disciplines for various applications. As previously mentioned, using IPF to update a multiway table

given marginal constraints, is equivalent to solving an entropy maximization problem, having those same constraints (Johnston and Pattie, 1993).

Intuitively, IPF operates on a two-way table that has *i* rows and *j* columns and an initial state ($\mathbf{T^0}$). The objective is to update all cells T_{ij}^{0} in the table to reflect a new state T_{ij}^{n} using pre-defined totals of rows ($\sum_{j} T_{ij}^{n}$) and columns ($\sum_{i} T_{ij}^{n}$) that are available for a new state n. The calculation of table $\mathbf{T^n}$ is achieved through a number of iterations, each of which is executed in two steps. In Step 1, IPF calculates row proportions by dividing the new state row totals by the initial state row totals (from $\mathbf{T^0}$) and applying those proportions to each cell T_{ij}^{0} in the corresponding row of the initial table $\mathbf{T^0}$. This will result in a new table $\mathbf{T^1}$ that replaces table $\mathbf{T^0}$. In Step 2, IPF calculates column proportions by dividing the new state column totals by the initial state column totals (from $\mathbf{T^1}$) and applying these proportions to each cell T_{ij}^{1} in the corresponding column of the initial table $\mathbf{T^1}$. This will result in a new table $\mathbf{T^2}$ that still reflects the initial state. $\mathbf{T^2}$ is used as input to the second iteration where steps 1 and 2 are repeated starting from $\mathbf{T^2}$ as the initial state (i.e. $\mathbf{T^2}$ is denoted as $\mathbf{T^0}$). The procedure repeats steps 1 and 2 in an iterative fashion until the calculated proportions of row and column totals are equal to 1, or an iteration limit is reached. At this point, the cells of the original table are updated and now reflect the new state, where the row and column totals conform to the pre-defined row and column totals of the new state n. The result is a new table $\mathbf{T^n}$ reflecting the new state. As will be discussed later, IPFSR uses IPF as an intermediate step while creating the synthetic micro population.

The Synthetic Reconstruction (IPFSR) method, detailed in Huang and Williamson (2002), creates members of the synthetic population one by one, using a set of conditional probabilities. That is to say, an individual is assigned characteristics sequentially, where the probability of being assigned a given characteristic is conditional on previously assigned characteristics. The first characteristic assigned to an individual is based on the unconditional probability of that characteristic occurring in the population. If the synthetic population is to contain n characteristic variables (such as age, sex, income and geographic location) then the probabilities of each characteristic occurring in the

population must be stored in an n-dimensional table, referred to as a multiway table. Usually, the multiway table contains the entire population counts of individuals with each possible combination of characteristics, such that the sum of all cells in the table is the desired synthetic population size. For instance, a given cell could contain the number of individuals in the population who are old, male, rich, and residing in "zone 3". A series of Monte Carlo simulations are performed to assign characteristics sequentially to a synthetic individual, using the counts contained in the multiway table. It is important to note that changing the order in which characteristics are assigned may affect the resulting synthetic population, and hence this issue should be carefully considered.

Generally speaking, a multiway table of characteristic-set counts will not be available, and one needs to be estimated using the Iterative Proportional Fitting (IPF) technique. IPF requires as its input a sample from the population, as well as tabulations of the distributions of variables to be included in the synthetic population. A detailed description of how IPF is used to create the required multiway tables can be found in Beckman et al (1996) as well as Huang and Williamson (2002).

Population synthesis techniques are best understood with a numerical example. Suppose we have a region of 30,000 individuals divided into 3 mutually exclusive and exhaustive zones with 10,000, 15,000 and 5,000 individuals, respectively. The available public data for the study area includes: aggregate tabulations with the break down of zonal population by Sex, Age and Income as shown in Tables 2.4a, 2.4b and 2.4c; and a 2.5% micro sample of 750 individuals with information about the sex, age and income of each member in the sample, as shown in Table 2.4d. Given the information in tables 2.4a – 2.4d, both the CO and IPFSR methods can be employed to create a list of 30,000 individuals, where each member in the list has attributes of sex, age and income that will, when aggregated, conform to the zonal totals.

Table 2.4a: Example Sex Tabulation

Sex	Zone 1	Zone 2	Zone 3	Total
M	4,500	8,000	3,000	15,500
F	5,500	7,000	2,000	14,500
Total	**10,000**	**15,000**	**5,000**	**30,000**

Table 2.4b: Example Age Tabulation

Age	Zone 1	Zone 2	Zone 3	Total
Young	4,000	4,200	900	9,100
Middle	2,500	7,800	3,000	13,300
Old	3,500	3,000	1,100	7,600
Total	**10,000**	**15,000**	**5,000**	**30,000**

Table 2.4c: Example Income Tabulation

Income	Zone 1	Zone 2	Zone 3	Total
Poor	2,000	5,000	0	7,000
Middle	5,000	6,000	2,300	13,300
Rich	3,000	4,000	2,700	9,700
Total	**10,000**	**15,000**	**5,000**	**30,000**

Table 2.4d: Example 2.5% micro sample

ID	Sex	Age	Income
Person 1	M	Young	Middle
Person 2	M	Old	Middle
Person 3	F	Old	Rich
...	...	...	...
Person 750	F	Middle	Poor

In the given example, using the IPFSR technique, the first task is to create a multi-way table with estimated counts (or probabilities) of each possible combination of Sex, Age, Income and Zone characteristics. This requires multiple applications of IPF. To begin with, we are given the two-dimensional tables of (Sex × Zone), (Age × Zone) and (Income × Zone). From these we can derive the one-dimensional tabular distributions of Sex, Age, Income and Zone. Since there are 4 characteristics to be constrained for in the population, there are a total of 6 two-dimensional tables. Those which are not given (Sex × Age, Sex × Income, Age × Income) must be estimated using

IPF, where the corresponding one-dimensional tabulations are used as the column and row constraints during the process. For instance, if we begin by creating the Sex × Age table, then the one-dimensional tabulations of Sex and Age become the constraints in the IPF process, while the initial state of the table is determined by filling the cells of Sex × Age with counts from the sample. Therefore, the cells in the initial table will sum to 750.

Once all of the 2-dimensional tables have been generated, the 4 possible 3-dimensional tables are created. These are: (Sex × Age × Income); (Sex × Age × Zone); (Sex × Income × Zone); (Age × Income × Zone). In the case of (Sex × Age × Income), the already created 2-dimensional tables (Sex × Age), (Sex × Income) and (Age × Income) become the constraints for the IPF process (equivalent to the row and column totals in the creation of 2-dimensional tables). The initial state of the 3-dimensional table is determined from the counts in the sample. The only exception to this is in the creation of 3-dimensional tables containing Zone, such as (Sex × Age × Zone). Here, the initial state of the table cannot be taken from counts in the sample, because the sample does not include information on the zonal characteristics of its members. In these cases, the initial 3-dimensional table is filled with 1s in every cell.

Finally, the 4-dimensional table of (Sex × Age × Income × Zone) is created, using the 3-dimensional tables as constraints, and an initial state of 1s in every cell. It is from this table that the synthetic reconstruction portion of IPFSR proceeds. Each individual in the population is created, through a series of Monte Carlo simulations based on the conditional probabilities in the 4-dimensional table. For instance, if Sex is the first characteristic to be assigned to each individual, then the number of males and females are taken from the 4-dimensional table, a number between zero and the total population count is drawn, and if it falls below the number of males, then the first individual is assigned the male sex. Given that the individual is male, the numbers of young, middle aged and old will be derived from the 4-dimensional table, and another Monte Carlo simulation will be run, assigning the individual to one of those categories. These draws from conditional probability distributions continue until the individual has been assigned a value for all 4 characteristics. Following this, the remaining individuals in the population

are created in the same way as the first. We should again note that the order in which characteristics are assigned can influence the resulting population, so this order must be designed with care.

2.3.2 The Combinatorial Optimization Technique

The Combinatorial Optimization technique and its variations are far less common in the literature than Synthetic Reconstruction techniques. One major effort to synthesize populations using these techniques has been undertaken by the National Centre for Social and Economic Modelling (NATSEM), centered at the University of Canberra, Australia (see for example: Williams, 2001; Melhuish et al, 2002; Harding et al, 2004). Another effort at population synthesis, using Combinatorial Optimization, from which the techniques described in this chapter are more directly derived, can be found in Voas and Williamson (2000) as well as Huang and Williamson (2002). In the latter paper, the CO and IPFSR methods are directly compared in their abilities to produce synthetic microdata. It is concluded that although both methods produce reliable synthetic microdata, there is less variation amongst populations produced using the CO method, as compared with those created using the IPFSR method. Therefore, the CO method is deemed superior to the IPFSR method.

The CO method synthesizes the population on a zone by zone basis. We will again refer to the example presented in Section 2.3.1. Operating on the first zone, the method starts by performing a first draw to select (with replacement) a sub-set of 10,000 individuals at random from the micro sample (listed in Table 2.4d) to match the total population size in that zone. Using the drawn sub-set, the method then generates three distributions similar in form to those shown in Tables 2.4a, 2.4b and 2.4c using the age, sex and income attributes of the selected 10,000 individuals. In this first draw of individuals, there might be: 2,100 males and 7,900 females; 2,500 young, 3,000 middle aged and 4,500 old; 3,000 poor, 3,000 middle income and 4,000 rich. The fit of this list of individuals to the 3 tabulations is assessed using the overall Relative Sum of Squared Z-scores (*RSSZ*) proposed by Voas and Williamson (2001, pp. 187) and Huang and

Williamson (2002, pp. 54-57). The statistic is formulated by adding the relative Sum of Squared Z-scores (SSZ_k) for all input tabulations k as follows:

$$RSSZ = \sum_k SSZ_k \quad \ldots(1)$$

Where:

$$SSZ_k = \sum_i F_{ki}(O_{ki} - E_{ki})^2$$

$$F_{ki} = \begin{cases} \left(C_k O_{ki} \left(1 - \frac{O_{ki}}{N_k} \right) \right)^{-1}, & \text{if } O_{ki} \neq 0 \\ \frac{1}{C_k} & \text{if } O_{ki} = 0 \end{cases} \quad \ldots(2)$$

O_{ki} is the observed (from the sub-set) count for the i^{th} cell of the k^{th} tabulation (characteristic)

E_{ki} is the expected (known) count for the i^{th} cell of the k^{th} tabulation

N_k is the total count of tabulation k

C_k is the 5% χ^2 critical value for tabulation k (with n-1 degrees of freedom, for a table with n cells).

It should be noted that C_k is used in equation 2 to make the Sum of Squared Z-scores (SSZ_k) for the input tabulations relative when formulating $RSSZ$. This is required since different tabulations may not have equal numbers of cells (i.e. equal degrees of freedom). SSZ_k has been proposed by Voas and Williamson (2001, pp. 185) as a robust statistic to assess the goodness of fit when comparing tabular data while creating synthetic populations.

When the distributions formed from the sample sub-set exactly fit the constraining tabulations, the value of *RSSZ* equals zero. Huang and Williamson (2002, pp. 18, 20) provide several justifications for the use of the *RSSZ* statistic during the sample sub-set selection process. First, the statistic provides a relative measure of the fit of a sub-set to different tables, regardless of the number of cells in those tables. That is to say that the fit of a sub-set to one table can be directly compared to the fit of that sub-set to another table, though the two tables may have different sizes (as in the case of age and sex in this example). Second, the statistic calculated for each constraining table and then summed provides a measure of the overall fit of the sub-set to the constraining tables, treating each table with equal importance. Third, the statistic is focused on the fit of individual counts (cell by cell) and not only the fit of an entire table.

If the value of *RSSZ* is large, then the selected sub-set is a poor fit to the constraining tabulations. Consequently, one of the chosen individuals from the sub-set is swapped with another from the 2.5% sample at random, with replacement. This results in a new sub-set that is used to form a new set of distributions that are again assessed against the constrained distributions listed in Tables 2.4a – 2.4c, using *RSSZ*. If the new value of *RSSZ* is reduced, the new swap is deemed superior and the switch of individuals is maintained. If the new sub-set does not have an improved fit to the tabulations, then the original sub-set is maintained. Following this, a new random switch of one individual from the sub-set with one from the sample is assessed, and this process continues until an iteration limit, or a fit threshold is reached. The latter is set to 1 such that if the synthesized population can produce an *RSSZ* value of less than or equal to 1, then the last drawn sub-set is deemed appropriate to represent the population of Zone 1. The populations for Zones 2 and 3 are created in the same way, using the same sample of 750 individuals. The order in which the zones are synthesized is not important.

The rationale for choosing an *RSSZ* value of less than 1 is to ensure a significant goodness of fit. A perfect match will occur if SSZ_k is 0 for all k, which consequently will result in an *RSSZ* value of 0. However, setting a value of 0 as a tolerance for SSZ_k is not practical and might not allow the CO method to converge. As a result we aim for a very

small tolerance value that is close to 0 (say $\eta = 0.1$) and check if the calculated SSZ_k is less than η. Since *RSSZ* is obtained by adding up all SSZ_k's, then the result should be a small value that is less than 1 and close to zero. From a statistical stand point, the calculated *RSSZ* can be assigned a significance level. Following Voas and Williamson (2001), *RSSZ* can be evaluated using a Chi-square test, where the degrees of freedom of the *RSSZ* statistic equal the sum of the degrees of freedom of all the input k tabulations. As an extreme case, when the degrees of freedom equal 1, *RSSZ* is significant if it has a value that is less than 3.84, which is the critical Chi-square value at the 95% confidence level. Aiming for a value of less than 1 (which is significantly smaller than the critical Chi-square value under any condition) will guarantee that the synthesized population closely mimics the actual population.

2.4 Comprehensive Micro Populations

Given a geographic study area, and a population of some kind occurring over the study area (individuals, households or buildings for example), several entities may be defined. First, we define a "complete population data set" as an exhaustive list of the individual components of a population, where each individual is associated with a set of attributes, one of which is spatial. Secondly, we define a "comprehensive population data set" as two or more complete population data sets, which are linked together in a hierarchical fashion. For instance, a longitudinal data set of individuals in a city at two different time periods (t1, t2) could be considered a 'comprehensive population'. Here individuals at t1 are each linked to a single individual at t2, provided the individual resides in the city at both time periods. Complete population data sets making up a comprehensive population may be referred to as 'elemental' populations.

In this work, a primary goal is to create a comprehensive population for the URM-Microsim model. In order to meet the model requirements, this must consist of four elemental populations, namely: individuals; households; dwelling units; and buildings. Each member of these elemental populations must be endowed with attributes as required by URM-Microsim (see Tables 2.1a, 2.1b, 2.3a, 2.3b). Furthermore,

hierarchical relationships must exist between the elemental populations such that: individuals belong to households; households belong to dwelling units; and dwelling units belong to buildings. To realize this goal, a series of linking rules are implemented on the initially independent elemental populations (see Chapter 4).

The creation of comprehensive populations for use in microsimulation models has not received much attention in the literature. Two notable exceptions are Moeckel et al. (2003), and Guo and Bhat (2007), who use an IPF based technique to create a comprehensive population of individuals and households. Guo and Bhat point out that:

> "...past efforts of population synthesis have accounted for only the household-level contingency table during sample household selection, leaving the individual-level variables uncontrolled (i.e., the resulting gender distribution in the synthesized population is likely to deviate from the known gender distribution given by SF1). The deviation could severely affect the accuracy of the subsequent microsimulation outcome. Thus, a method that controls both household- and individual-level distributions is needed."

This implies that previous microsimulation models of residential mobility have made use of a rudimentary comprehensive population of individuals and households, where the characteristics of individuals may not fit known aggregate tabulations. In this work, we test a method of creating comprehensive populations that differs from that suggested by Guo and Bhat, 2007, and instead relies on linking pre-synthesized elemental populations. One notable advantage of proceeding in this manner is that multiple elemental populations can be included in a given comprehensive population.

Chapter Three: Comparing Methods of Population Synthesis

3.1 Introduction

The two population synthesis techniques most commonly cited in the literature are the Synthetic Reconstruction (IPFSR) technique, and the Combinatorial Optimization (CO) technique. In this chapter, we implement both of these techniques, in order to test their ability to recreate a small, complete population of firms for the City of Hamilton, Ontario, in the year 1990. From the complete firm population (11, 499 in total), different levels of input data are extracted. The techniques are implemented with these different levels of input data, and outputs are compared to the entire population, in order to explicitly test their quality. The purpose of this chapter is fourfold: to implement the IPFSR and CO techniques for general use; to compare the two techniques, by measuring each one's ability to recreate the known population; to ensure that for both techniques, higher quality input data yields higher quality synthesized populations; to gain an idea of the minimum input data requirements for each technique to produce synthetic populations of reasonable quality. These objectives are realized through a series of comparisons of the outputs from both techniques, using various levels of input data, to the known population.

Huang and Williamson (2002), also compare the CO and IPFSR techniques; however their work was concerned with comparing the two methods without considering variations in the size of the micro sample and input tabulation levels. Our work extends their efforts to provide an in depth insight into the robustness of the two techniques under various conditions. Furthermore, we have the luxury of being able to compare synthetic outputs with a representation of the actual population, while Huang and Williamson (2002) did not have full information on their target population.

The remainder of this chapter is organized as follows. Section 3.2 presents the methods of analysis adopted to fulfill the objectives of the study. Section 3.3 provides empirical results and discussion, while Section 3.4 contains conclusions of the study.

3.2 Methods of Analysis

Programs to execute the CO and IPFSR methods were written in C++ and are collectively called the Synthpop program. For the CO method, tests were conducted to determine the minimum number of optimal iterations to perform in order to reasonably assure convergence. For synthesis of the 1990 Hamilton firm population, 6000 iterations were found to be sufficient. Similarly, for the IPFSR method, the minimum number of optimal iterations to be used as the cut-off point during the IPF portion of the program was determined to be 100. In both cases, the reason for limiting iterations was to find a balance between proper convergence of the techniques and the computer time required to run them. The optimal number of iterations refers to the cut-off point after which no remarkable gain in the quality of the generated synthesized population is realized. It should be noted that the number of iterations should not be compared between the two methods because each corresponds to a different process. In general, the number of iterations required for each method depends on the complexity of the input data used and the size of the synthesized population being created. Increasing the number of constraining characteristics or the number of categories in a given constraining characteristic usually requires an increase in computer runtime as well as the optimal number of iterations, for either method. For this analysis, the computer runtimes to generate a synthetic firm population were 12 and 10 minutes for the CO and IPFSR methods, respectively, using a Pentium 4 computer with a 2.0 GHz CPU.

3.2.1 Firm Population and Attributes

The 1990 firm population consisted of 11, 499 firms spread across the 127 Census Tracts (CTs) comprising the city of Hamilton, Ontario (see Figure 3.1).

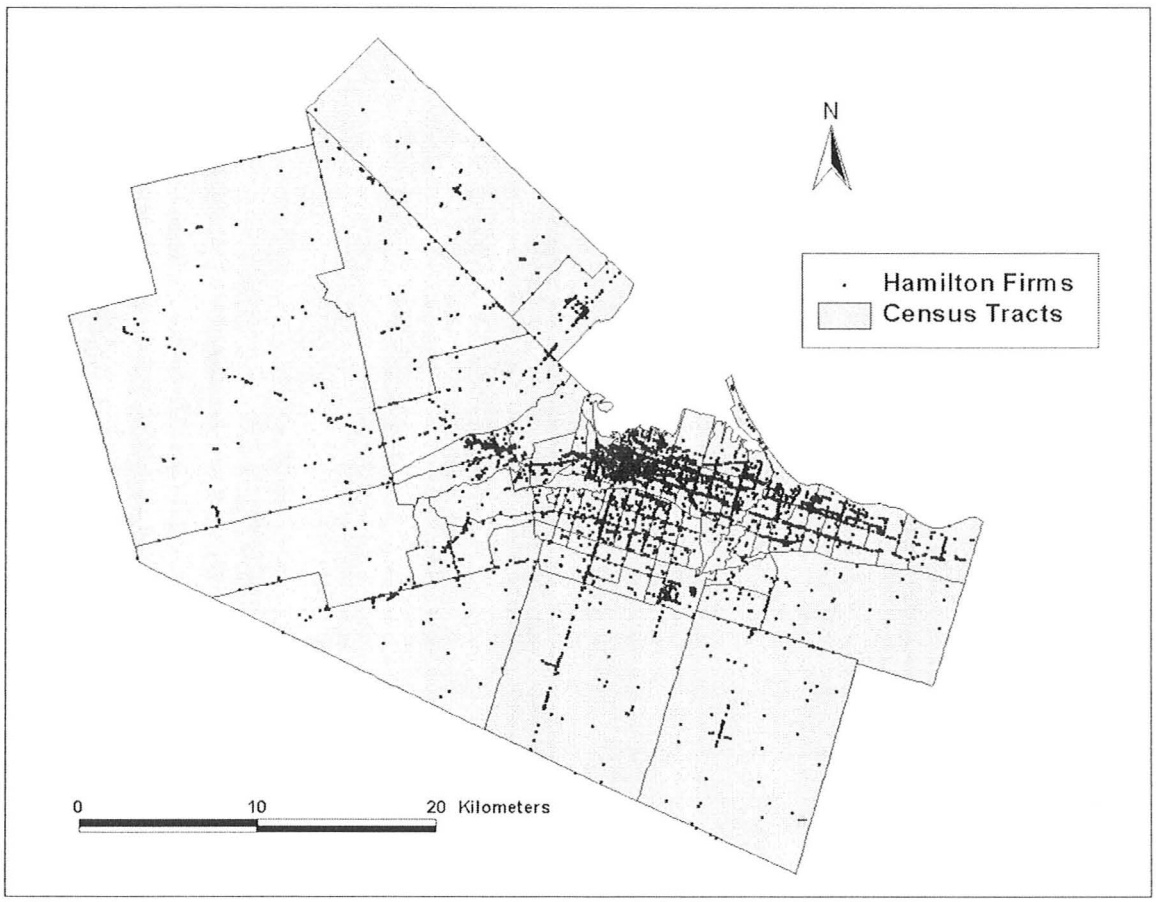

Figure 3.1: Firm locations in the Hamilton CMA, 1990

The attributes of each firm were: number of employees; census tract (CT); 3-digit Standard Industrial Classification (SIC) code. When synthetic populations were created using the IPFSR technique, firm attributes were assigned in the following order: number of employees; SIC; CT.

3.2.2 Input Data Derived from the Firm Population

The values of the attributes 'number of employees' and '3-digit SIC' were recoded to create representative cross-tabulations, as they would be available in practice. The values from the number of employees attribute were reclassified into 6 ordinal categories representing discrete employment ranges. The new attribute is referred to as EmpCat. Two reclassification schemes were used to represent the 3-digit SIC. The first

divided the 3-digit SIC codes into 14 mutually exclusive categories, and is referred to as SIC-E. The second scheme divided the 3-digit SIC codes into 68 mutually exclusive categories, and is referred to as SIC-2d. The SIC-E and SIC-2d codes are commonly used means of representing firm industrial classifications, with SIC-E codes providing less detail than SIC-2d codes. The CT attribute was not reclassified from its original 127 categories. From these attributes over the entire population, representative cross-tabulations were derived, as shown in Table 3.1a. It is useful to note that although a given firm could theoretically take on any of the 408 categories in cross-table 5 (see Table 3.1a), in the actual population only 335 of these categories are represented. In addition to the tabulations, eight different samples were taken randomly from the firm population, ranging in size from 1% to 100% of the entire population.

Table 3.1a: Two-way tabulations derived from the Hamilton firm population

Cross-Table	Description	Categories
1	Sic-E × CT	14 by 127
2	Sic-2d × CT	68 by 127
3	EmpCat × CT	6 by 127
4	EmpCat × SicE	84
5	EmpCat × Sic2d	408
6	(EmpCat × SicE) × CT	84 by 127
7	(EmpCat × Sic2d) × CT	408 by 127

Table 3.1b: Tabulation sets used as input to the synthesizing process

Name	Tabulation set	Relative Complexity
Level A	(EmpCat × CT) and (SIC-2d × CT)	Low
Level B	((EmpCat × SicE) × CT) and (SIC-2d × CT)	Medium
Level C	((EmpCat × Sic2d) × CT)	High

As input to the synthesizing process, three sets of tabulations were used, as listed in Table 3.1b. Note that the detail of the tabulations increases from set A to set C, where more detailed tabulation sets are considered to be those which provide more information on the structure of the population.

3.2.3 Comparing Actual and Synthesized Populations

For each combination of input tabulations and sample sizes, two synthetic populations were created using the CO method, and similarly for the IPFSR method. The reason for creating two populations for each combination of inputs was to give some indication of the variance in the accuracy of these populations. In all, 48 populations were synthesized using each method, yielding 96 synthesized populations in total. Both the synthetic populations as well as the actual population were then represented using the format of cross-table 7 (see Table 3.1a). Since there are 127 census tracts and 408 (EmpCat × Sic2d) categories, the tables representing the synthetic and actual populations contained 127 × 408 = 51,816 cells. Synthetic populations were then compared to the actual population using the Freeman-Tukey statistic, which is defined as follows:

$$FT^2 = 4\sum_i \sum_j \left(\sqrt{S_{ij}} - \sqrt{A_{ij}}\right)^2 \qquad \ldots(1)$$

Where S_{ij} is the ij^{th} cell from the synthesized population table and A_{ij} is the ij^{th} cell from the actual population.

The inspiration for using the Freeman-Tukey statistic comes from Voas and Williamson (2001, pp. 180-181), which provides a detailed discussion of goodness-of-fit measures for the evaluation of synthetic microdata. The Freeman-Tukey statistic follows a χ^2 distribution, with degrees of freedom equal to one less than the number of cells in the tables being compared, in our case 51,816 – 1 = 51,815, giving a 5% critical χ^2 value of 52,346. An FT^2 value greater than 52,346 indicates that the null hypothesis (that the two tables are independent) can be rejected, at a 95% level of confidence. At the same time, an FT^2 value of 0 indicates that the two tables match perfectly, and in general, the closer an FT^2 value is to 0, the better the match between the two tables. It is important to note that since both the synthesized and actual populations are being represented according to the C level of tabulation detail (see Table 3.1b), the FT^2 statistic is measuring the fit of synthesized populations to a representation of the actual population. Nonetheless, the representation of actual and synthesized populations in terms of (EmpCat × Sic2d) × CT contains enough detail to acceptably describe firm characteristics, for most purposes. A

further advantage of the Freeman-Tukey statistic is that the presence of zeros in either the synthetic or actual table is not problematic, as is the case with other statistics following the χ^2 distribution (Voas and Williamson, 2001, p. 181), which allows for a fairly detailed representation of the population.

In addition to the 96 synthetic populations created with the CO and IPFSR methods, two random populations were also created. These populations were random in the sense that the firms belonging to each census tract were randomly assigned EmpCat × Sic2d categories. However, the number of firms assigned to each CT was consistent with the actual population, mimicking the CO method outputs in this respect. The reason for creating these two random populations was to compare them to the actual population using the FT^2 statistic, and determine the sensitivity of the statistic to an arbitrarily created population. If the results of the FT^2 tests between these random populations and the actual population proved to be values less than the 5% critical χ^2 value, then the discernment of the FT^2 statistic would be brought into question.

Although 2 populations were synthesized with each method for each input combination, a more detailed analysis of the variations of populations produced with each method was desired. To accomplish this, 50 new populations were synthesized with each method, using the 5% population sample and level A tabulation detail ((EmpCat × CT) and (SIC-2d × CT)). Each synthetic population was then compared to the actual population in a similar manner as above. The results of these comparisons were used to assess the variance in the FT^2 values observed from each methods output. The choice of input sample size and tabulation detail to be used for the runs was, as mentioned, level A tabulations with a 5% sample. This combination was chosen due to its similarity to data that can be obtained in practice from publicly available sources.

3.2.4 Refining Synthetic Population Attributes

From each set of 50 synthesized populations, the population with the lowest FT^2 value (the "best" population) was chosen, to conduct further analysis on each method's ability to generate reliable estimates of the value of attributes associated with members of

the population over space. For this, we tested the continuous variable "*Size*", which reflects the number of employees of the firm. This variable is distinct from EmpCat, which was used during the synthesizing process. Here, we constrain the "*Size*" variable to 200 or fewer employees, in order to minimize the effect of outliers. The distribution over the census tracts of the "*Size*" variable resulting from each method's output were then compared to that of the actual population, minus those firms having greater than 200 employees. In the case of the population created with the CO method, each member of the population can be linked directly to a member of the 5% sample, and hence a value for "*Size*" can be assigned. Alternatively, Monte Carlo simulations could be used to estimate the size of the firm based on the value of EmpCat associated with the firm. Here, a cumulative probability distribution for each EmpCat category is developed from the "*Size*" variable of the existing 5% micro sample. Monte Carlo simulations are then run to generate an estimate of the size variable for each firm, using the constructed cumulative distribution and the synthesized EmpCat value. If the Monte Carlo simulations are executed n times, then the estimated size of firm i is calculated as the average of all simulated sizes, that is:

$$SS_i = \frac{\sum_{t=1}^{n} SS_i(t)}{n} \qquad \ldots(2)$$

Where $SS_i(t)$ is the size of firm i in simulation t.

Equation 2 was effective in generating a *Size* value for the members of the firm population synthesized using IPFSR. This is because unlike populations generated with the CO method, members of IPFSR populations are associated with an EmpCat value, but not an original sample member from which a value of *Size* can be taken. Monte Carlo simulations were executed 100 times and a value of SS_i for each firm i was estimated, with n = 100. For consistency and comparison purposes, equation 2 was also used to estimate the *Size* value for the firms of the CO population.

3.3 Results and Discussion

3.3.1 Sets of 48 (Varying Tabulation Levels and Sample Sizes)

The result of comparisons between the synthetic populations produced using the Combinatorial Optimization method and the actual population can be found in Table 3.2. These results are reproduced in graphical form in Figure 3.2a.

Table 3.2: Synthetic-Actual population comparisons (FT^2 critical value 52346)

Method	Sample Size	FT^2 for level A	FT^2 for level B	FT^2 for level C
CO	1%	33156	27288	18965
CO	1%	33171	27535	18966
IPFSR	1%	33956	26903	7869
IPFSR	1%	33831	27221	8201
CO	2.5%	23541	17488	9381
CO	2.5%	24048	17718	9381
IPFSR	2.5%	27332	21180	7707
IPFSR	2.5%	27332	20909	8048
CO	5%	22037	13440	5437
CO	5%	21806	12987	5440
IPFSR	5%	27541	19409	7921
IPFSR	5%	27093	18675	8119
CO	7.5%	21744	12443	3722
CO	7.5%	21304	12475	3709
IPFSR	7.5%	27776	18605	7837
IPFSR	7.5%	27555	18438	8032
CO	10%	20889	11751	2676
CO	10%	20889	11520	2694
IPFSR	10%	26688	17956	8139
IPFSR	10%	26559	18084	7837
CO	20%	20561	11012	1491
CO	20%	20563	10609	1525
IPFSR	20%	23670	16454	8071
IPFSR	20%	23592	16546	7828
CO	50%	19481	10231	568
CO	50%	19637	9929	698
IPFSR	50%	8131	8217	7940
IPFSR	50%	8126	7995	7929
CO	100%	19291	9899	540
CO	100%	19897	10183	615
IPFSR	100%	7802	8036	8162
IPFSR	100%	8097	8085	7941

Several general trends are immediately evident. First, as the level of tabular detail increases (i.e. increasing the cross-classification details) so does the accuracy of produced populations. At every sample size, populations produced with the tabular level A are less accurate than those produced with tabular level B. Similarly, populations produced with tabular level B are less accurate than those produced with tabular level C. In fact, the best A level population (which makes use of a 100% sample) is less accurate than the worst C level population (using only a 1% sample). This is not a complete surprise, however, since the C level tabulation is equivalent to the actual population representation, while the A level tabulations provide a far less accurate description of the population. It should be noted, however, that the above findings will hold so long as the cells in the cross-classification tables do not contain very few or no cases. If the latter occurs, then the additional cross-tabular information will provide unstable population estimates. In our analysis, none of the three levels of input tabular information (A, B and C) suffers this limitation.

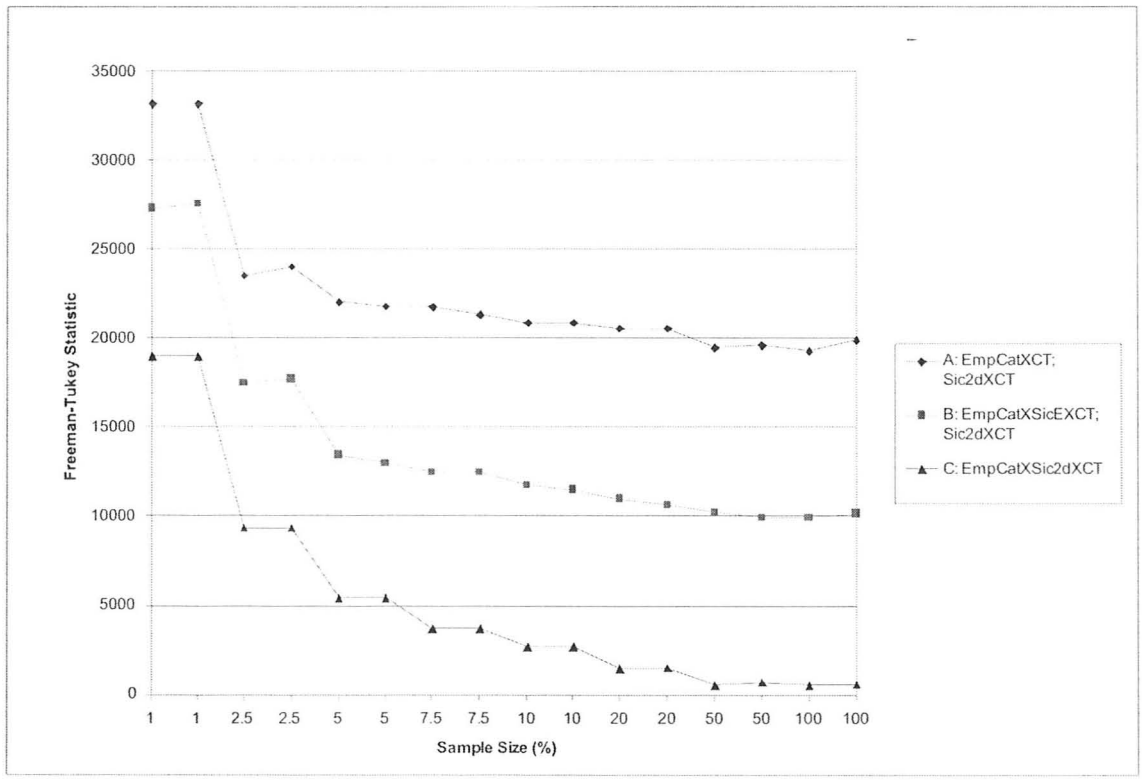

Figure 3.2a: FT^2, comparing CO outputs against the actual population

A second general trend observed in Figure 3.2a is that, at every level of tabular detail, there is an increasing relationship between the sample size used to create synthetic populations, and the accuracy of those populations. With the most detailed tabular input (level C) and the largest sample input (100%), the CO method produces a synthetic population which is almost identical to the actual population table ($FT^2 = 615$). All of this implies that there is a consistency to the CO method, where higher quality input yields higher quality synthetic populations. The result of comparisons between the synthetic populations produced using the Synthetic Reconstruction method and the actual population can again be found in Table 3.2. These results are reproduced in graphical form in Figure 3.2b. The results are similar to those from the CO method.

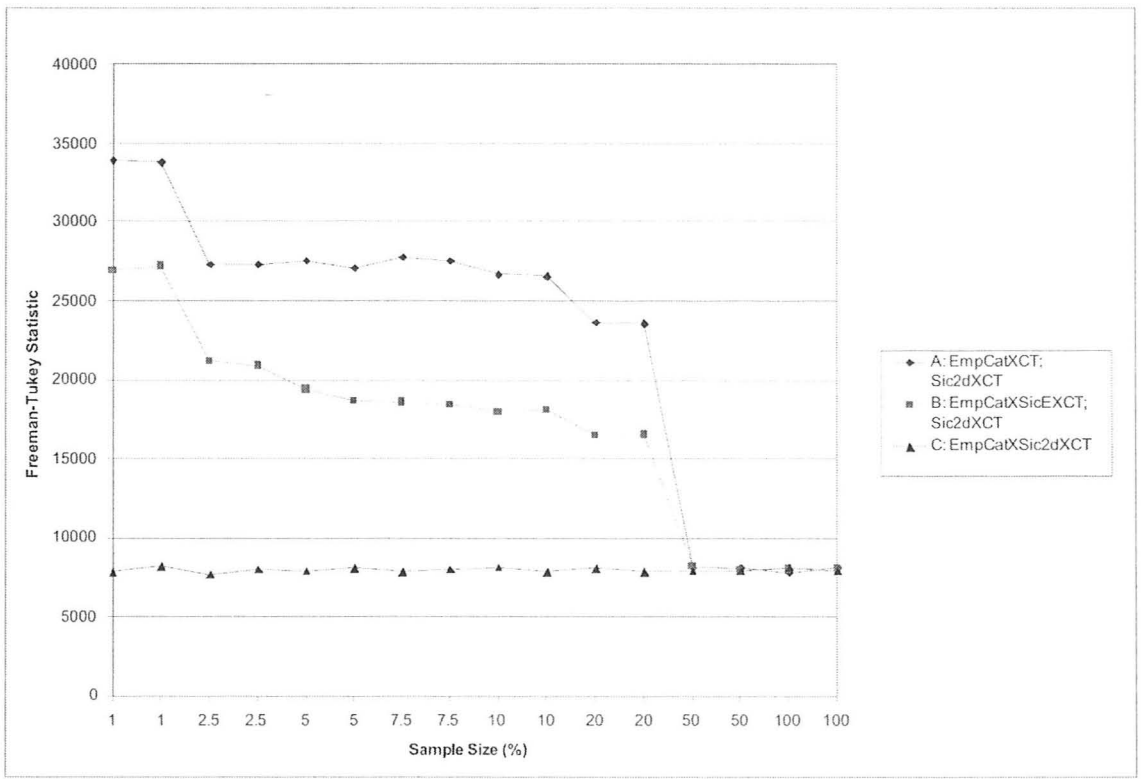

Figure 3.2b: FT^2, comparing IPFSR outputs against the actual population

Although the accuracy of synthetic populations produced with both the CO and IPFSR methods generally increases as the sample size increases, the gains earned by a

unit increase in sample size are not uniform. For the CO method, output accuracy increases drastically when the sample size is increased between 1% and 5%. From this point on, however, only gradual, small increases in accuracy can be made through further increases in sample size. These results hold true for all levels of tabulation detail, as can be clearly seen in Figure 3.2a. Therefore, if there is a cost associated with the collection of sample data, we recommend a 5% sample be used as input to the CO method. In the case of the IPFSR method, a slightly different pattern can be observed (see Figure 3.2b). Output accuracy increases drastically for sample size increases between 1% and 2.5%, and again for increases between 20% and 50%. For sample size increases between 2.5% and 20%, relatively little gain in accuracy is observed. This result holds true for all levels of tabular detail excepting C which, as discussed earlier, is insensitive to changes in sample size. We conclude that for the IPFSR method, a 2.5% sample input is sufficient where data collection is costly.

Table 3.3 contains some summary statistics describing the comparison results (sets of Freeman-Tukey statistics) of the 96 populations synthesized using the CO and IPFSR methods. The sample variance of the CO runs is 78,562,751, while that of the IPFSR runs is 77,423,594.

Table 3.3: FT^2 results from CO and IPFSR outputs, except the 50 run sets

Summary Statistic	CO results	IPFSR results
Max	33171	33956
Min	540	7707
Mean	14049	15945
Sample Variance	78562751	77423594

Despite this slight shortcoming of the CO results, the minimum FT^2 statistic from the CO outputs is 540, while that number is 7802 for the IPFSR outputs. This shows that given high levels of input information, the CO method is capable of producing more accurate synthetic populations than the IPFSR method. The maximum FT^2 statistics for CO and IPFSR outputs are 33,171 and 33,956, respectively. This shows that the worst CO outputs are better than the worst IPFSR outputs, and the best IPFSR outputs are worse

than the best CO outputs. It is important to note, however, that the 5% Chi-square critical value for the FT^2 statistic used in these comparisons is 52,346, meaning that all of the populations synthesized with both methods are acceptably similar to the actual population, according to the Freeman-Tukey test. On the other hand, the two random populations which were produced (see Table 3.4) have FT^2 values that exceed the 5% Chi-square critical value, lending further credit to the results of both the CO and IPFSR programs. Again referring to Table 3.3, the mean FT^2 statistic values for the CO and IPFSR methods were 14,049 and 15,945, respectively. This means that the 48 populations produced with the CO method were superior, on average, to those produced by the IPFSR method.

Table 3.4: FT^2 results from randomly synthesized populations

Run	FT^2	5% Critical χ^2	Fits? (Y/N)
R1	70769	52346	N
R2	70916	52346	N

The results of outputs from the CO and IPFSR method are directly compared in Figures 3.3a, 3.3b and 3.3c for levels of tabulation input A, B and C respectively. At level A (Figure 3.3a), the CO method outperforms the IPFSR method for all sample sizes, except for 50% and 100%. This implies that for a very large input sample size, and poor input tabulation detail, IPFSR provides superior synthetic populations to CO.

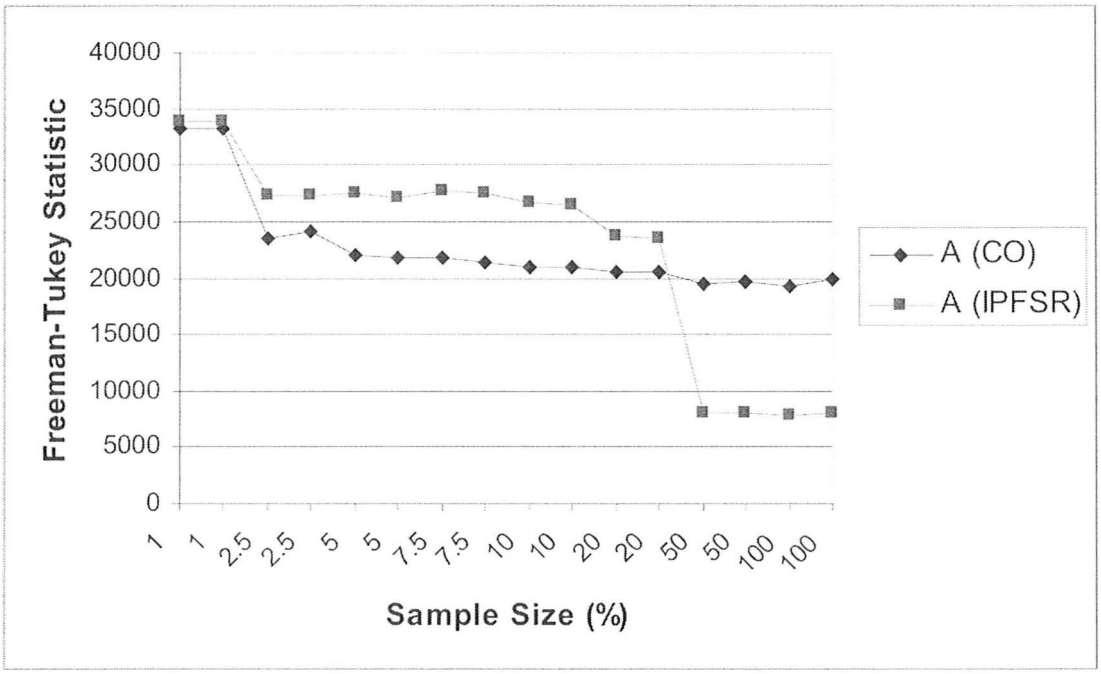

Figure 3.3a: FT^2, comparing IPFSR and CO outputs, tabulation level A

At level B (Figure 3.3b), we see a similar result as Figure 3.3a, with the IPFSR method outperforming the CO method only for sample sizes of 50% and 100%. In this case, however, the difference in FT^2 values between CO and IPFSR outputs for sample sizes 50% and 100% is negligible.

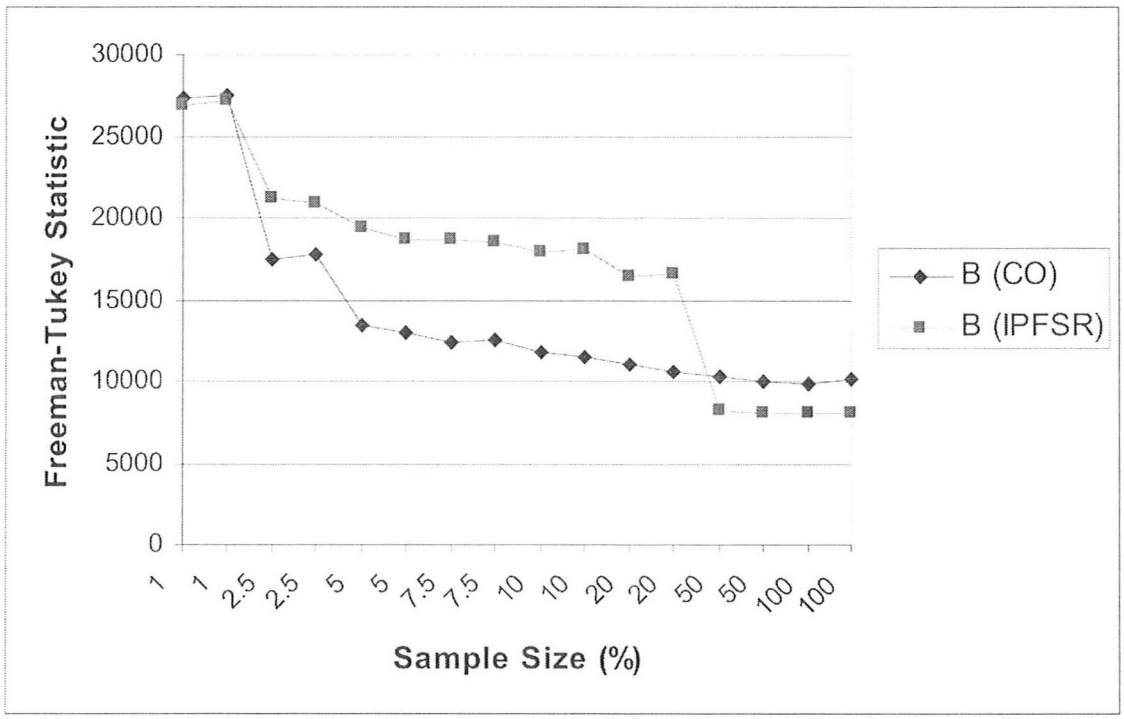

Figure 3.3b: FT^2, comparing IPFSR and CO outputs, tabulation level B

At level C (Figure 3.3c), the CO method outperforms the IPFSR method for all sample sizes larger than 2.5%. Given that level C tabulations are equivalent to the actual population representation, and that a 100% sample could hence be derived from it, the most important comparison between CO and IPFSR methods at this tabulation level is between the 100% sample size outputs. Here the CO method outperforms the IPFSR method, meaning that with full information, the CO method is preferable for population synthesis. In the more likely scenarios of tabulation levels A and B, with a relatively small sample size (20% or smaller), the CO method outperforms the IPFSR method. Hence, we recommend the CO method over the IPFSR method when full population information is available, as well as when limited population information is available. The only case where IPFSR is recommended is when tabular information is of low detail, and the sample size is large (greater than 20%).

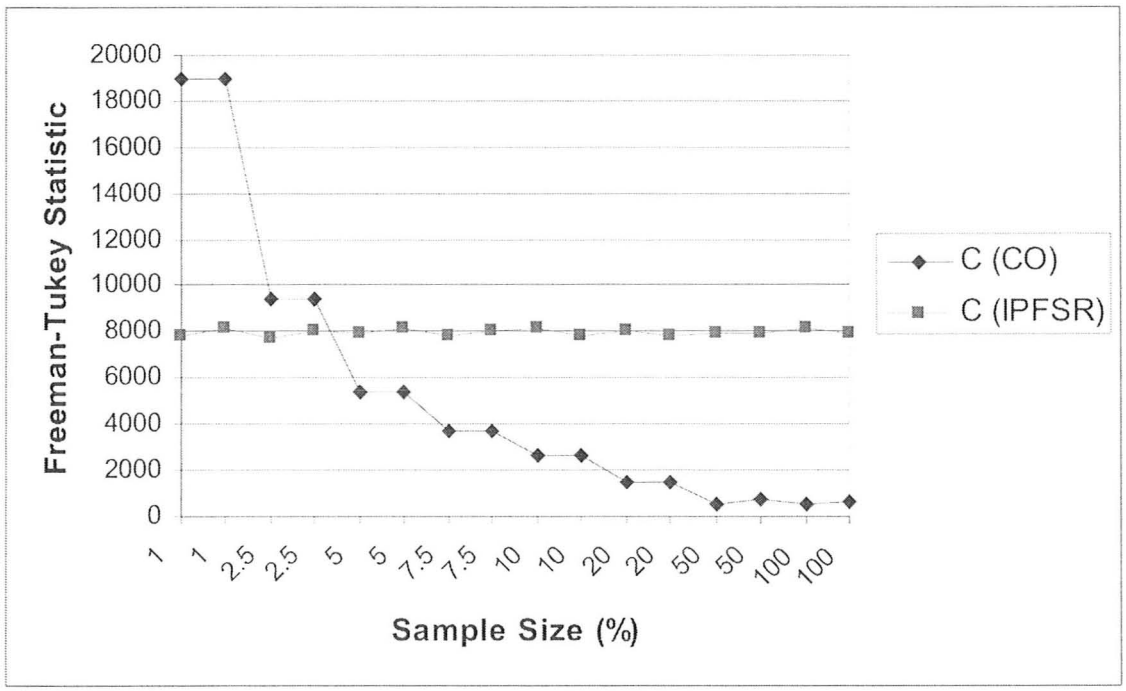

Figure 3.3c: FT^2, comparing IPFSR and CO outputs, tabulation level C

3.3.2 Sets of 50 (Tabulation Level A, 5% Sample Size)

For the two sets of fifty populations produced with an input of tabulation detail A and 5% sample size, summary statistics of the resulting FT^2 values can be found in Table 3.5. Of particular note, the mean of the CO produced populations is 21,539, while the mean of the IPFSR populations is 27,301. Furthermore, the maximum FT^2 value from the CO populations is 21,928, while the minimum FT^2 value from the IPFSR populations is 26,942. Thus, CO outputs at this level of input data are consistently closer to the tabular representation of the actual population than their IPFSR counterparts. This result is important, because the level of input data used in these runs is typical of what can be obtained by a researcher in practice. The FT^2 values from the two output sets can be found in graphical form in Figure 3.4. Although it cannot be distinguished from Figure 3.4, the standard deviation of the CO outputs is larger than that of the IPFSR outputs. In particular, the standard deviation of the CO outputs is 180, while that of the IPFSR outputs is 174. Of course, the fact that all CO produced populations in the set have

significantly lower FT^2 values than the most accurate IPFSR population outweighs the advantage these latter populations have in terms of standard deviation.

Table 3.5: FT^2 results from IPFSR and CO 50 run sets

Summary Statistic	CO	IPFSR
Mean	21539	27301
Median	21553	27304
Standard Deviation	180	174
Sample Variance	32419	30260
Kurtosis	-0.6049	0.0834
Skewness	-0.1791	0.0271
Range	774	816
Minimum	21155	26942
Maximum	21928	27759
Count	50	50

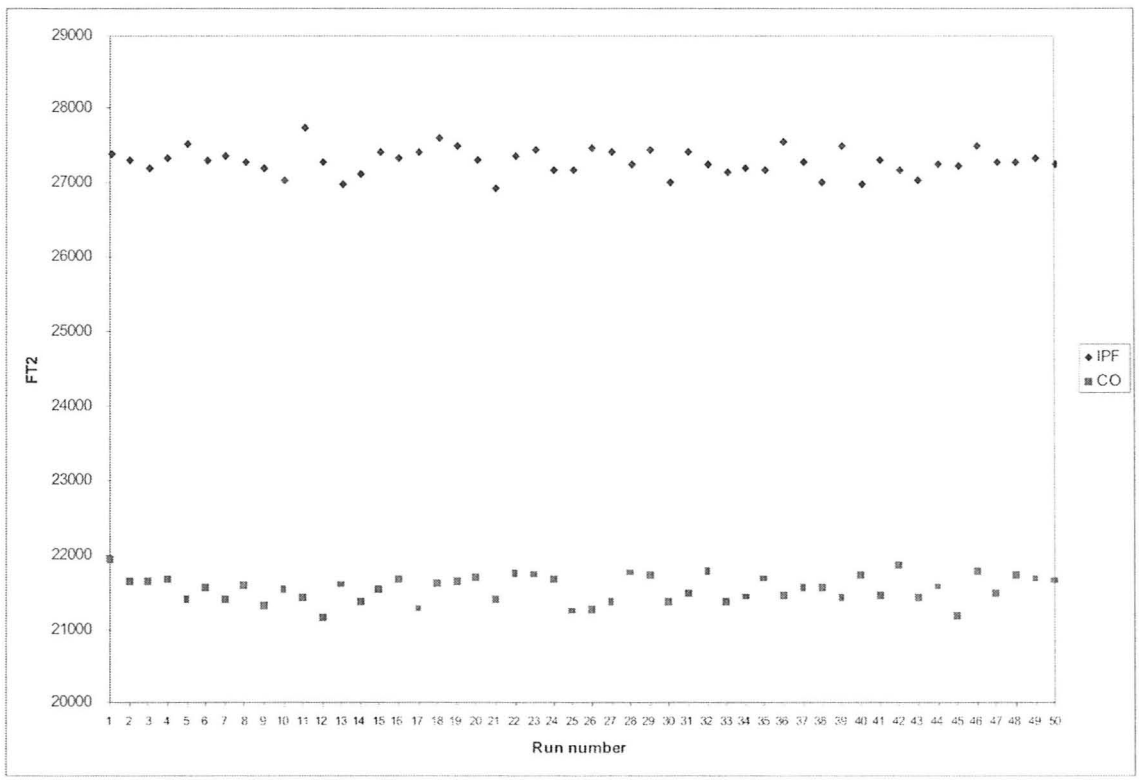

Figure 3.4: FT^2, comparing IPFSR and CO produced populations, created with tabulation level A and 5% sample size

3.3.3 Refined *Size* Variable Comparisons

The population from the set of 50 CO outputs having the lowest FT^2 value (the best population) was selected to ascertain how it reproduced the variable "*Size*" (the number of firm employees) for firms having fewer than 201 employees, and similarly with a population from the 50 IPFSR outputs. This is to ascertain how well the synthesized populations are capable of re-creating the continuous "*Size*" distribution of the majority of the population, that is to say, the population without outliers. Summary statistics on the set of "*Size*" values derived from the CO and IPFSR populations via Monte Carlo simulations, as well as those derived from the CO population via linking to the original sample (referred to as 'sample linked'), as well as those from the actual population are presented in Table 3.6a.

Table 3.6a: "*Size*" distribution statistics, Size ≤ 200

Summary Statistic	Actual Population	CO (by joining to 5% sample)	CO (Average of 100 simulated sizes)	IPFSR (Average of 100 simulated sizes)
Mean	10.8	11.2	11.1	11.0
Median	4	4	3	3
Mode	2	2	3	3
Standard Deviation	20.0	21.6	19.8	19.4
Sample Variance	400.2	464.6	390.6	378.3
Kurtosis	28.5	24.2	13.2	13.7
Skewness	4.8	4.6	3.6	3.7
Range	199	174	108	109
Minimum	1	1	2	2
Maximum	200	175	111	111
Sum	123090	127387	126581	124961

The mean value of number of employees is 11.1, 11.2, 11.0 and 10.8 for the CO, CO sample linked, IPFSR and actual population respectively. The standard deviations of the "*Size*" values are 19.8, 21.6, 19.4 and 20.0 for the CO, CO sample linked, IPFSR and actual population respectively. So, all three methods create "*Size*" distributions with means and standard deviations close to the actual population values. The ranges of

"*Size*" values from the CO, CO sample linked, IPFSR and actual populations are 111, 175, 111 and 200 respectively. Here the CO and IPFSR ranges are the same, because they both depend on the range of values in the 5% sample. Table 3.6b shows correlations between the synthesized and actual distributions of the "*Size*" values over space.

Table 3.6b: Correlation of "*Size*" distributions over space, Size ≤ 200

Distributions Being Compared	**Correlation**
Actual Population Vs. CO (joining with 5% sample)	0.9951
Actual Population Vs. CO (Avg. 100 simulated sizes)	0.9965
Actual Population Vs. IPFSR (Avg. 100 simulated sizes)	0.9798

The correlation between the CO and actual distribution is 0.9965, while that value is 0.9798 between the IPFSR and actual distribution. The correlation between the CO sample linked and actual distribution is 0.9951. Across all of the above measures, the "*Size*" distribution is best recreated by the CO distribution created via Monte Carlo sampling. The residuals of each of the three "*Size*" distributions from the actual population are depicted graphically over space in Figures 3.5a, 3.5b and 3.5c. Although certain census tracts are consistently under or over predicted across each figure, it is in Figure 3.5b (IPFSR residuals) that the largest number of census tracts are extremely over or under predicted. In general, the distributions of the "*Size*" variable are best replicated by the CO method, using Monte Carlo simulations based on the sample distribution of firm sizes.

Figure 3.5a: '*Size*' Residuals Over Space, CO Monte Carlo

Figure 3.5b: '*Size*' Residuals Over Space, IPFSR Monte Carlo

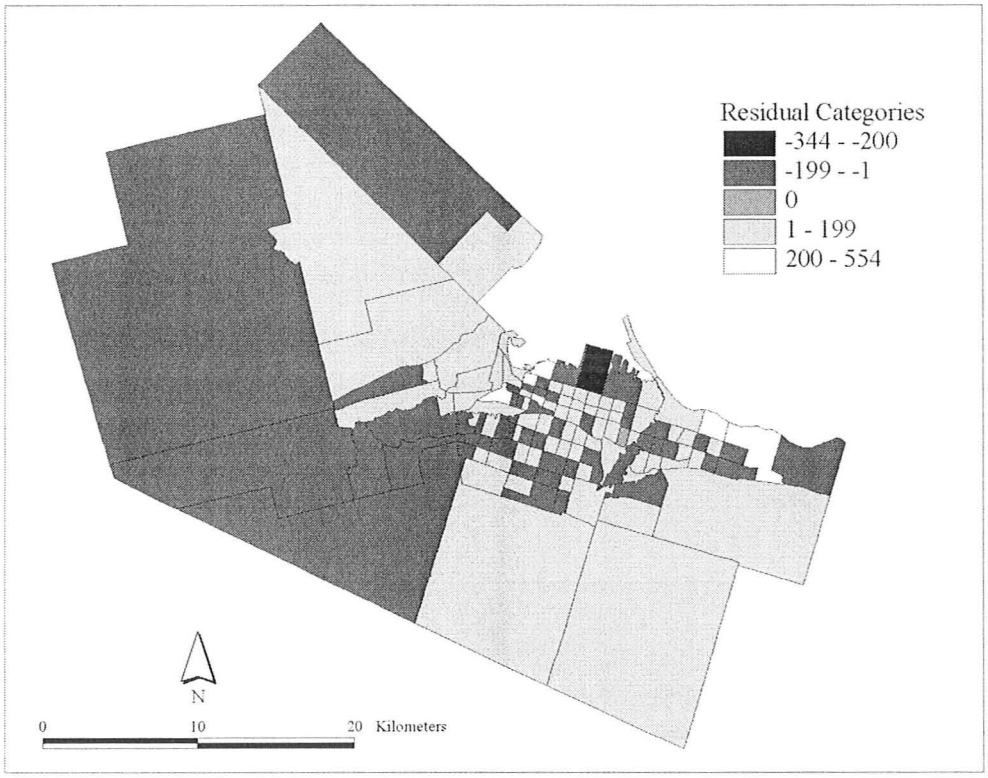

Figure 3.5c: '*Size*' Residuals Over Space, CO Sample Linked

3.4 Conclusions

It has been shown in this chapter that both the CO and IPFSR methods are capable of synthesizing a small population of firms that, according to statistical tests, is similar to the actual population of firms. This result holds true even when the coarsest levels of data are used as input to the methods. That being said, as the levels of tabulation detail used as input to the methods increases, the resulting populations are found to be more accurate. Similarly, as input sample size increases, resulting populations experience gains in accuracy. Increases in tabulation detail influence resulting population accuracy more than increases in sample size, for both methods. In general, using similar input data, the CO method produces populations that are more accurate than those produced via the IPFSR method. We therefore recommend the CO method as a viable alternative over the IPFSR method for the synthesis of population datasets.

There are several limitations to this study, which should be mentioned. First, our work was done on a relatively small population. The latter corresponds to population sizes in the order of tens of thousands of individuals. Where large populations, in the order of hundreds of thousands or millions of individuals, need to be synthesized, the results of this chapter may not completely apply. However, the biggest issue with large populations is computing time. Secondly, our firm population did not contain many attributes. Synthesizing a large population whose members have many attributes is the next logical step, building on the results of this chapter.

In Chapter 4 of this work, the CO method is used to synthesize a population of individuals for the City of Hamilton. This constitutes an excellent test of the CO method's capabilities, since the population of individuals is relatively large (457,325 members) and is constrained by six detailed tabulations. Following this, and the synthesis of a population of households for the City of Hamilton, the two synthesized populations are 'linked' to form a comprehensive synthetic population.

As future work, methods of reducing the computer run-time required by the IPFSR and CO methods will be explored. For the IPFSR method, the most time consuming element is the multiple applications of IPF required to create a complete multiway table. This task can equivalently be solved using entropy maximization techniques, which may be more time efficient (see: Johnston and Pattie, 1993). For the CO method, the most time consuming element is the many swaps required to bring a random sub-set of sample members into line with constraining tabulations, for a given zone. However, once this task has been completed for one zone, the resulting list of members could be used as a starting point for neighbouring zones (especially those zones whose constraining distributions are highly correlated with the completed zone). This would allow for a more explicit incorporation of space into the CO method, as well as possible time savings.

Chapter Four: A Comprehensive Synthetic Population for Hamilton, Ontario

4.1 Introduction

In this Chapter, a comprehensive synthetic population is created for the residential mobility microsimulation model "URM-Microsim" (see: Svinterikou & Kanaroglou, 2006; Svinterikou, 2007). The model was originally used in the European context, for the City of Mytilene, Greece. Here, the model is to be adapted to the Canadian context, through the input of a synthetic population for the City of Hamilton, Ontario, in the year 1996. For this purpose, the input synthetic population must consist of four hierarchically related elemental populations, namely: individuals; households; dwelling units; and buildings. The synthesized attributes of each elemental population are chosen to meet the requirements of URM-Microsim. The population is created for the year 1996.

Section 4.2 provides an overview of the required comprehensive population, as well as the methods employed to create it. Section 4.3 describes the synthesis of elements of the synthetic population, namely: individuals; households; dwellings; and buildings. Section 4.4 discusses the procedures used to link the elements of the synthetic population into a hierarchically related 'comprehensive' population. Section 4.5 provides some validation of the comprehensive synthetic population, as well as its elements. Finally, section 4.6 presents conclusions of the study, as well as future research avenues.

4.2 Overview

The URM-Microsim model requires complete sets of individuals, households, dwelling units and buildings over a given study area. Each individual belongs to a household and each household belongs to a dwelling unit, which in turn belongs to a building. Buildings belong to a specific location in the study area, which indirectly specifies the location of individuals, households and dwellings. Each elemental population set making up the comprehensive population is endowed with a set of attributes. For instance, individuals each have attributes of age, sex and income, among others. It is important to note that the attributes of each elemental population are chosen

(or synthesized) because they are required as input by the URM-Microsim model (although some additional attributes are included for the purpose of creating a richer population for possible future research). When a given attribute is categorical (as in the case of structure-type for buildings), it is important that the categories are equal to, or can be translated into the particular categorization scheme required by URM-Microsim. The exact attributes required by URM-Microsim for each complete population can be found in Tables 4.1a, 4.1b, 4.1c and 4.1d.

The URM-Microsim model uses the comprehensive input population as a base upon which to develop simulations of the future state of the population. Individuals are aged, and may experience a variety of demographic events such as the birth of a child, or a move to another residence. A series of probabilistic rules based on the theory of residential mobility and demographics are applied to the base population in order to determine the events which will befall its members. Note that events occur not only to individuals, but to members of each elemental population. For instance, buildings may be demolished or constructed, in turn eliminating or creating associated dwelling units. The latter events are driven by the land development process.

Table 4.1a: Attributes of Individuals required by URM-Microsim

Attribute	Description
Sex	
Date of Birth	
Marital Status	
Position in household	Household head, child, couple, flat-mate
Education	Highest level attained
Employment Status	
Income	
Occupation	If employed
Industry	If employed

Table 4.1b: Attributes of Households required by URM-Microsim

Attribute	Description
Size	Number of household members
Income	Sum of members' incomes
Tenure	Does household rent or own dwelling?

Table 4.1c: Attributes of Dwelling Units required by URM-Microsim

Attribute	Description
Size	Floor space
Number of Rooms	
Market value	If owned
Rent value	If rented (monthly)
Availability	For rent or sale?

Table 4.1d: Attributes of Buildings required by URM-Microsim

Attribute	Description
Type	Indication of structure
Floor space	In square feet
Number of floors	
Number of dwellings	
Year of construction	

As a prerequisite to creating the comprehensive population for Hamilton, the four complete population data sets (Individuals, Households, Dwellings, and Buildings) are first created individually. This is necessary because the techniques and input data at our disposal do not allow for direct synthesis of the comprehensive population as a whole. The population of Individuals and Households are both synthesized using the Combinatorial Optimization (CO) method. Input data for this task are taken entirely from the 1996 Canadian Census. (See Chapter 2 for a detailed description of the CO method). Members of the Individual and Household populations are spatially resolved to the Census Tract (CT) level, which divides the City of Hamilton into 127 mutually exclusive geographic areas. The population of Buildings is derived from City of Hamilton parcel data, which are compiled from various resources at the Center for Spatial Analysis, at

McMaster University. In it basic form, the parcel data include attributes on the amount of floor space, the property code (which indicates the structural type of the building), and X, Y spatial coordinates. For each building, a set of dwelling units is created, using the building's attributes to derive the number of dwellings as well as dwelling attributes. For instance, dwellings take on the same spatial coordinates of the building to which they belong, while the floor space and property code of a building help to determine the number of dwellings created, as well as the number of rooms per dwelling.

Once the four complete population data sets have been created, they are linked together hierarchically in order to yield the desired comprehensive population. As mentioned in the description above, dwelling units are naturally linked to buildings from their creation. Therefore, the remaining tasks are to link households to dwelling units, and to link individuals to households. In the case of linking households to dwelling units, several rules apply. Among these: a dwelling unit must be located in the same CT as a household; and the number of household members should not exceed the number of rooms in a dwelling, in most cases. When linking individuals to households, a multitude of conditions must be respected. Among these: the relationships that exist between household members must be logical (for example, the age of a son or daughter of the 'household maintainer' must be within a certain range, given the age of the household maintainer); the individuals must be located in the same CT as the household.

Figure 4.1 provides a graphic representation of the relationship between individuals, households, dwellings, buildings and parcels. Here, two theoretical parcels are presented, each containing one residential building. While Building 1 contains only one dwelling unit (as is typical for a single family home), Building 2 contains 3 dwellings. Note that Dwelling 4 is vacant, and hence is not associated with a household. Further note that each household contains one or more individuals.

It is worth mentioning some supporting literature at this point. In particular, the data structure used in the LOCSIM microsimulation model (see: Hooimeijer and Oskamp, 1996, 2000; Oskamp, 1997) is similar to that required by URM-Microsim.

Here, although the database is adequately described, very few details are provided on the mechanisms used to link individuals to households, and households to dwelling units.

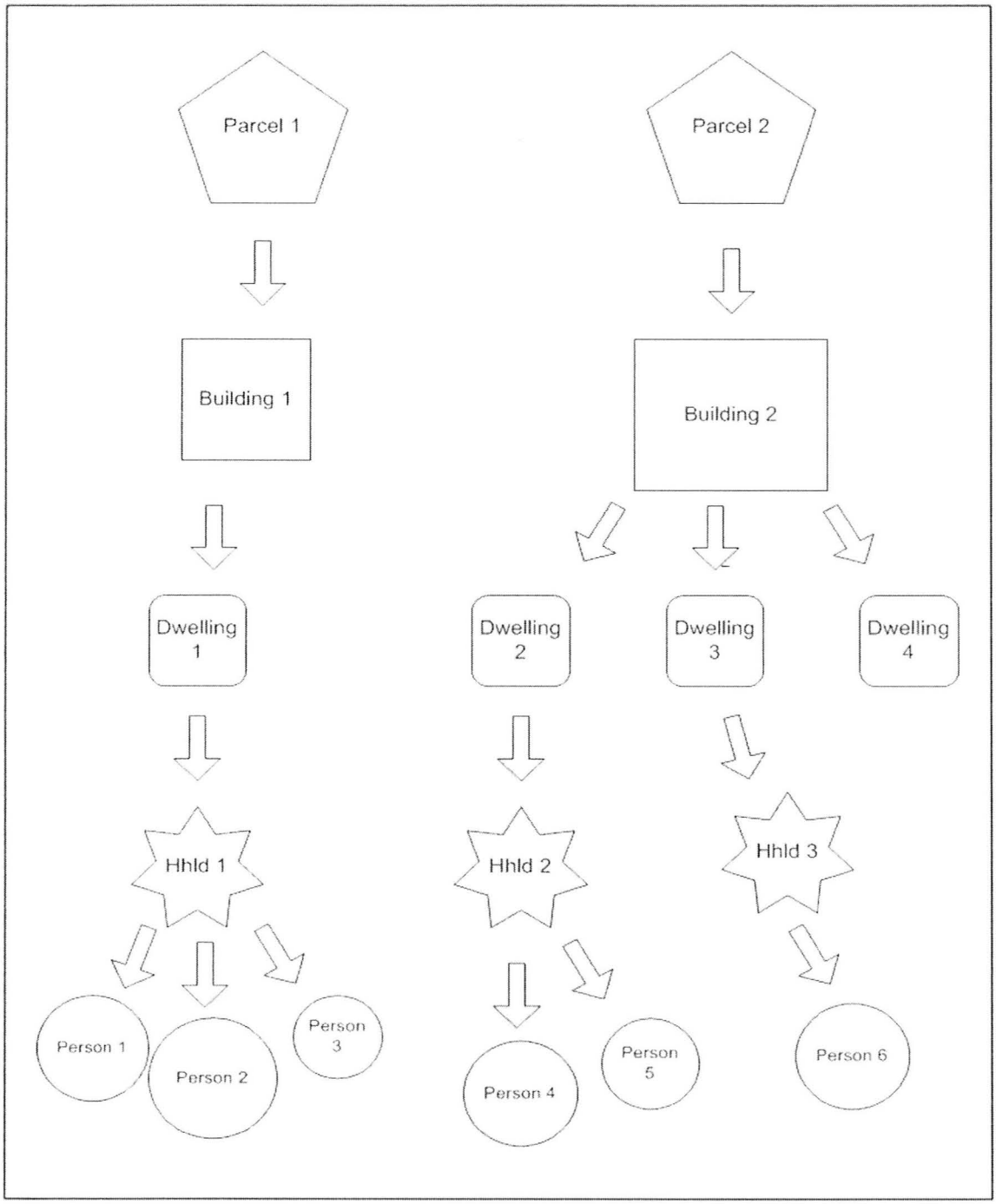

Figure 4.1: Relationships between elements of the comprehensive synthetic population

Other work exploring the nature of inter-household relationships, such as Rindfuss et al. (2004) compliments the task of devising rules to link individuals to households. Although it is tangential to our particular goals in this work, Ermisch and Francesconi (2000) provide an interesting example of an attempt to link households to land parcels. For the most part, the linkages between elemental components of the comprehensive synthetic population were made in an ad-hoc fashion, in order to meet the needs of the URM-Microsim model.

An in-depth description of each stage of the creation of the 1996 comprehensive population for Hamilton, Ontario can be found in sections 4.3 and 4.4.

4.3 Synthesizing the Elements of the Comprehensive Population

4.3.1 Synthesizing Individuals

A population of individuals for the City of Hamilton, in the year 1996, was synthesized using the Combinatorial Optimization (CO) method (see Ch.2 for a detailed description of the CO method). The required input data for the CO method is a set of population characteristic tabulations over space, as well as a non-spatial micro-sample from the population, both of which were obtained from publicly available census data. Population characteristics are expressed as categorical variables for which every member of the population takes on a value, such as age or income. Population characteristic tabulations show the distribution of these categorical variables, for the actual population. In general, tabulations are chosen to constrain characteristics of the population which are considered important for a given study. For instance, if gender is an important consideration in a study, then any population synthesized for the study should constrain for gender during the synthesizing process, to ensure that the correct number of males and females are created in each zone of the study area. For the Hamilton population, taking into consideration the requirements for the URM-Microsim model, the following six population characteristics were constrained for during the CO method: Sex by Age by Relationship to Person 1; Marital Status; Sex by Income; Sex by Employment; 5-Year

Mobility Status; Citizenship. Note that 'Person 1' is a designation of individuals from the census, also referred to as the 'household maintainer'. For each population characteristic, the 'constraining tabulations' related to these variables show the variables' distribution in the actual Hamilton population, for each CT. Due to data blurring and masking in the 1996 Census, 3 CTs were omitted from the synthesized population. These CTs had a combined size of approximately 250 individuals, while the remaining 124 CTs contained 457,325 individuals. Tables 4.2a, 4.2b, 4.2c, 4.2d, 4.2e, and 4.2f (located in Appendix VI) list the categories of each constraining variable, and show their distribution in the actual 1996 Hamilton population. It is important to emphasize that these constraining characteristics were chosen because they are fundamental to the URM-Microsim model.

In all cases except for 'Sex by Age by Relationship to Person 1', the constraining tabulations were derived from publicly available Canadian Census data, for 1996. The tabulation for 'Sex by Age by Relationship to Person 1' was acquired from Statistics Canada, and consists of data collected in the 1996 Census which was not made publicly available. This tabulation played a key role in the synthesis of individuals, due in part to its detail. In particular, the sex and age of individuals, as well as their relationship to their household maintainer ('Person 1') is specified from a set of 180 categories. Age is broken into 18 groups of 5-year spans (0-4, 5-9, ... , 80-84, 85+), and labeled as A1 through A18, respectively. Relation to Person 1 originally contained 10 categories, but was regrouped into the following 5 categories for our purposes: Person 1; Person 1's spouse or common-law partner; Person 1's son or daughter; Other relatives of person 1; Person not related to Person 1. These are labeled as R1 through R5, respectively. The regrouping of 'Relation to Person 1' to 5 categories shrunk the total number of categories in 'Sex by Age by Relationship to Person 1' from 360 to 180, given the 18 age, and 2 sex categories. This smaller tabulation was more desirable for input to the CO method. See Table 4.2f (located in Appendix VI) for the particular categories included in the 'Sex by Age by Relation to Person 1' tabulation. The category codes can be read as the sex classification, followed by the age group, followed by the relationship classification. For

example, F A7 R3 refers to individuals who are female, 30 – 34 years of age, and the daughter of Person 1 in their household. See Tables 4.3a and 4.3b (located in Appendix VI) for the distributions of the original and regrouped 'Relation to Person 1' variables in the actual Hamilton population. These provide a fascinating look at the structure of individuals within their households. The number of Person 1's is 177,005 which correspond to the number of households in the study area. There are 104,860 individuals reporting to be the 'spouse or common-law partner' of Person 1, which implies that an equal number of Person 1's live with their spouse. This leaves 72,145 'single' Person 1s, which account for approximately 41% of households in the study area. There are 152,945 sons or daughters of Person 1. Together with Person 1s and their spouses, these three categories of individuals account for just over 95% of the population. This fact further justifies the decision to regroup 'Relation to Person 1' from 10 down to 5 categories (Tables 4.3a to 4.3b in Appendix VI), since under 5% of the population is affected. Although it is beyond the scope of this work, it would be interesting to study the 'Relation to Person 1' variable over time, to examine temporal changes in household living arrangements. Tables 4.3c and 4.3d (located in Appendix VI) show the sex and age group distributions of the 1996 Hamilton population. When sex, age and relationship to Person 1 are combined into the variable 'Sex by Age by Relation to Person 1', we are provided with a detailed look at the population structure in each CT making up the study area. This information is not only useful for synthesizing individuals, but invaluable when assigning individuals to households (see section 4.4.1).

The publicly available constraining tabulations, whose categories are shown in Tables 4.2a through 4.2e (located in Appendix VI), were altered to match certain aspects of the 'Sex by Age by Relation to Person 1' tabulation. This was necessary since the CO method requires consistency amongst its input tabulations. The 'raw' tabulations provided by Statistics Canada usually do not agree on the number of individuals per CT, the number of males and females per CT, or the number of persons under the age of 15 per CT. In all cases, the number of individuals per CT was standardized to match the values in 'Sex by Age by Relation to Person 1'. This was accomplished by applying a

multiplier to the values of constraining tabulations, for each CT. The values were then rounded, and if any error remained, it was nullified by subtraction or addition from the most populous category. A similar method was used to standardize for sex and age. As an example, the tabulation 'Sex by Employment' (Table 4.2b in Appendix VI) required these further standardizations, since it specifies not only the sex of individuals, but the sex of individuals under the age of 15.

The micro-sample of individuals used during the synthesizing process contains 17,158 records, or approximately 3.75% of the actual population of Hamilton, for the year 1996. Each record contains values for all of the constraining variables, as required by the CO method. It is interesting to note that 'Relationship to Person 1' is included as an attribute of sample members, despite the fact that its tabulation for the entire city was not made publicly available. Additional attributes, collected as part of the census, are also provided for each record; however, none of these attributes are spatial. In total, each record contains 128 attribute values, including the 6 which correspond to the categories of the constraining variables. Once a synthesized population has been created, some or all of the additional attributes from the micro-sample may be included as attributes of the synthetic individuals. This is possible because the synthetic individuals are chosen directly from the micro-sample, meaning that each synthetic individual can be linked to a record in the micro-sample. Since these additional attribute variables were not constrained for during the synthesizing process, there is no guarantee that their distributions will match those of the actual population. In general, the more highly correlated an additional variable is with one or more of the constrained variables, the greater the probability that the distribution of the additional variable in the synthetic population will match that of the actual population. There are several reasons why 'additional' variables of interest are not constrained for during the synthesizing process. For one, if the tabulations of such variables over the study area are not available, then the variable cannot be constrained for. Also, the CO method will converge quicker if fewer (or less complex) tabulations are constrained for during the process. In this work, although the URM-Microsim model requires nine specific attribute variables for

individuals (see Table 4.1a), only six were constrained for (Sex by Age by Relationship to Person 1; Marital Status; Sex by Income; Sex by Employment; 5-Year Mobility Status; Citizenship). The choice of constraining variables represents a compromise which was deemed sufficient for the purposes of synthesizing the population of individuals. In cases where a large number of attribute variables are required for a synthetic population, a more quantitative approach to deciding which variables to use as constraints is possible, by employing methods of factor analysis. Factor analysis allows for a determination of which variables are correlated with each other, informing the choice of a minimum number of constraining characteristics from the set of required variables. In our case, the overall number of variables required by URM-Microsim is relatively small, and the three non-constrained variables were assumed to be correlated with one or more of the six constrained variables. Note that validation of the three non-constrained variables is presented in section 4.5.1.

The three non-constrained variables (or 'additional variables') which were added to the population via linking with the micro-sample were: HLOSP; IND80P; and OCC91P. HLOSP indicates the highest level of schooling individuals have obtained. IND80P characterizes an individual's industrial classification (1980 SIC), while OCC91P classifies the occupation of individuals. See Tables 4.4a, 4.4b and 4.4c (located in Appendix VI) for details on the categories pertaining to each of the additional variables. The inclusion of the three additional variables as well as the six constrained for variables, as attributes of the individuals, satisfies the requirements of URM-Microsim. The final synthetic population of individuals contained the following attributes: Individual ID; CT; Original sample number (i); Citizenship; Sex by Employment; 5-year Mobility Status; Sex by Income; Marital Status; Sex by Age by Relationship to Person 1; Highest Level of Schooling; Industry; Occupation.

Due to the large amount of time required by the CO method to synthesize all of the individuals in the City of Hamilton, 8 sub-populations of individuals were synthesized instead. In order to do this, the input tabulations were each broken into 8 sub-tabulations, representing approximately 14 to 16 CTs each, and containing fewer

than 65,000 individuals. The original micro-sample was used for each sub-population. In this way, the CO method could be run for each of the sub-populations simultaneously, using several different computers or a computer endowed with multiple processors. The average time required to complete a sub-population of individuals was 137 hours, using a computer with a 2 GHz CPU. The collective time required to synthesize the entire population was 1094 hours; or approximately 46 days. Synthesizing the individuals in this way had the extra advantage that each sub-population could be opened in Microsoft Excel, for quick and easy analysis. Furthermore, when performing any kind of operation on the individuals, such as linking them to households, one of the sub-populations could be used as a sample for testing purposes.

Given the large amount of time required by the CO method to synthesize a population of individuals for a city the size of Hamilton, it is natural to pursue methods of decreasing computer run-time. One alternative, within the framework of the CO method, is to increase the accuracy of the initial draw of individuals from the micro-sample, for a given CT. Since the CO method operates on each CT independently, the final population of a given CT can be used as a starting draw for neighbouring CTs. When a finished CT varies in total population size from an unfinished neighbouring CT, a randomly selected multiple of the finished population could be taken as the initial draw for the neighbouring CT. This modification of the CO method has the added advantage of a more explicit incorporation of space, given the assumption that neighbouring CTs will have similar populations, in terms of constrained attributes. Increasing the speed of the CO method will be the subject of future work.

4.3.2 Synthesizing Households

A population of households for the City of Hamilton was synthesized using the CO method, and generally following the technique used to synthesize the population of individuals (see Section 4.3.1). In this case, the tabulations and micro-sample used as input to the CO method were all publicly available, from the 1996 Canadian Census. The constraining tabulations were: Size (the number of household members); Income (the

collective incomes of household members); Structure (the structure of the building in which the household resides); and Tenure (whether the household is owned or rented). Tables 4.5a, 4.5b, 4.5c and 4.5d (located in Appendix VI) list the categories of each constraining variable, respectively, and show their distribution in the actual 1996 Hamilton population. An examination of the tables shows that there are 177,005 households making up the population. The number of households in each CT was determined by the number of Person 1's, from the 'Sex by Age by Relation to Person 1' tabulation used for the synthesis of individuals (see section 4.3.1). That is to say, constraining tabulations for the households were normalized by the distribution of Person 1's over the CTs, keeping the distributions of constraining tabulations in proportion.

The micro-sample used in the synthesis of households contained 6544 records, which is approximately 3.7% of the total household population. No 'additional' attributes from the sample were added to the population. When compared to the synthesis of individuals, the household synthesis required fewer constraining tabulations, generally having fewer categories. Furthermore, the size of the micro-sample, as well as the overall population size for households, was smaller than those used for the individuals. Because of this, the time required to synthesize households using the CO method was only 4 hours, and the entire population was completed in one run.

The final synthetic population of households contained the following attributes: Household ID; CT; Original sample number (h); Tenure; Structure (Census Type); Size; Income classes.

4.3.3 Dwellings & Buildings

The creation of dwellings and buildings was based on parcel data that were available for the year 2001. Parcels are small geographic divisions of space, which correspond to property lines. Our data contained a total of 141,857 parcels. Of these, 114,041 contained residential buildings, accounting for approximately 80% of Hamilton parcels. Each developed parcel is used to represent a building. Based on the attributes of the parcel data, the following 13 attributes were assigned to each building: id; CT (the CT

containing the centroid of the parcel); X-coordinate (of the centroid); Y-coordinate; area (square footage of the building); date of construction; Pluc1 (a code indicating the land-use type of the parcel); ResBld (a dummy variable taking on 1 for residential, 0 for non-residential); Property Code (a detailed code reflecting land use); number of floors; Census Type (indicating the building's structure); number of dwellings; number of rooms per dwelling. Although most of these attributes were taken directly from the original parcel attributes, or easily derived from them, several require further explanation.

ResBld indicates whether a building is considered 'residential' based on the following criteria. The building must have been built prior to 1996 (since the data contains buildings constructed as recently as 2001). The area of the building must be greater than zero. The Pluc1 code must equal 100, a classification which indicates residential land use. Finally, the building must be located in a CT which is included in the study area (note that 3 CTs were left out of the analysis, see section 4.3.1).

The number of floors and Census Type attributes were determined from the Property Code. For example, buildings with Property Code 301 (Single family detached, not on water) were assigned 2 floors, and Census Type 1 (Single-detached house). In some cases, the Property Code specified the number of floors, as in the case of Property Code 332: Residential property with two self-contained units (typically a duplex). In most cases however, a plausible number of floors was assigned, without further analysis. This lack of precision is not detrimental to our study, given that no other attributes are created using the number of floors as input, nor is the variable especially emphasized in the URM-Microsim model. The Census Type variable is essentially a simplification of the Property Code variable, into 7 categories. The Census Type variable is assigned to buildings, because it matches with the 'Structure' variable from the population of households. This becomes an asset when households are linked to dwelling units (which inherit Census Type from their buildings).

The number of dwellings and number of rooms per dwelling are assigned to buildings based on their Property Code and area. It is assumed that the average room area in low-density buildings (those with a small number of dwellings) is 130 square feet,

while that figure is 100 square feet in high-density buildings. In many cases, the number of dwelling units is indicated by Property Code, and all that remains to be done is to divide the area of the building by the number of dwelling units it contains followed by the average area of rooms in that building type, in order to determine the number of rooms per dwelling. For instance, buildings with Property Code 301 (Single family detached, not on water) contain 1 dwelling unit. Dividing the buildings area by 130 (all figures in square feet), provides the number of rooms per dwelling. In some cases, a standard number of rooms per dwelling were assigned, without considering the total area of dwellings. This was only done in cases where a reasonable estimate could be made, or when the average area of rooms was deemed difficult to determine. For instance, buildings with Property Code 350 (Row housing, with 3 to 6 units under one title) were assigned 5 dwelling units, with 8 rooms apiece. In other cases, the Property Code did not specify the number of dwelling units in any meaningful way. In these cases, the total number of rooms in the building is estimated by dividing the building's area by the appropriate average-room-area. Following this, an assumption is made on the number of rooms per dwelling. The total number of dwellings in the building is determined by dividing the total number of rooms by the number of rooms per dwelling. For instance, in the case of Property Code 340 (Multi-residence, more than six self-contained units; does not include row housing), the average room size is assumed to be 100 square feet, while the number of rooms per dwelling is assumed to be 4. Although the assumptions mentioned in this discussion are intuitively reasonable, a more careful treatment is possible, where unknown values of 'number of dwellings', 'number of rooms per dwelling' and 'average room size' are drawn from representative distributions.

Following the initial assignments of dwellings to buildings, the vacancy rate per CT is controlled for, which requires some modifications to the number of dwellings in some buildings. The reason for altering the number of dwellings, as opposed to the number of households per CT, is that the information on households is derived directly from the census, while dwelling information was estimated (as described above) and is hence less reliable. Although information on the vacancy rate of dwellings is unknown at

the CT level, at the city level, the average vacancy rate for 1996 was approximately 3% (Source: Canadian Mortgage and Housing Corporation (CMHC)). The number of dwellings per CT, in the synthesized population, was modified to reflect this vacancy rate, given the known number of households per CT. For each CT, a series of semi-random draws were made from the set of buildings. These draws had a greater probability of selecting a 'large' building (one containing more than 5 dwellings), than a 'small' building. Selected buildings would have 1 dwelling either added or removed, in order to approach a 3% CT vacancy rate; this selection and modification process terminated when the approximate desired vacancy rate was achieved. The reason for focusing these modifications primarily on large buildings was because these often had Property Codes with vague indications of the number of dwellings, making it more likely that these synthesized values were incorrect to begin with. Furthermore, altering the number of dwellings in a large building is less likely to change the character of the building, as described by the Property Code, than it is for smaller buildings.

The synthetic dwelling population consisted of a list of the dwellings in each residential building (after modification to account for vacancy rates), along with several attributes inherited from the buildings. The attributes assigned to each dwelling were: id; Building id; CT; Census Type; Property Code; number of rooms. As previously mentioned, the Census Type attribute is a simplification of Property Code, and indicates the building structure in which dwellings are located. The Census Type variable is compatible with the Structure variable from the population of households, which facilitates the linking of household and dwelling populations.

4.4 Creating the Comprehensive Population

4.4.1 Linking Individuals to Households

In order to link the synthetic populations of individuals and households, the households were conceptualized as containers, needing to be filled by an appropriate set of individuals. An ad-hoc procedure was devised to perform the linkage. The algorithm

describing this procedure was programmed in the Python language, and will be referred to during the course of our arguments. The general idea behind the linking algorithm is to assign individuals to each household in such a way that the attributes of individuals in a given household do not conflict with each other, or those of the household. The algorithm proceeds by filling households one-by-one to realize this goal. A set of rules, which will be elaborated upon, determine what constitutes a conflict between various attributes of the populations. In some cases, strict adherence to the rules greatly decreased the probability that the linking process would converge; in these cases, a probabilistic approach was adopted, where bending or outright breaking of the rules could occur, with a certain fixed probability. Nonetheless, the relaxation of particular rules was implemented in a way that would still enable the procedure to produce credible results. Figure 4.2 illustrates the basic algorithm for linking individuals to households.

Before diving into the technical aspects of linking individuals to households, a brief general description is merited. In each CT in the study area, there are a fixed number of individuals, and a fixed number of households. For each household, a number of individuals must be selected, which add to the household size. Individuals are selected without replacement, leaving fewer individuals to select from for subsequent households. To begin with, for each household, a 'household head' (also known as Person 1) is selected. In each CT, there are as many Person 1s in the population of individuals as there are households. Households are treated differently based on whether they contain one, two or three plus members. In the case of one person households, the selected individual must be a Person 1, and in the majority of cases, should not be married. For two person households, there are few restrictions on the first selected individual, other than that they be Person 1. If Person 1 is married, then in the large majority of cases, the second chosen individual should also be married, and should be the 'spouse of Person 1' (according to the 'relationship to Person 1' attribute, which all individuals possess). If the second selected individual is the spouse of Person 1, then probabilistic restrictions are placed on the age and sex of the second individual, in relation to the age and sex of

Person 1. If the second individual is a child of Person 1, then the second individual must fall into an age range determined by the age of Person 1. For 'three plus' sized

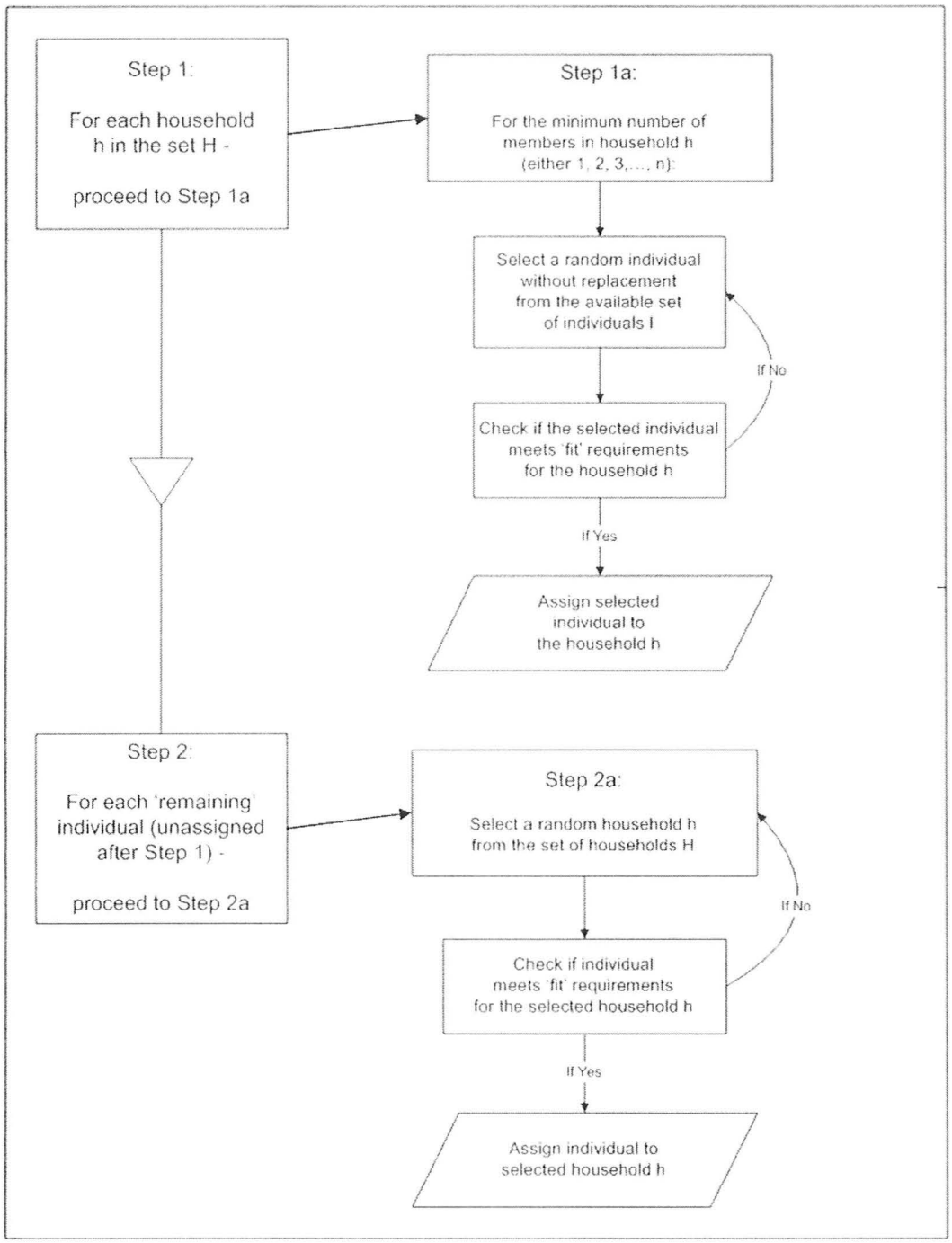

Figure 4.2: Simplified Individual-Household Linking Algorithm

households, the first two individuals are chosen in the same manner as for two person households, and then 'subsequent' individuals are chosen. Subsequent individuals can usually not be the spouse of Person 1, however there is a small probability of this occurring. If subsequent individuals are the child of Person 1, they must meet certain age criteria, based on the age of Person 1. For all households, once a set of members has been chosen, their collective income is checked against the household income. Sets of individuals whose collective income does not fall within an acceptable range of the household income are dissolved, and the process of choosing a set of individuals for the household is started anew.

The attributes of synthetic individuals are: Individual ID; CT; Original sample number (i); Citizenship; Sex by Employment; 5-year Mobility Status; Sex by Income; Marital Status; Sex by Age by Relationship to Person 1; Highest Level of Schooling; Industry; Occupation. See section 4.3.1 for more details on these attributes.

The attributes of synthetic households are: Household ID; CT; Original sample number (h); Tenure; Structure (Census Type); Size; Income range. See section 4.3.2 for more details on the household attributes. For convenience, Individual and Household attributes are listed in Tables 4.6a and 4.6b respectively.

Table 4.6a: Attributes of the Population of Individuals

Individual ID
Census Tract (CT)
Original sample number (from the micro-sample of individuals)
Citizenship (Canadian or non-Canadian)
Sex by Employment Status
5-year Mobility Status
Sex by Income range
Marital Status
Sex by Age by Relationship to Person 1
Highest Level of Schooling obtained
Standard Industrial Classification (1980)
Occupation (in the Labour Force)

Table 4.6b: Attributes of the Population of Households

Household ID
Census Tract (CT)
Original sample number (from the micro-sample of households)
Tenure (does the household rent or own their dwelling)
Structure of the households dwelling (Census Type)
Size (the number of household members)
Income range

The greatest difficulty in linking individuals to households is that, if a household requires a set of individuals with "rare" attributes, relative to the pool of available individuals, there is no guarantee that such individuals will exist. This can occur when an individual with universally rare attributes is required, at any stage of the linking process. Alternatively, this situation can arise when most individuals have already been assigned to households, leaving those remaining with a limited variety of attribute sets. For example, there are only a small percentage of wealthy, 1-person households per CT. Similarly, there are only a relatively small percentage of wealthy individuals per CT. When it comes time to assign an individual to such a household, an appropriate one may not remain in the pool of un-assigned individuals. A number of strategies were adopted during the linking process to address this difficulty. First of all, difficult to assign households were treated early on in the linking process, in order to provide the largest possible pool of available individuals for selection. In particular, households were treated in increasing order of size, and decreasing order of wealth, as a result of the observation that wealthy households having fewer members were the most difficult to fill. Secondly, as previously mentioned, most rules dictating possible sets of individuals for a given household were softened. That is to say that these rules were only strictly enforced a certain proportion of the time, as determined by the situation. Thirdly, households were initially filled with the minimum possible amount of individuals (see Figure 4.2). The possible household sizes were: 1; 2; 3; 4-5; 6+. In the case of households having 3 or fewer members, complete filling took place. In the case of households having 4-5 or 6+ members, 4 and 6 individuals were initially assigned, respectively. The idea behind this third strategy was to leave as large a pool of available individuals as possible, until each

household was filled with the minimum amount of individuals. Subsequently, any remaining individuals would have a large choice of 4-5 and 6+ households to join, increasing the likelihood that an appropriate household could be found. If any of the remaining individuals could not be assigned to a household, then they would be removed from the synthetic population of individuals.

Appendix I contains the main body of code used during the linking process. Note that lines in the code containing comments have '#' as their first character. The main thrust of the code up until line 255 is to create a useable list of households, referred to as the household use-list. The household use-list differs slightly from the original list of synthetic households in several ways. First, the household use-list is sorted by CT, and sorted in decreasing order of total-income and increasing order of size within each CT. Second, the household use-list ensures that there will not be more households in a given CT than Person 1s in the population of individuals. This is important, because during the linking procedure, one-and-only-one individual with the Person 1 attribute must be assigned to each household. If a deficit of Person 1s exists for a given CT, assigning the last few households a set of appropriate individuals becomes impossible, as all available Person 1s will have previously been assigned. Although the number of households synthesized per CT was constrained by the number of Person 1s per CT (see section 4.3.2), the CO method cannot guarantee that synthesized distributions will match the constraining tabulations exactly. The household use-list omits 'extra' households from the original synthetic population of households, with no bias as to which types of household are eliminated. If there happen to be extra Person 1s in a CT, then these will eventually be eliminated from the population of individuals, when they fail to be assigned to any households.

Beginning at line 255 of the main body of code (Appendix I), the core linking algorithm is implemented. Households in the household use-list are treated sequentially. First, information on the attributes of the selected household is gathered. Following this, at line 278, one of three possible functions is called to select an appropriate set of individuals for the household, based on household size divisions: 1; 2; 3 or more. Here,

the set of chosen individuals will be of minimum size (4 or 6 for 4-5 and 6+ households, respectively).

Appendix II contains the code for the function that picks an individual to belong to size 1 households. Random individuals are chosen (line 6), until one is found which meets four conditions:

- The individual must be from the same CT as the household (lines 9-11).
- The individual must not live with a spouse, as indicated by their Marital Status attribute (lines 13-21).
- The individual must be a Person 1, as indicated by their Relation to Person 1 attribute (lines 23-25).
- The individual's income should fall within the household income range (lines 28-46).

In the last case, to ensure convergence of the function, some leeway was given; individuals with an income 'close' to the household income range were also accepted. Once an individual has been selected by this function, it is removed from the available population of individuals, to prevent it from being selected for another household (Appendix I, lines 321-325).

Appendix III contains code for the function that picks two individuals, for households of that size. The first individual is chosen in a similar manner as above (lines 5-22); however in this case, individuals need only meet two criteria. First, they must be located in the same CT as the household. Second, the individual must be a Person 1. Following successful selection of a first individual, the second individual is treated (lines 25-117). As a first step, information on the age, sex and marital status of the first selected individual is gathered. Then, individuals from the available pool are randomly chosen, until one is found which meets six conditions:

- The individual's CT must match that of the household.
- The individual can't be a Person 1. This condition has the advantage of ensuring that the first individual will not be chosen again.

- Depending on the marital status of the first individual, the second individual should be the spouse of Person 1, as indicated by their 'Relation to Person 1' attribute (lines 55-72).
- Should the second individual be the spouse of Person 1, then they should, in the majority of cases, be the opposite sex of their spouse (lines 75-85).
- If the second individual is a spouse of Person 1, then their age should be close to that of the first individual (lines 88-98).
- The sixth and final criterion imposes an age restriction on the second individual, in the case that they are the son or daughter of Person 1 (lines101-111).

For spouses, age restrictions ensured that they be no more than 15 years apart. For children of Person 1, an age at least 15 years younger than Person 1, but no more than 45 years younger was imposed. Once two individuals have been selected by this function and returned to the main linking code, they are checked, 80% of the time, for adherence to the household income range (Appendix I, line 292). The function inc-2plus takes the individual's income ranges, and ensures that the minimum possible collective income of the individuals is smaller than the high end of the household income range, and vice versa. If the individuals fail to conform to the household income range when required, the individuals will be thrown back into the pool of individuals, and the function run anew. If however, the individuals pass the income test, then they are assigned the household ID, and deleted from the pool of available individuals.

The function for selecting synthetic individuals for households with 3 or more members proceeds exactly as above for selection of the first two individuals. For subsequent individuals, the function calls another function solely responsible for choosing one additional individual for a household with 2 or more existing individuals. This function will be referred to as addit-mems, and can be found in Appendix IV. As can be seen on line 10, the addit-mems function ensures that additional household members meet five criteria:

- The individual must reside in the same CT as the household being filled.
- The individual cannot be Person 1.
- The individual cannot be Person 2 (the spouse of Person 1). This rule exists because the second person chosen for a household should be the spouse of Person 1, by design, should Person 1 have a spouse.
- If the individual is a son or daughter of Person 1, then age restrictions apply.
- Finally, the individual may not already be a member of the household, either through previous selection of the second individual, or subsequent individuals already selected by the addit-mems function.

The addit-mems function is called until all remaining individuals have been selected for the household in question. However, in the case of 4-5 and 6+ member households, addit-mems is only called until either 4 or 6 individuals are selected, respectively. This strategy is adopted so that a large pool of 'extra' individuals will remain, and have a large number of potentially suitable 4-5 and 6+ households for assignment.

Once each household has been assigned its minimum number of members, the remaining individuals are examined and treated in sequence (Appendix I, lines 329-385). Remaining Person 1s and Person 2s (spouses of Person 1) are eliminated from the synthetic population at this point. Person 1s are eliminated because exactly one is required for each household, all of which have been filled. It is assumed that any households requiring a Person 2 would have already selected one, in most cases, and hence Person 2s are eliminated. Individuals that are not eliminated are treated with the function 'end-game', which selects an appropriate household for the individual. The code for the end-game function can be found in Appendix V. The function selects a random household for a given individual, contingent on the household meeting several criteria. First, the household must contain either 4 or 6+ individuals. This prevents individuals from being assigned to full households. In the case of 4-5 member households, a maximum of 1 individual can be added, while for 6+ households, no limit

is imposed. Second, the household must be located in the same CT as the individual. The relatively lax set of rules imposed by the end-game function ensures convergence, and although this was our intention, several additional rules can be envisioned. First, children of Person 1 (Person 3s) could be assigned only to households where Person 1s age is in accord. Second, Person 2s could be assigned to households where no previous Person 2 has been assigned, and where the spouses conform on age, sex and marital status criteria. The end-game function is responsible for the assignment of approximately 3.5% of individuals, and so the overall error introduced by the function is minimal.

The linking of individuals to households is a fascinating task, due to its novelty as well as the variety of possible solutions it engenders. Although a specific approach to linking was adopted here, variations of this approach, as well as altogether different approaches can certainly be envisioned. This could prove to be an interesting area for future research.

4.4.2 Linking Households to Dwellings

The process of linking households to dwellings is a simplified version of the individual-household linking procedure (see section 4.4.1 above). In this case, dwelling units are envisioned as containers, which may be filled by at most one household. Each household is assigned to a dwelling, based on the household and dwelling unit attributes. There are three main criteria which determine whether a given dwelling unit is suitable for a given household. First, the dwelling unit must be located in the same CT as the household. Second, the structure in which the household lives must match the structure of the dwelling unit. The structure attribute is referred to as 'Census Type' (see sections 4.3.2 and 4.3.3 for more details). Finally, in 90% of the cases, the number of household members should not exceed the number of rooms per dwelling unit. This imposes a reasonable restraint on the size of households residing in small living quarters.

Once all households have been assigned a dwelling unit, the vacancy rate in each CT, as well as the entire study area, will be 3%. Furthermore, the household-dwelling linkage finalizes the comprehensive population, such that each element of the population

(or elemental population) is linked in a meaningful, hierarchical way. Dwellings are linked to Buildings by virtue of their creation (see section 4.3.3), while Individuals were linked to Households in section 4.4.1; and this section completes the comprehensive population by linking Households to Dwellings.

4.5 Validation of the Synthesized Populations

Given the comprehensive population for the City of Hamilton, Ontario described in previous sections of this chapter, a question naturally arises: "how close is this synthetic population to the 'actual' population?" Since we do not have complete information on the actual Hamiltonian population, this question cannot be fully answered. Nonetheless, several measures comparing the synthetic comprehensive population to the actual population can be made, in order to validate the quality of the former. To this end, measures of each elemental population (those comprising the comprehensive population) are made, as well as measures of the accuracy of linkages between elemental populations. Note that the most important aspects and attributes of the comprehensive synthetic population are those which are required by the URM-Microsim model; and these will be focused on during validation.

4.5.1 Validation of Individuals and Households

For each variable that was constrained for during the synthesis of individuals, a measure of fit to the actual population distribution of the variable is taken. In particular, the Relative Sum of Squared Z-scores (RSSZ) statistic is used. A detailed description of the RSSZ statistic can be found in Ch. 2, as well as Huang & Williamson (2002). When the synthetic distribution of a variable exactly fits the constraining tabulations, the value of RSSZ equals zero, and if two different synthetic distributions are compared to the constraining tabulations, the better of the two will have a value of RSSZ closer to zero. In general, a value of RSSZ less than one implies an excellent fit between synthesized and actual distributions of a variable. During the synthesizing process, RSSZ measures are taken for each of the 6 constraining variables, for each of the 124 CTs in the study

area. Table 4.7 contains summary statistics on the set of RSSZ values, across all CTs, for each constrained variable in the synthesized population of individuals. These give an indication of the fit between synthetic and actual distributions of each variable, across the entire study area. Of particular note, the average RSSZ values for each variable are less than one. The two highest averages are 0.0003 and 0.0605, for Sex by Income and Sex by Age by Relation to Person 1, respectively. This is an understandable result, since these variables are more complex than their counterparts, having more categories in their domains (see section 4.3.1 for variable details). Accordingly, the two highest maximum RSSZ values are held by the same two variables. Only 'Sex by Age by Relation to Person 1' has a maximum RSSZ value above one, namely 1.063.

Table 4.7: Summary Statistics of RSSZ values; constrained variables; individuals

	Citizen-ship	Marital status	5 Year Mobility	Sex by Employment	Sex by Income	Sex by Age by Relation to Person 1
Avg	0	0.000218	1.15E-05	4.63E-05	0.000293	0.060516
Stdev	0	0.000520	8.41E-05	0.000147	0.000500	0.153144
Max	0	0.003298	0.000835	0.001076	0.003446	1.063040
Min	0	0	0	0	0	6.73E-04

An important measure of how well the synthesized population of individuals fits each CT is the sum of RSSZ values across all constraining variables, for each CT. During the synthesis procedure, which operates independently on each CT, it is this sum of RSSZ values which is minimized. Table 4.8 contains summary statistics on these summed RSSZ values. Of particular note, the average and standard deviation are 0.061 and 0.153 respectively. These results imply an excellent fit between the synthetic and actual distributions of the constrained variables, for each CT. The RSSZ values are displayed graphically over the study area in Figure 4.3. Darker CT's are those with higher values, indicating poorer fit between the actual and synthetic distributions of constrained variables. Some descriptive observations can be made; the main concentration of darker CTs is located to the west of the Central Business District (CBD),

with the highest RSSZ CT belonging to the municipality of Ancaster. With the exception of one periphery CT to the north-west of the CBD, and a small cluster to its east, the remaining CTs are in the extreme low end of the RSSZ spectrum, implying excellent fit.

Table 4.8: Summary Statistics of summed RSSZ values, by CT

Statistic	Individuals	Households
Avg	0.061085	0.001705
Stdev	0.153118	0.018988
Max	1.063220	0.211441
Min	0.000673	0
Count	124	124

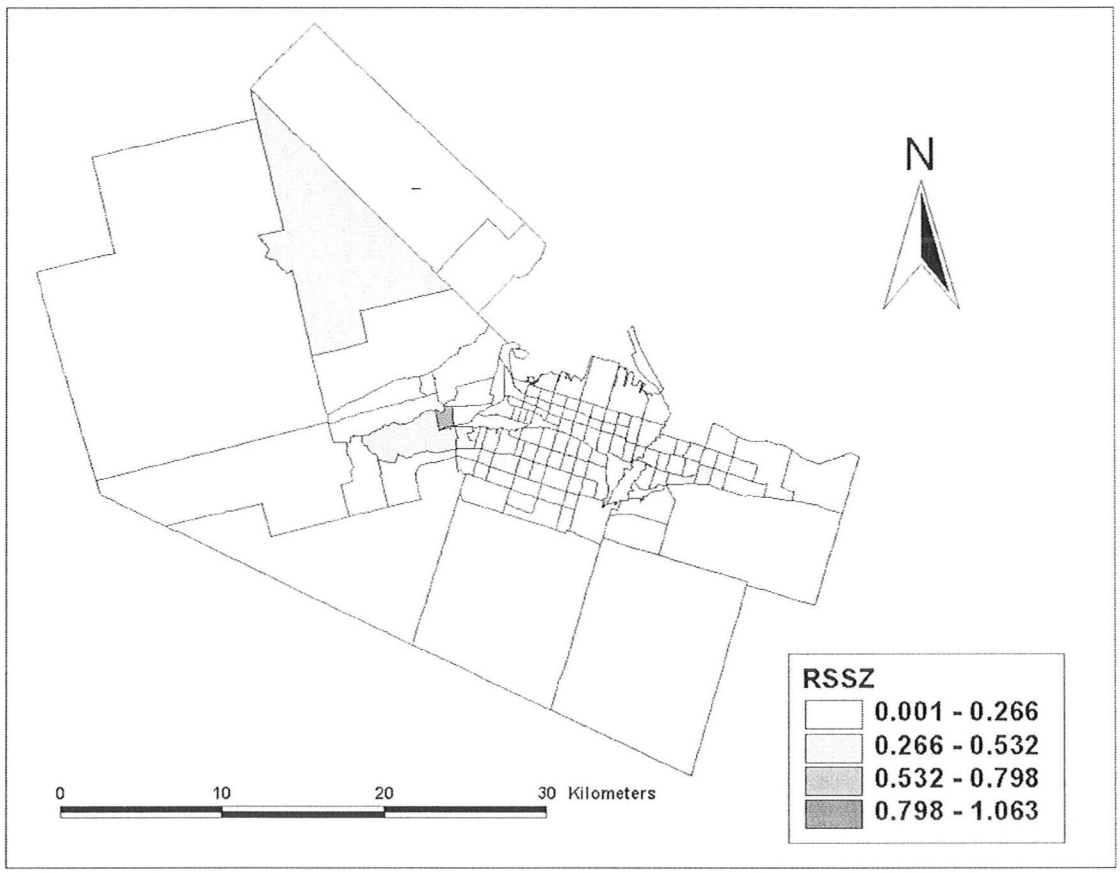

Figure 4.3: RSSZ sums, over Hamilton Census Tracts; individuals

Since the purpose of the synthesis process was to replicate the actual distribution of constrained variables in the synthetic population, the results above are expected. Less certain is the fit between the synthetic and actual distributions of 'added' variables (namely: Highest Level of Schooling; Industry; and Occupation). These variables were not constrained for during the synthesizing process, but were added afterwards from the original sample of individuals, due to their importance in the URM-Microsim model (see section 4.3.1). Summary statistics on the set of correlations between actual and synthetic distributions of these variables, over space, can be found in Table 4.9. For the variable 'Highest Level of Schooling', the average correlation is 0.908, which implies a strong fit over space. For 'Industry' and 'Occupation', the average correlation values are 0.883 and 0.829 respectively, still implying a strong fit in both cases. These results suggest that 'Highest Level of Schooling' is more closely related to the set of constrained variables than 'Industry' or 'Occupation'.

Table 4.9: Summary Statistics of Additional variable correlations over space

	Level of Schooling	**Industry**	**Occupation**
Avg	0.908068	0.882860	0.828720
Stdev	0.086354	0.091028	0.068445
Max	0.991716	0.987706	0.957613
Min	0.571307	0.519171	0.588729
Count	124	124	124

For each of the added variables, the minimum correlation value is above 0.5, further demonstrating an acceptable fit between actual and synthetic distributions of these variables over space. Figures 4.4a, 4.4b and 4.4c show the correlations of the variables 'Highest Level of Schooling', 'Industry' and 'Occupation', respectively, over the CTs. For 'Highest Level of Schooling', areas with relatively low correlations are found primarily in the areas bordering the CBD to the east and west. The CBD as well as the peripheral CTs show high correlations. For 'Industry' (Figure 4.4b), we see the same areas of low correlations as in Figure 4.4a, however more of the CBD itself contains low correlation tracts, as well as the majority of the periphery. For 'Occupation' (Figure

4.4c), high and low correlation tracts are spread heterogeneously throughout the study area.

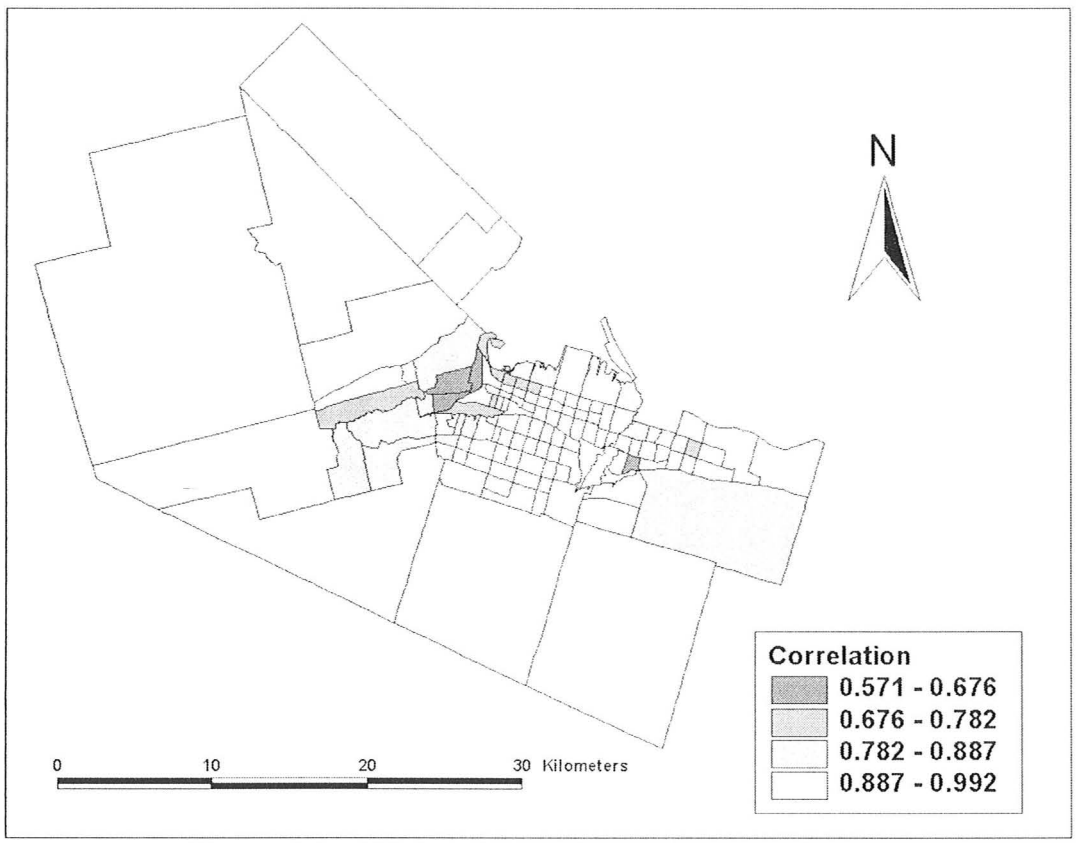

Figure 4.4a: 'Highest Level of Schooling' correlations

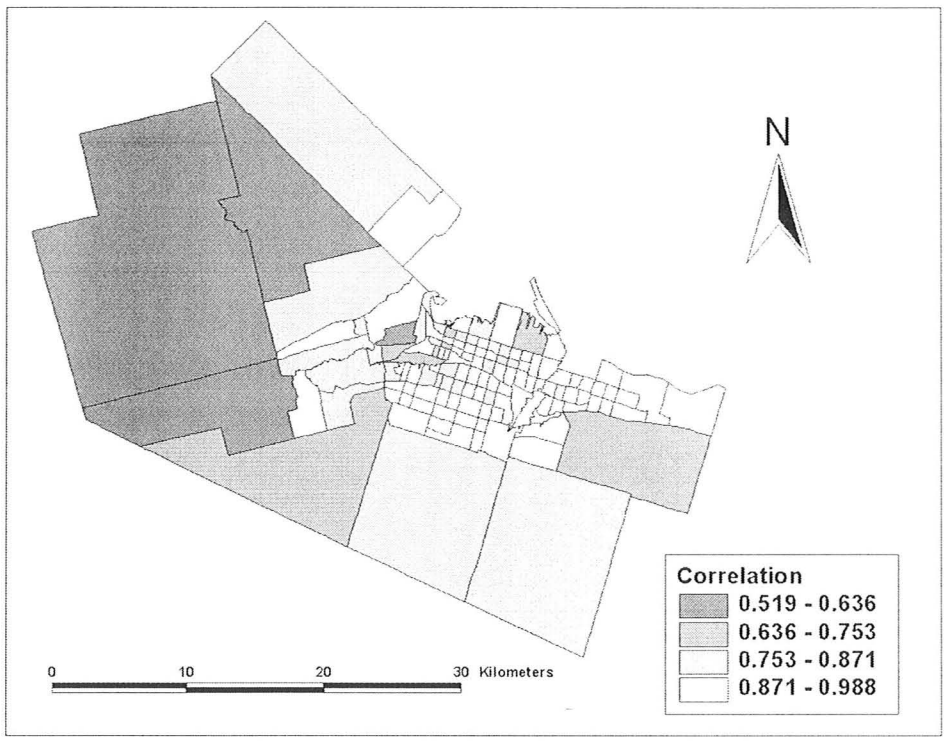

Figure 4.4b: 'Industry' correlations

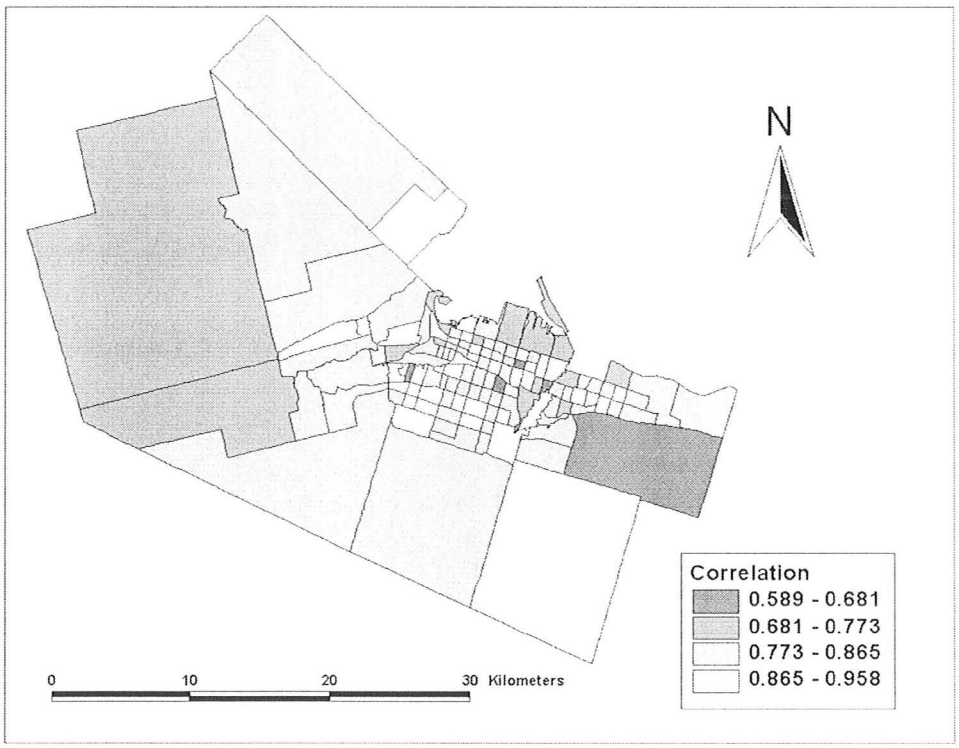

Figure 4.4c: 'Occupation' correlations

Once the synthesis process is complete, it is simple to include "additional" variables, such as those discussed above, in the population. In order to gain another perspective on the validity of the population, we add the variable 'Mode' to the population of individuals, for the purpose of examination. The variable 'Mode' categorizes the type of transportation that individuals take to work. 'Mode' is not required by the URM-Microsim model, but is generally of interest in transport research. Summary statistics on the 'Mode' correlations over space can be found in Table 4.10. As can be seen from the average of 0.989 and minimum value of 0.785, the fit of the synthetic 'Mode' distribution against the actual distribution is highly significant, over the study area. This is most likely the result of 'Mode' being highly correlated with one or more of the constraining variables, and also perhaps the result of 'Mode' being relatively uniformly distributed over the study area. Correlation values for 'Mode' can be seen graphically in Figure 4.5. Interestingly, the only low correlation CTs are clustered to the north-west of the city centre. This may be due to heterogeneity in the population in that location, with respect to their mode choices.

Table 4.10: Summary Statistics of 'Mode' and 'Language' correlations over space

Statistic	Mode	Language
Avg	0.988950	0.999200
Stdev	0.026471	0.002900
Max	0.999803	1
Min	0.784837	0.974500
Count	124	124

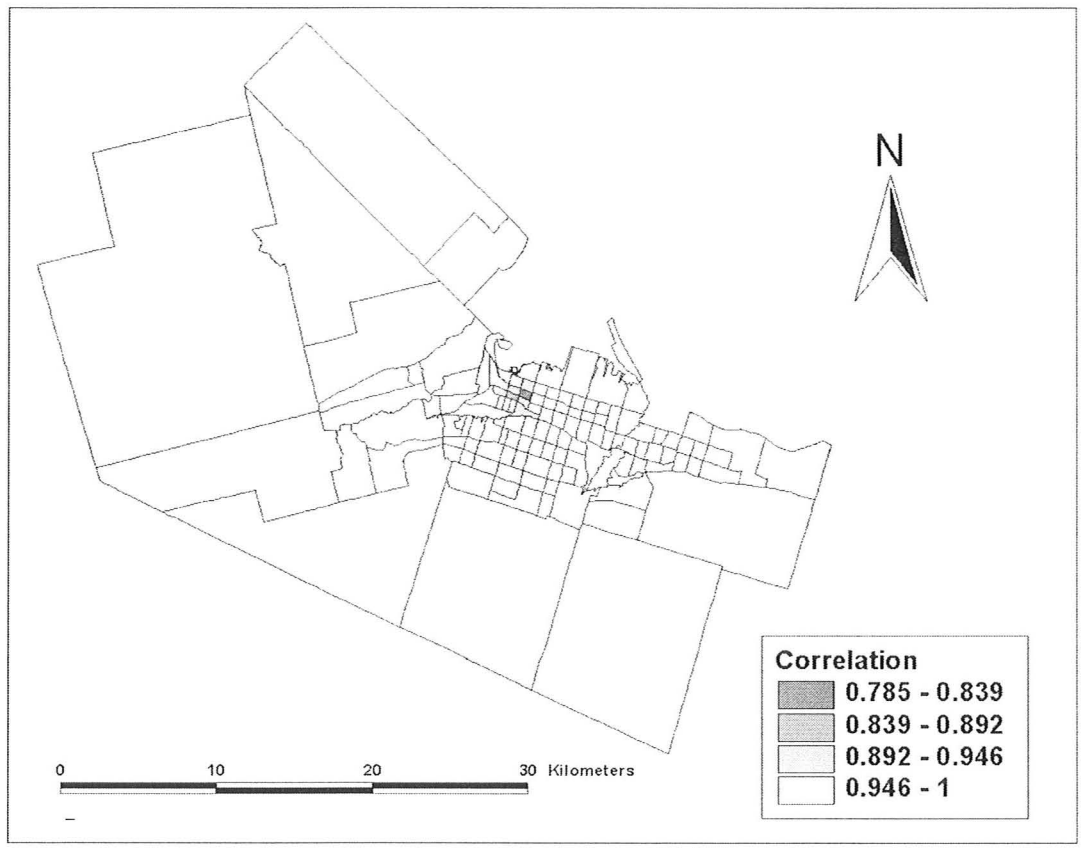

Figure 4.5: 'Mode' correlations

Another variable which is not required by the URM-Microsim model, but whose analysis nevertheless provides an interesting perspective on the synthesized population of individuals, is 'Language'. This variable categorizes individuals into those who speak English, those who speak French, those who speak both, and those who speak neither. Although 'Citizenship' was constrained for during the synthesis of individuals, it would appear that the 'Language' variable is not highly correlated to any of the constraining variables, and hence that its correlations over space with the actual population distributions would not be high. Summary statistics on the 'Language' correlations over space can be found in Table 4.10. Interestingly, the average correlation value is 0.999, while the minimum value is 0.975. The high correlations observed in the cases of the variables 'Mode' and 'Language' suggest two things. First, the six variables used to constrain the synthesis of individuals were well chosen, in that they are highly correlated

with other variables of potential interest to geographers. Second, variables such as 'Mode' and 'Language' have distributions which are relatively uniform over the study area, meaning that they may be well replicated even by synthesized populations whose constraining variables are poorly correlated with them.

During the linking of individuals to households (see Section 4.4.1), approximately 3.3% of the original individuals were omitted from the population, leaving a final population of 442,124 individuals. The eliminated individuals primarily consisted of Person 2s (spouses of Person 1). In total, approximately 10% of the original Person 2s were eliminated. This constitutes the largest known error introduced into the population of individuals by the linking process.

During the synthesis of households, the following variables were constrained: Tenure; Structure (Census Type); Size; Income range. Notably, every variable required by the URM-Microsim model was constrained for. Therefore, the most important measure of validity for synthetic households is the summed RSSZ, over all constraining variables, for each CT. Table 4.8 contains summary statistics of these values. The average RSSZ value is 0.002, while the maximum value is 0.211. This indicates that the distributions of constrained and actual variables fit extremely well, for all CTs in the study area. In fact, the summed RSSZ is zero in all cases except for one, where it is 0.211.

A small number of households were eliminated from the final comprehensive synthetic population during the linking of individuals to households (see Section 4.4.1). This elimination was carried out to ensure that there would be an equal number of households and Person 1s in each CT. In particular, the number of households was decreased from 177,005 to 175,020; approximately a 1% reduction.

4.5.2 Validation of Dwellings and Buildings

Validation of Dwellings and Buildings is difficult, due to lack of data. In fact, if additional data on these populations were available, it would have been incorporated into their creation. Several points can be made however. First, the number of dwellings is

based on the number of synthetic households, resulting in a 3% vacancy rate in every CT in the study area. The number of dwellings, along with most dwelling characteristics, can be easily changed if additional data becomes available. In this case, the population of individuals and households need not be altered, except for the re-linking of households to the updated set of dwellings. Second, most building characteristics were not synthesized per-se, but derived directly from city parcel data (see Section 4.3.3), precluding validation at this stage.

4.5.3 Validation of Individual-Household Linkages

When children of Person 1 were added to households, during the linking process, there was a stipulation that Person 1 be at least 15 years older than the child, and no more than 45 years older. As with the majority of linking rules, this was enforced in the majority of instances. If a 'Spouse of Person 1' already existed in the household, no additional constraint was put on children's ages. However, the age of a 'Spouse of Person 1' was constrained by the age of Person 1, indirectly constraining children's ages by the age of the 'Spouse of Person 1'. In order to ensure that children's ages are reasonably distributed with regards to Person 1's age, and the age of Person 1's spouse, the maximum, minimum and average differences between parents and children's ages were calculated, for each applicable household. In the case of single parent households, where no 'Spouse of Person 1' is present, the averages of maximum, minimum and average difference between Person 1's age and the ages of children in the household are presented in Table 4.11. Note that there are 41,541 single parent households, accounting for almost 24% of total households. The average mean difference is, 30.7 years, which is reasonable. When parent-child age differences are disaggregated by sex, the female values are slightly higher than male values, for all measures. For example, the female average mean difference is 31.1 years, while that figure is 30.5 years for males. Although comparative data of this type from the actual Hamilton population is not publicly available, some comparison can be made with Statistics Canada data detailing the mean age of mothers directly after giving birth, in a given year. In particular,

Statistics Canada reports that the mean age of mothers in Ontario, for the year 2004, was 29.9 years. Also of interest, single parent families headed by a female account for 38.5% of single parent households. In the actual population of Hamilton, this figure is approximately 85%, according to the 1996 Canadian census. Also, the number of single-parent families is overestimated in the synthetic population; specifically, there are 41,541 single-parent families in the synthetic population, and 19,620 of these in the actual population.

Table 4.11: Parent-Children age differences in single-parent households

	Avg. Maximum	Avg. Minimum	Avg. Mean	Total Number
All	34.2	28.6	30.7	41,541
Male	34.0	28.4	30.5	25,532
Female	34.5	28.9	31.1	16,009

For two parent households, the averages of maximum, minimum and average difference between the parents' ages and the ages of children in the household are presented in Table 4.12. The average mean difference is 32.5 years, slightly higher than that figure for single parent households. For female parents in two-parent households, the average difference is 31.5 years, while this figure is 33.5 for male parents. In contrast to the single-parent figures, for two-parent households, female parents have slightly lower values of average minimum, maximum and mean age difference than their male counterparts. Note that there are 44,643 two-parent households in the synthetic population, accounting for approximately 26% of households in the study area. According to 1996 Canadian census data, there were 63,110 two-parent households (with children) in the actual Hamilton population. The overestimation of single-parent families, along with the underestimation of two-parent families, can be attributed to the exclusion of a number of 'Spouses of Person 1' from the population of individuals during the linking of individuals to households (see Section 4.4.1). Special attention will be paid to this matter in any future comprehensive population synthesis efforts.

Table 4.12: Parent-Children age differences in two-parent households

	Avg. Maximum	Avg. Minimum	Avg. Mean	Total Number
All	35.8	30.4	32.5	44,643
Male	36.9	31.5	33.5	42,913
Female	34.9	29.5	31.5	46,373

A number of 1996 Canadian Census Tabulations exist which specify the make up of households (the relationships between a household's members). Since these tabulations were not used in the creation of individuals, nor the linking of individuals to households, they provide an excellent means of validating these linkages. One such census tabulation provides the distribution of family types by the number of children present in the household. In this case, the family types are: married-couple family; common-law couple family; and single-parent family. The divisions of children for each family are: zero; one; two; three or more. Summary statistics on the correlations of this variable's distribution between the actual and synthesized populations, over space, can be found in Table 4.13. Of note, the average, maximum and minimum correlations are 0.852, 0.983 and 0.613, respectively. These values indicate that the distribution of 'family structures by number of children' over space have been well replicated in the comprehensive synthetic population. The correlations can be seen graphically in Figure 4.6. Note that CTs with the lowest correlations are located bordering the CBD, to the North and South-West.

Table 4.13: Summary Statistics of 'Family type by children' correlations over space

Avg	0.852488
Stdev	0.07265
Max	0.982648
Min	0.612859
Count	124

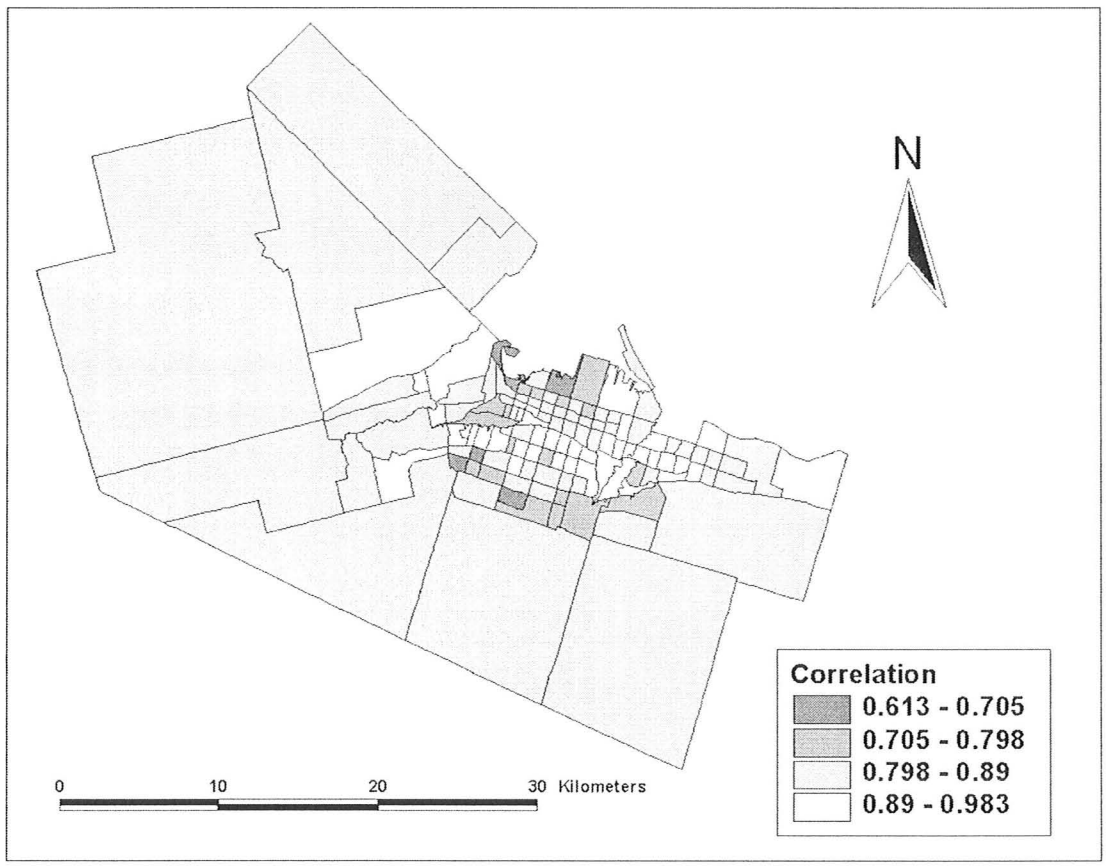

Figure 4.6: 'Family type by number of children' correlations

Another census tabulation of interest depicts the distribution of 'census families' by the number of employed persons per family. Here, 'census families' are defined as couple families (married or common-law) with or without children, and single-parent families having at least one child. In the synthetic population, 382,083 individuals (approximately 86%) belong to 'census families'. The tabulation 'Census family type by employment' provides the distribution of census families among the following categories: Couple family with no member in the labour force; Couple family with some members in the labour force; Couple family with both spouses in the labour force; Lone parent family with no member in the labour force; Lone parent family with some members in the labour force; Lone parent family with the parent in the labour force. Note that these categories are not mutually exclusive, since a family with all of the 'parents' in the labour force is also considered to have 'some members in the labour force'. Summary statistics on the

correlations of this variable's distribution between the actual and synthesized populations, over space, can be found in Table 4.14. The average, maximum and minimum correlations are 0.805, 0.941 and 0.419, respectively. These values indicate that the distribution of 'Census family type by employment' over space have been acceptably replicated in the comprehensive synthetic population. The correlations are displayed graphically in Figure 4.7. As can be seen, relatively few of the census tracts have low correlation values, and these are concentrated to the North of the CBD.

Table 4.14: Summary Stats, 'Census families by employment' correlations over space

Avg	0.805023
Stdev	0.107251
Max	0.941221
Min	0.418642
Count	124

During the assignment of individuals to households, the collective incomes of household members were constrained by the household income range (see Section 4.4.1). For convergence purposes, this constraint was partially or completely relaxed at times. In order to assess the error introduced by relaxing this constraint, we look at the accumulated error of all households, per CT. Incomes for both individuals and households are presented in the synthesized population as ranges of values. For a given household, the income set of its members does not 'fit' if: the sum of the minimum individual incomes is greater than the maximum household income; the sum of the maximum individual incomes is less the minimum household income. In the case where the members of a household have an income set which does not fit the household income range, we define the error as the minimum sum which would cause the incomes to fit.

Figure 4.7: 'Census family type by employment' correlations

Summary statistics on the income errors for all households in the population can be found in Table 4.15. Of particular note, the average and standard deviation of the errors are $11,289 and $18,972 respectively. The relatively large standard deviation is due to the presence of outliers in the error set. These outliers also account for the range of error values being relatively large, namely $230,001. When the average household income errors are calculated for each CT, we find that the maximum and minimum values are $15,933 and $5,714 respectively. These average error values are displayed graphically over space in Figure 4.8. Note that higher values appear to be clustered together to the south and east of the CBD. All of the CTs with low average error values are located in the central portion of the City, containing the CBD. When households are treated by a model, such as URM-Microsim, the sum of individuals' incomes should be used as the total household income, for consistency.

Table 4.15: Summary Statistics, Household Income Errors, Entire population ($ Cdn)

Avg	11289.61
Stdev	18972.08
Max	230001
Min	0
Count	175020

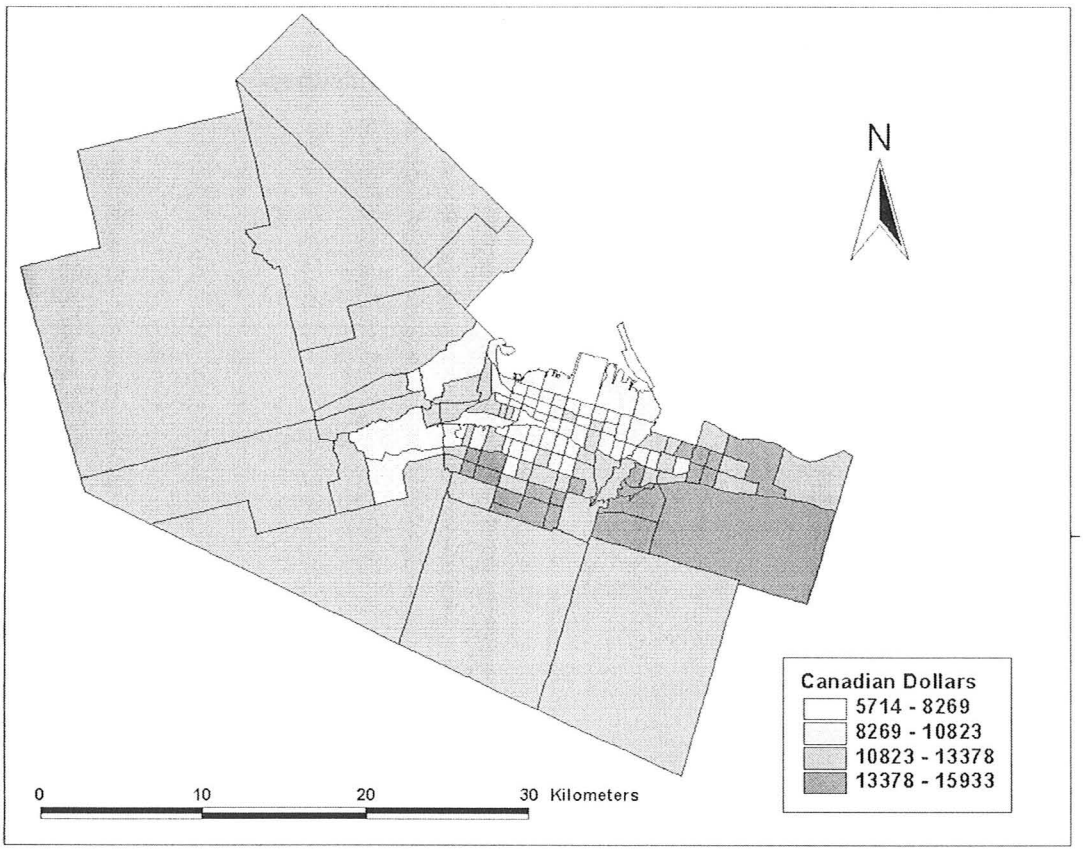

Figure 4.8: Average income assignment errors by CT

4.6 Conclusions

We have successfully synthesized a comprehensive population for the City of Hamilton, Ontario, which meets the criteria for input into the URM-Microsim model. In doing so, we have adapted URM-Microsim to the City of Hamilton, and hence the Canadian context. It has yet to be seen how URM-Microsim will perform under these new conditions; an issue which is left for future work.

The techniques used to create the Hamilton comprehensive population are general enough that they may be easily applied to any city in Canada; and perhaps to cities in the United States and Europe as well. This is encouraging, since it allows for wide ranging application of URM-Microsim, as well as other similar models. Once some experience has been gained implementing URM-Microsim, we will be able to look at the comprehensive population synthesis process with hindsight, and possibly improve upon it. As it stands, some improvements to the linking process between individuals and households can be envisioned, in particular, a more complete assignment of Person 2s (spouses of Person 1) to households is desirable.

From a run-time perspective, the most time consuming element of the entire synthesis effort is the synthesis of individuals using the Combinatorial Optimization (CO) method. As mentioned in Chapter 2, cutting edge techniques to reduce these run times are the subject of ongoing research, and will no doubt be incorporated into future synthesis efforts.

Validation of the Hamilton comprehensive population proved to be encouraging on the whole. In particular, variables which were added to the population of individuals after implementation of the CO method were highly correlated with their actual population distributions. Validation carried out on the nature of synthetic family structures yielded, for the most part, positive results. In particular, several family characteristics which were not constrained for during the linking of individuals to households were reasonably replicated in synthetic households. Issues that require improvement include: constraining the sex of parents in single-parent households, as well as the overall number of such households; ensuring a complete assignment of 'Spouses of Person 1' to households, while eliminating any multiple assignments. Where buildings and dwellings are concerned, a more detailed set of input data would improve the quality of the synthetic population. However, the input data used here satisfies our needs for this project.

Chapter Five: Conclusions

Several important conclusions can be reached as a result of this thesis. To begin with, in Chapter Three, two different methods for synthesizing populations were compared in their abilities to replicate a known, small population of firms. The two methods in question were the Synthetic Reconstruction technique (IPFSR) and the Combinatorial Optimization method (CO). In both cases, the required input data consisted of a representative sample of population members, without spatial identifiers, as well as tabular distributions of desired population characteristics, for each zone in the study area. This basic form of input data is commonly available in practice (for instance from publicly available Canadian Census sources), which speaks to the practicality of both methods. From the results of experiments carried out in Chapter Three, we learned that both the CO and IPFSR methods produced synthetic microdata which conformed to the target population, according to statistical tests. That being said, the quality of microdata produced using the CO method proved to be higher than that of IPFSR produced microdata. This result may be explained in part by the stochastic nature of the Monte Carlo draws used to select attributes for synthetic individuals in IPFSR. Experiments were also carried out to test techniques of adding additional attributes to synthesized population members, after the fact. For the CO method, this could be accomplished by linking population members to the original sample members, such that population members inherited attributes from the sample which were not constrained for during synthesis. For both the CO and IPFSR methods, additional attributes could also be added through Monte Carlo draws from a conditional probability distribution. Both techniques for adding additional attributes were successful in their abilities to recreate distributions of the additional attributes over space. In the case of linking population members with the original sample, better results are achieved when the attributes being added are highly correlated with attributes that were constrained for during the synthesis process.

Using the knowledge gained from Chapter Three, a synthetic population of individuals and households was created for the City of Hamilton, ON, as described in Chapter Four. The CO method was used to synthesize each population, which constituted a test of the CO method's abilities, given the larger population sizes, and greater number of constrained attributes, than was the case in the Chapter Three synthesis effort. One observation made from the synthesis of individuals was that a large amount of computer run-time was required for the process to converge, given the large size of the population being synthesized. This led to an interest in creating more efficient methods of population synthesis which can handle large populations in less time; and this will certainly be the subject of future research. Once the populations of individuals and households were created, a series of linking rules were developed to join members of the individual population with members of the household population, forming 'households' of synthesized individuals. Here, convergence was found to be difficult to obtain, and a series of techniques, including the use of 'soft' rules (rules which were honored in only the majority of cases) were employed. In general, a variety of similar linking methodologies could be used in this situation, and the exploration of such will be the subject of future work. Following the linking of individuals to households, a synthetic set of dwellings was created, based on parcel and building data. Another set of linking rules was employed to join the population of households to dwellings, which effectively created a hierarchical set of populations, where individuals belonged to households; and households belonged to dwellings, which in turn belonged to buildings which were fixed spatially, at the parcel level of geography. We refer to this sort of hierarchically linked population as a 'comprehensive' population. The motivation for creating this particular comprehensive population was to obtain input data for a microsimulation model of residential mobility. In particular, the population was created as input to the URM-Micosim model, and population attributes, as well its general form, were chosen to meet the demands of this model.

The methods used to create the Hamilton, ON comprehensive synthetic population can be replicated to create input data for other microsimulation models, or

indeed refined, to produce better quality comprehensive populations. It is in these methods that the major novelty of this study is inherent, as creation of this sort of data has very little precedence in the literature. In general, microsimulation models are a relatively new area of research, leaving room for the refinement of techniques to provide such models with high quality input data. That being said, the quality of the comprehensive population was scrutinized in the latter part of Chapter Three, and was found to be, for all intents and purposes, acceptable.

References

Abu-Lughod, J.L. and M.M. Foley (1960) "Consumer strategies" in Housing choices and Housing Constraints, Part 2, N.N. Foote, J.L. Abu-Lughod, M.M. Foley and L. Winnick, editors. New York: McGraw-Hill, pp. 387-447.

Adair, A., S. McGreal, and A. Smyth (2000) "House price and accessibility: the testing of relationships within the Belfast urban area" *Housing Studies*, Vol. 15, No. 5, pp. 699-716.

Alonso, W. (1960) "A theory of the urban land market" *Papers and Proceedings, Regional Science Association,* Vol. 6, pp. 149-158.

Alonso, W. (1964) "Location and land use" Cambridge, MA: Harvard University Press.

Anderson, W., P. Kanaroglou and E. Miller (1994) "Integrated Land Use and Transportation Model for Energy and Environmental Analysis: A report on Design and Implementation", Unpublished Report, McMaster University, Hamilton, ON, Canada.

Anderson, W.P., P.S. Kanaroglou, E.J. Miller, and R.N. Buliung (1996) "Simulating Automobile Emissions in an Integrated Urban Model" *Transportation Research Record*, Vol. 1520, pp. 71-80.

Arentze, T., and H.J.P. Timmermans (2000) "ALBATROSS – A Learning-Based Transportation-Oriented Simulation System" Eindhoven: European Institute for Retailing and Services Studies (EIRASS).

Arentze, T., and H.J.P. Timmermans (2004) "A Learning-Based Transportation-Oriented Simulation System" *Transportation Research Part B*, Vol. 38, 7, pp. 613-633.

Arentze, T., H. Timmermans, and F. Hofman (2007) "Creating Synthetic Household Populations: Problems and Approach" TRB 2007 Annual Meeting proceedings.

Auld, J., A. Mohammadian, and K. Wies (2007) "Population Synthesis With Region-Level Control Variable Aggregation" Paper presented at the 87th annual Transportation Research Board Meeting, January 2008, Washington D.C.

Baccaini, B. (1997) "Commuting and residential strategies in the Ile-de-France: individual behavior and spatial constraints" *Environment and Planning A*, Vol. 29, pp. 1801-1829.

Ballas, D., G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter (2005) "SimBritain: A Spatial Microsimulation Approach to Population Dynamics" *Population, Space and Place*, 11, 13-34.

Barrios Garcia, J.A., and J.E. Rodriguez Hernandez (2007) "Housing and Urban Location Decisions in Spain: An Econometric Analysis with Unobserved Heterogeneity" *Urban Studies*, Vol. 44, No. 9, pp. 1657-1676.

Beckman, R.J., K.A. Baggerly, and M.D. McKay (1996) "Creating Synthetic Baseline Populations" *Transportation Research A*, Vol. 30, No. 6, pp. 415-429.

Beer, A. (1999) "Housing investment and the private rental sector in Australia" *Urban Studies*, Vol. 36, pp. 255-269.

Boehm, T.P., and K.R. Ihlanfeldt (1986) "Residential Mobility and Neighborhood Quality" *Journal of Regional Science*, Vol. 26, pp. 411-424.

Boots, B.N., and P.S. Kanaroglou (1988) "Incorporating the Effects of Spatial Structure in Discrete Choice Models of Migration" *Journal of Regional Science*, Vol. 28, No. 4, pp. 495-509.

Brown, H.J. (1975) "Changes in workplace and residential locations" *Journal of the American Institute of Planners* Vol. 41, pp. 32-39.

Brown, L.A., and E.G. Moore (1970) "The intra-urban migration process: a perspective" *Geografiska Annaler* B, Vol. 52, pp. 1-13.

Burton, I. (1963) "The Quantitative Revolution and Theoretical Geography" *The Canadian Geographer*, Vol. 7, Issue 4, pp. 151-162.

Chang, J. (2006) "Models of the Relationship between Transport and Land-use: A Review" *Transport Reviews*, Vol. 26, No. 3, pp. 325-350.

Chong, P., and C. Jianquan. (2007) "Using multi-agent system for residential expansion models – A case study of Hongshan District, Wuhan City" *Chinese Geographical Science*, Vol. 17, No. 3, pp. 210-215.

Clapp, J. M., and C. Giaccotto (1998) "Residential hedonic models: A rational expectations approach to age effects" *Journal of Urban Economics*, Vol. 44, pp. 415-437.

Clark, W.A.V., M.C. Deurloo, and F.M. Dieleman (1994) "Tenure changes in the context of micro-level family and macro-level economic shifts" *Urban Studies*, Vol. 31, pp. 137-154.

Clark, W.A.V., M.C. Deurloo, and F.M. Dieleman (2006) "Residential mobility and Neighbourhood Outcomes" *Housing Studies*, Vol. 21, No. 3, pp. 323-342.

Clark, W.A.V., M.C. Deurloo and F.M. Dieleman (1986) "Residential mobility in Dutch housing markets" *Environment and Planning A*, Vol. 18, pp. 763-788.

Clark, W.A.V., and Y. Huang (2003) "The life course and residential mobility in British housing markets" *Environment and Planning A*, Vol. 35, pp. 323-339.

Clark, W.A.V., and J.L. Onaka (1983) "Life cycle and housing adjustments as explanations of residential mobility" *Urban Studies,* Vol. 20, pp. 47-57.

Clark, W.A.V., and J.L. Onaka (1985) "An empirical test of a joint model of residential mobility and housing choice" *Environment and Planning A*, Vol. 17, No. 7, pp. 915-930.

Clark, W.A.V., and W.F.J. Van Lierop (1987) "Residential Mobility and Household Location Modelling" *Handbook of Regional Economics*, Vol. 1, P. Nijkamp (ed.), North-Holland Publishing Co., Amsterdam, pp. 97-132.

Clarke, M. (1995) "A Microsimulation approach to demographic and social accounting, in Social and demographic accounting" G.J.D. Hewings and M. Madden, editors. Cambridge: Cambridge University Press, pp. 195-221.

Clarke, M., and E. Holm (1987) "Microsimulation methods in spatial analysis and planning" *Geografiska Annaler* B, Vol. 69, pp. 145-164.

Clarke, M., P.A. Longley, and H.C.W.L. Williams (1989) "Microanalysis and simulation of housing careers: subsidy and accumulation in the UK housing market" *Papers of the Regional Science Association*, Vol. 66, pp. 105-122.

Colwell, P., C. Dehring, and G. Turnbull (2002) "Recreation demand and residential location" *Journal of Urban Economics*, Vol. 51, pp. 418-428.

Connerly, C.E. (1986) "The impact of neighbourhood social relations on prospective mobility" *Social Science Quarterly*, Vol. 67, pp. 186-194.

Coupe, R.T., and B.S. Morgan (1981) "Towards a fuller understanding of residential mobility: a case study in Northampton, England" *Environment and Planning A*, Vol. 13, pp. 201-215.

Crouchley, R., R.B. Davies, and A.R. Pickles (1982) "Identification of some recurrent choice processes" *Journal of Mathematical Sociology*, Vol. 9, pp. 63-73.

Davies Withers, S. (1998) "Linking household transitions and housing transitions: a longitudinal analysis of renters" *Environment and Planning A*, Vol. 30, pp. 615-630.

Davies, R.B., and A.R. Pickles (1983) "The estimation of duration-of-residence effects: a stochastic modelling approach" *Geographical Analysis*, Vol. 15, pp. 305-317.

Deane, G. (1990) "Mobility and adjustments: paths to the resolution of residential stress" *Demography*, Vol. 27, pp. 65-79.

Deming, W.E., and F.F. Stephan (1940) "On a least squares adjustment of a sampled frequency table when the expected marginal tables are known" *Annals of Mathematical Statistics*, Vol. 11, pp. 427-444.

Detang-Dessendre, C., and I. Molho (2000) "Residence Spells and Migration: A Comparison for Men and Women" *Urban Studies*, Vol. 37, No. 2, pp. 247-260.

Devisch, O.T.J., H.J.P. Timmermans, T.A. Arentze, and A.W.J. Borgers (2004) "Towards a Generic Multi-Agent Engine for the Simulation of Spatial Behavioural Processes" in Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning, Kluwer Academic Publishers, Netherlands, pp. 145-160.

Doling, J. (1976) "The family life cycle and housing choice" *Urban Studies*, Vol. 13, pp. 55-58.

Duncan, G.J., and S.J. Newman (1976) "Expected and actual residential mobility" *Journal of the American Institute of Planners*, Vol. 42, pp. 174-186.

Dynarski, M. (1985) "Housing demand and disequilibrium" *Journal of Urban Economics*, Vol. 17, pp. 42-57.

Ellickson, B. (1981) "An alternative test of the hedonic theory of housing markets" *Journal of Urban Economics*, Vol. 9, pp. 56-79.

Erickson, M.A., J.A. Krout, H. Ewen, and J. Robison (2006) "Should I Stay of Should I Go? Moving Plans of Older Adults" *Journal of Housing for the Elderly*, Vol. 20, No. 3, pp. 5-22.

Ermisch, J. (1996) "The demand of housing in Britain and population ageing: microeconometric evidence" *Economica*, Vol. 63, pp. 383-404.

Ermisch, J., and M. Francesconi (2000) "The Increasing Complexity of Family Relationships: Lifetime Experience of Lone Motherhood and Stepfamilies in Great Britain" *European Journal of Population*, Vol. 16, pp. 235-249.

Ettema, D., and H. Timmermans (2006) "Multi-agent modelling of urban systems: a progress report of PUMA System" *Stadt Region Land*, Vol. 81, pp. 165-171.

Ewert, U. and A. Prskawetz (2002) "Can regional variations in demographic structure explain regional differences in car use? A case study in Austria" *Population and Environment*, Vol. 23, pp. 315-345.

Feijten, P. and C.H. Mulder (2002) "The timing of household events and housing events in the Netherlands: a longitudinal perspective" *Housing Studies*, Vol. 17, pp. 773-792.

Fischer, M.M., and P. Nijkamp, (1987) "From Static towards Dynamic Discrete Choice Modelling, A State of the Art Review" *Regional Science and Urban Economics*, Vol. 17, pp. 3-27.

Fletcher, M., P. Gallimore, and J. Mangan (2000) "Heteroskedasticity in hedonic house price models" *Journal of Property Research*, Vol. 17, No. 2, pp. 93-108.

Flowerdew, R., and A. Al-Hamad (2004) "The relationship between marriage, divorce and migration in a British dataset" *Journal of Ethnic and Migration Studies*, Vol. 30, pp. 339-351.

Fokkema, T., and L. Van Wissen (1997) "Moving plans of the elderly: a test of the stress-threshold model" *Environment and Planning A*, Vol. 29, No. 2, pp. 249-268.

Fokkema, T., J. Gierveld, and P. Nijkamp (1996) "Big Cities, Big Problems: Reason for the Elderly to Move?" *Urban Studies*, Vol. 33, No. 2, pp. 353-377.

Fransson, U. (1994) "Interrelationship between household and housing market: A microsimulation model of household formation among the young" *CYBERGEO*, No. 135. Institute for Housing Research, Uppsala University, Sweden. (http://193.55.107.3/durham/fransson/fransson.htm)

Fransson, U., and K. Makila (1994) "Residential Choice in a Time-Space Perspective: A Micro-Simulation Approach" *Neth. J. of Housing and the Built Environment*, Vol. 9, No 3.

Fredland, D. R. (1974) "Residential Mobility and Home Purchase" D. C. Heath: Lexington, Massachusetts.

Freeman, A. M. (1979) "Hedonic prices, property values and measuring environmental benefits: A survey of the issues" *Scandinavian Journal of Economics*, Vol. 81, pp. 154-171.

Frick, M., and K. Axhousen (2004) "Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results" Paper presented at the Swiss Transport Research Conference, March 25-26, 2004.

Gilbert, N., and K.G. Troitzsch (1999) "Simulation for the Social Scientist" Published by Open University Press, Buckingham, GB.

Ginsberg, R.B. (1971) "Semi-Markov processes and mobility" *Journal of Mathematical Sociology*, Vol. 1, pp. 233-262.

Ginsberg, R.B. (1973) "Stochastic models of residential and geographic mobility for heterogeneous populations" *Environment and Planning A*, Vol. 5, pp. 113-124.

Ginsberg, R.B. (1978) "Probability models of residence histories: analysis of times between moves" in Population mobility and residential change. W.A.V. Clark and E.G. Moore, editors. Evanston: Northwestern University, pp. 233-265.

Goldscheider, C. (1971) "Population, modernization, and social structure" Boston: Little, Brown & Co.

Golledge, R.G. (1980) "A behavioural view of mobility and migration research" *Professional Geographer*, Vol. 32, pp. 14-21.

Goodman, A. C. (1978) "Hedonic prices, price indices and housing markets" *Journal of Urban Economics*, Vol. 5, pp. 471-482.

Goodman, A.C. (1995) "A dynamic equilibrium model of housing demand and mobility with transaction cost" *Journal of Housing Economics*, Vol. 4, pp. 307-327.

Goodman, A.C. (2002) "Estimating equilibrium housing demand for 'stayers'" *Journal of Urban Economics*, Vol. 51, pp. 1-24.

Goodman, J.L. Jr. (1976) "Housing consumption disequilibrium and local residential mobility" *Environment and Planning A*, Vol. 8, pp. 855-874.

Gordon, I.R., and I. Molho (1995) "Duration dependence in migration behaviour: cumulative inertia versus stochastic change" *Environment and Planning A*, Vol. 27, pp. 1961-1975.

Goulias, K.G., and R. Kitamura (1992) "Travel Demand Forecasting with Dynamic Microsimulation" *Transportation Research Record*, Vol. 1357, pp. 8-17.

Greenberg, M., and M. Lewis (2000) "Brownfields redevelopment, preferences and public involvement: a case study of an ethnically mixed neighbourhood" *Urban Studies*, Vol. 37, pp. 2501-2514.

Guo, J. Y., and C. R. Bhat (2007) "Population Synthesis For Microsimulating Travel Behavior" *Transportation Research Record: Journal of the Transportation Research Board*, No. 2014, Transportation Research Board of the National Academies, Washington, D. C., pp. 92-101.

Hanushek, E.A., and J.M. Quigley (1978a) "An explicit model of intra-metropolitan mobility" *Land Economics*, Vol. 54, pp. 411-429.

Hanushek, E.A., and J.M. Quigley (1978b) "Housing market diseqilibrium and residential mobility" in Population mobility and residential change. W.A.V. Clark and E.G. Moore, editors. Evanston: Northwestern University, pp. 51-98.

Harding, A., R., A. Lloyd, and A. King (2004) "Assessing Poverty and Inequality at a Detailed Regional Level – New Advances in Spatial Microsimulation" Research Paper No. 2004/26, originally prepared for the UNU-WIDER Conference on Inequality, Poverty and Human Well-Being, May 30-31, 2003, Helsinki.

Harris, R. (1991) "Housing" in Canadian cities in transition. T. Bunting and P. Filion, editors. Toronto: Oxford University Press, pp. 350-378.

Henley, A. (1998) "Residential Mobility, Housing Equity and the Labour Market" *The Economic Journal*, Vol. 108, pp. 414-427.

Henneberry, J. (1998) "Transport investment and house prices" *Journal of Property, Valuation and Investment*, Vol. 16, No. 2, pp. 144-158.

Hooimeijer, P., and A. Oskamp (1996) "A simulation model of residential mobility and housing choice" *Journal of Housing and the Built Environment*, Vol. 11, No. 3, pp. 313-336.

Hooimeijer, P., and A. Oskamp (2000) "Locsim: microsimulation of households and housing market" Paper delivered at the 10th Biennial Conference of the Australian Population Association, Melbourne, Australia.

Huang, Z., and P. Williamson (2002) "A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata" Working Paper, Department of Geography, University of Liverpool.

Huff, J.O., and W.A.V. Clark (1978) "Cumulative stress and cumulative inertia: a behavioural model of the decision to move" *Environment and Planning A*, Vol. 10, pp. 1101-1119.

Hunt, J.D., D.S. Kriger, and E.J. Miller (2005) "Current Operational Urban Land-use-Transport Modelling Frameworks: A Review" *Transport Reviews*, Vol. 25, No. 3, pp. 329-376.

Ioannides, Y.M. (1987) "Residential mobility and housing tenure choice" *Regional Science and Urban Economics*, Vol. 17, pp. 265-287.

Ioannides, Y.M., and K. Kan (1996) "Structural estimation of residential mobility and housing tenure choice" *Journal of Regional Science*, Vol. 36, No. 3, pp. 335-363.

Irwin, M. (2004) "Why People Stay: The Impact of Community Context on Non-migration in the USA" *Population* (English Edition), Vol. 59, pp. 567-592.

Irwin, M., C. Tolbert, and T. Lyson (1999) "There's no place like home: non-migration and civic engagement" *Environment and Planning A*, Vol. 31, pp. 2223-2238.

Johnston, R.J., and C.J. Pattie (1993) "Entropy-Maximizing and the Iterative Proportional Fitting Procedure" *Professional Geographer*, Vol. 45, No. 3, pp. 317- 322.

Kain, J.F., and W.C. Apgar Jr. (1985) "Housing and neighborhood dynamics: a simulation study" Cambridge: Harvard University Press.

Kanaroglou, P., and D. Scott (2002) "Integrated Urban Transportation and Land-use Models for Policy Analysis" in Cities on the Move, pp. 44-73.

Kearns, A., and A. Parkes (2005) "Living in and Leaving Poor Neighbourhood Conditions in England" in Life in Poverty Neighbourhoods: European and American Perspectives. J. Friedrichs, G. Galster and S. Musterd, editors. Routledge.

Kendig, H.L. (1981) "Buying and Renting: Household Moves in Adelaide" Australian Institute of Urban Studies: Canberra.

Kendig, H.L. (1984) "Housing careers, life cycle and residential mobility: implications for the housing market." *Urban Studies*, Vol. 21, pp. 271-283.

Kim, J.H., F. Pagliara, and J. Preston (2005-1) "The Intention to Move and Residential Location Choice Behaviour" *Urban Studies*, Vol. 42, No. 9, pp. 1621-1636.

Kim, T-K., M.W. Horner, and R.W. Marans (2005-2) "Life Cycle and Environmental Factors in Selecting Residential and Job Locations" *Housing Studies*, Vol. 20, No. 3, pp. 457-473.

Kitamura, R. et al. (1996) "The sequenced activity mobility simulator (SAMS): an integrated approach to modelling transportation, land use and air quality" *Transportation*, Vol. 23, pp. 267-291.

Kulu, H. (2008) "Fertility and Spatial Mobility in the Life Course: Evidence from Austria" *Environment and Planning A*, Vol. 40, No. 3, pp. 632-652.

Lancaster, K.J. (1966) "A new approach to consumer theory" *Journal of Political Economy*, Vol. 74, pp. 132-157.

Lee, B.A., R.S. Oropesa, and J.W. Kanan (1994) "Neighborhood Context and Residential Mobility" *Demography*, Vol. 31, No. 2, pp. 249-270.

Leven, C.L., J.T. Little, and H.O. Nourse (1976) "Neighbourhood change: lessons in the dynamics of urban decay" New York: Praeger.

Lewis, R. (1991) "The segregated city: class residential patterns and the development of industrial districts in Montreal, 1861 and 1901" *Journal of Urban History*, Vol. 17, pp. 123-152.

Li, S. (2004) "Life course and residential mobility in Beijing, China" *Environment and Planning A*, Vol. 36, No. 1, pp. 27-43.

Long, L.H. (1972) "The influence of number and ages of children on residential mobility" *Demography*, Vol. 9, pp. 371-382.

Lu, M. (1999) "Do People Move When They Say They Will? Inconsistencies in Individual Migration Behavior" *Population and Environment: A Journal of Interdisciplinary Studies*, Vol. 20, No. 5, pp. 467-488.

Mackett, R.L. (1990) "Comparative analysis of modelling land-use transport interaction at the micro and macro levels" *Environment and Planning A*, Vol. 22, No. 4, pp. 459-475.

Mackett, R.L. (1990) "Exploratory Analysis of Long-Term Travel Demand and Policy Impacts Using Micro-Analytical Simulation." in Developments in Dynamic and Activity-Based Approaches to Travel Analysis, P. Jones, editor. Aldershot: Avebury, pp. 384-405.

Malpezzi, S. (2003) "Hedonic pricing models: a selective and applied review" in Housing Economics and Public Policy, A. O'Sullivan and K. Gibb, editors. Oxford: Blackwell Publishers, pp. 67-89.

Maoh, H., and P. Kanaroglou (2007) "Geographic clustering of firms and urban form: a multivariate analysis" *Journal of Geographical Systems*, Vol. 9, No. 1, pp. 29-52.

Maoh, H., and P. Kanaroglou (2007) "Business Establishment Mobility Behavior in Urban Areas: A microanalyical model for the City of Hamilton in Ontario, Canada" *Journal of Geographical Systems*, Vol. 9, pp. 229 – 252.

Maoh, H., P. Kanaroglou, and R. Buliung (2005) "Modelling the Location of Firms within an Integrated Transport and Land-use Model for Hamilton, Ontario" CSpA Working Paper, McMaster University.

Margulis, H. (2002) "Suburban housing resale prices and housing market restructuring" *Journal of Urban Affairs*, Vol. 24, pp. 461-477.

McCarthy, K. F. (1976) "The household life cycle and housing choices" *Papers of the Regional Science Association*, Vol. 37, pp. 55-80.

McGinnis, R. (1968) "A stochastic model of social mobility" *American Sociological Review*, Vol. 33, pp. 712-722.

McHugh, K.E., P. Gober, and N. Reid (1990) "Determinants of Short- and Long-Term Mobility Expectations for Home Owners and Renters" *Demography*, Vol. 27, No. 1, pp. 81-95.

McLeod, P.B., and J.R. Ellis (1982) "Housing consumption over the life cycle: an empirical analysis" *Urban Studies*, Vol. 19, pp. 177-185.

Melhuish, T., M. Blake, and S. Day (2002) "An Evaluation of Synthetic Household Populations for Census Collection Districts Created Using Spatial Microsimulation Techniques" Paper prepared for the 26th Australia and New Zealand Regional Science Association International (ANZRSAI) Annual Conference, Gold Coast, Queensland, Australia, 29 September – 2 October, 2002.

Meyer, J.W., and A. Speare Jr. (1985) "Distinctively elderly mobility: types and determinants" *Economic Geography*, Vol. 61, pp. 79-88.

Michelson, W. (1977) "Environmental choice, human behaviour, and residential satisfaction" New York: Oxford University Press.

Michelson, W. (1980) "Residential mobility and urban policy: some sociological considerations" *Residential Mobility and Public Policy, Urban Affairs Annual Reviews*, Vol. 19.

Miller, E., J. Hunt, J. Abraham, and P. Salvini (2004) "Microsimulating Urban Systems" *Computers, Environment and Urban Systems*, Vol. 28, pp. 9-44.

Miller, E.J., P.J. Noehammer, and D.R. Ross (1987) "A Micro-Simulation Model of Residential Mobility" Proceedings of the International Symposium on Transport, Communication and Urban Form: 2, Analytical Techniques and Case Studies, W. Young, editor. Clayton, Victoria: Monash University, pp. 217-234.

Moeckel, R., K. Spiekermann, and M. Wegener (2003) "Creating a Synthetic Population" Paper presented at 2003 CUPUM, Sendai, Japan.

Moeckel, R., C. Schurmann, and M. Wegener (2002) "Microsimulation of Urban Land Use" Paper presented at the 42nd European Congress of the Regional Science Association, Dortmund, August 27-31, 2002.

Moore, E.G. (1972) "Residential mobility in the city" Association of American Geographers, Resource Paper No. 13. Washington, D.C. Commission on College Geography.

Morrow-Jones, H.A., and M.V. Wenning (2005) "The Housing Ladder, the Housing Life-cycle and the Housing Life-course: Upward and Downward Movement among Repeat Home-buyers in a US Metropolitan Housing Market" *Urban Studies*, Vol. 42, No. 10, pp. 1739-1754.

Mulder, C.H., and M. Wagner (1993) "Migration and marriage in the life course: a method for studying synchronized events" *European Journal of Population*, Vol. 9, pp. 55-76.

Muth, R.F. (1969) "Cities and housing" University of Chicago Press.

Nijkamp, P., L.J.G. Wissen, and A. Rima (1992) "A household life cycle model for residential relocation behaviour" *Serie Research Memoranda* – 1992-77.

Norman, P. (1999) "Putting IPF on the Researcher's Desk" Working Paper, School of Geography, University of Leeds, United Kingdom.

Onaka, J., and W.A.V. Clark (1983) "A disaggregate model of residential mobility and housing choice" *Geographical Analysis*, Vol. 15, pp. 287-304.

Onaka, J.L. (1983) "A multiple-attribute housing disequilibrium model of residential mobility" *Environment and Planning A*, Vol. 15, pp. 751-765.

Oskamp, A. (1995) "A Microsimulation Approach to Household and Housing Market Modelling" A paper presented to the 1995 Annual Meeting of the American Association of Geographers, Chicago, March 15-18, PDOD Paper No. 29.

Oskamp, A. (1997) "Local Housing Market Simulation: A Micro Approach" NETHURD Publications, Amsterdam.

Ostrovsky, Y. (2004) "Life Cycle Theory and the Residential Mobility of Older Canadians" *Canadian Journal on Aging Supplement*, pp. 23-37.

Parkes, A., A. Kearns, and R. Atkinson (2002) "What makes people dissatisfied with their neighbourhoods?" *Urban Studies*, Vol. 39, pp. 2413-2438.

Perez, P.E., F.J. Martinez, and J. de D. Ortuzar (2003) "Microeconomic formulation and estimation of a residential location choice model: implications for the value of time" *Journal of Regional Science*, Vol. 43, No. 4, pp. 771-789.

Permentier, M., M. van Ham, and G. Bolt (2007) "Behavioural responses to neighbourhood reputations" *Journal of Housing and the Built Environment*, Vol. 22, No. 2, pp. 199-213.

Phipps, A. G. (1989) "Residential Stress and Consumption Disequilibrium in the Saskatoon Housing Market" *Papers of the Regional Science Association*, Vol. 67, pp. 71-87.

Pickles, A.R. (1983) "The analysis of residence histories and other longitudinal panel data: a continuous time mixed Markov renewal model incorporating exogenous variables" *Regional Science and Urban Economics*, Vol. 13, pp. 271-285.

Pickles, A.R., and R.B. Davies (1991) "The empirical analysis of housing careers: a review and general statistical modelling framework" *Environment and Planning A*, Vol. 23, pp. 465-484.

Pickvance, C.G. (1974) "Life cycle, housing tenure and residential mobility: a path analytic approach" *Urban Studies*, Vol. 11, pp. 171-188.

Quigley, J., and D. Weinberg (1977) "Intra-Urban Residential Mobility: A Review and Synthesis" *International Regional Science Review*, Vol. 2, No. 41.

Rindfuss, R.R. et al. (2004) "Household-Parcel Linkages in Nang Rong, Thailand" in People and the Environment: Approaches for Linking Households and Community Surveys to Remote Sensing and GIS. Springer, US.

Rosen, S. (1974) "Hedonic prices and implicit markets: product differentiation in pure competition" *Journal of Political Economy*, Vol. 82, pp. 34-55.

Rossi, P.H. (1955) "Why families move: a study in the social psychology of urban residential mobility" Glencoe: The Free Press.

Rouwendal, J., and E. Meijer (2001) "Preferences for housing, jobs, and commuting: a mixed logit analysis" *Journal of Regional Science*, Vol. 41, No. 3, pp. 475-505.

Salvini, P., and E.J. Miller (2005) "ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems" *Networks and Spatial Economics*, Vol. 5, pp. 217-234.

Scott, A.J. (2000) "Economic geography: the great half-century" *Cambridge Journal of Economics*, Vol. 24, pp. 483-504.

Short, J.R. (1978) "Residential mobility" *Progress in Human Geography*, Vol. 2, pp. 419-447.

Simmons, J.W. (1968) "Changing residence in the city: a review of intra-urban mobility" *Geographical Review*, Vol. 58, pp. 622-651.

Simpson, L., and M. Tranmer (2005) "Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software" *The Professional Geographer*, Vol. 57, No. 2, pp.222-234.

Simpson, L., and M. Tranmer (2005) "Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software" *The Professional Geographer*, Vol. 57, No. 2, pp. 222-234.

Smith, L., R. Beckman, K. Baggerly, D. Anson, and M. Williams (1995) "TRANSIMS: Project Summary and Status May 1995" Los Alamos National Laboratory Report prepared for U.S. Department of Transportation and U.S. Environmental Protection Agency.

Speare, A. Jr. (1974) "Residential satisfaction as an intervening variable in residential mobility" *Demography*, Vol. 11, pp. 173-188.

Speare, A. Jr., and F.K. Goldscheider (1987) "Effects of marital status change on residential mobility" *Journal of Marriage and the Family*, Vol. 49, pp. 455-464.

Speare, A. Jr., S. Golstein, and W.H. Frey (1975) "Residential mobility, migration, and metropolitan change" Cambridge: Ballinger Publishing.

Spiekermann, K., and M. Wegener (2007) "Environmental feedback in urban models" Accepted for publication in: *International Journal of Sustainable Transport*.

St. John, C., M. Edwards, and D. Wenk (1995) "Racial differences in intraurban residential mobility" *Urban Affairs Review*, Vol. 30, pp. 709-729.

Stapleton, C.M. (1980) "Reformulation of the family life-cycle concept: implications for residential mobility" *Environment and Planning A*, Vol. 12, No. 10, pp. 1103-1118.

Strassmann, W.P. (1991) "Housing market interventions and mobility: an international comparison" *Urban Studies*, Vol. 28, 759-771.

Strauss, K. (2008) "Re-engaging with rationality in economic geography: behavioural approaches and the importance of context in decision-making" *Journal of Economic Geography*, Vol. 8, pp. 137-156.

Svinterikou, M. (2007) "Microsimulation Models in Geography Using Object-Oriented Programming: An Application to Residential Mobility" PhD Thesis, University of the Aegean, Lesvos, Greece.

Svinterikou, M., and P. Kanaroglou (2007) "A Microsimulation Approach to the Modelling of Urban Population and Housing Markets Within an Object-Oriented Framework." Working Paper.

Terna, P. (1998) "Simulation Tools for Social Scientists: Building Agent Based Models with SWARM" *Journal of Artificial Societies and Social Simulation*, Vol. 1, No. 2.

Thorns, D. C. (1981) "The implications of differential rates of capital gain from owner occupation for the formation and development of housing classes" *International Journal of Urban and Regional Research*, Vol. 5, pp. 205-17.

Tiebout, C.M. (1956) "A Pure Theory of Local Expenditures" *Journal of Political Economy*, Vol. 64, No. 5, pp. 16-24.

Timmermans, H. (2003) "The Saga of Integrated Land Use-Transport Modelling: How Many More Dreams Before We Wake Up?" Keynote paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, Switzerland.

van Oort, F., A. Weterings, and H. Verlinde (2003) "Residential amenities of knowledge workers and the location of ICT-Firms in the Netherlands" *Tijdschrift voor Economische en Sociale Geografie*, Vol. 94, no. 4, pp. 516-523.

van Wissen, L.J.G. (2002) "Demography of the Firm: A Useful Metaphor?" *European Journal of Population*, Vol. 18, No. 3, pp. 263-279.

Varady, D.P. (1983) "Determinants of residential mobility decisions: the role of government services in relation to other factors" *Journal of the American Planning Association*, Vol. 49, pp. 184-199.

Vencatasawmy, C.P. et al. (1999) "Building a spatial microsimulation model" Paper presented at the 11th European Colloquium on Quantitative and Theoretical Geography in Durham, England, September 3-7.

Voas, D., and P. Williamson (2000) "An evaluation of the combinatorial optimization approach to the creation of synthetic microdata" *International Journal of Population Geography*, Vol. 6, No. 6, pp. 349-366.

Voas, D., and P. Williamson (2001) "Evaluating Goodness-of-Fit Measures for Synthetic Microdata" *Geographical and Environmental Modelling*, Vol. 5, No. 2, pp. 177-200.

Von Thünen, J.H. (1826) "Der isolierte staat in beziehung auf landwirtschaft und nationokonomie" Hamburg: Perthes.

Waddell, P. (1996) "Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, Tenure, and Location Choices" Presented at the 1996 Lincoln Land Institute TRED Conference.

Waddell, P. (2000) "A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim" *Environment and Planning B: Planning and Design*, Vol. 27, pp. 247-263.

Wagner, P., and M. Wegener (2007) "Urban Land Use, Transport and Environmental Models: Experiences with an Integrated Microscopic Approach" *disP*, Vol. 170, pp. 45-56.

Walker, J. L. (2004) "Making Household Microsimulation of Travel and Activities Accessible to Planners" Paper presented at TRB 2005 Annual Meeting.

Wegener, M. (1998) "The IRPUD Model: Overview" http://www.raumplanung.uni-dortmund.de/irpud/pro/mod/mod_e.htm (Accessed May, 2008).

Wegener, M., and K. Spiekermann (1996) "The potential of microsimulation for urban models" London, Pion, pp. 149-163.

Weidner, T., R. Donnelly, J. Freedman, J.E. Abraham, and J.D. Hunt (2006) "TLUMIP – transport land use model in Portland – current state" *Stadt Region Land*, Vol. 81, pp. 91-102.

Wheaton, W.C. (1977) "A bid rent approach to housing demand" *Journal of Urban Economics*, Vol. 4, pp. 200-217.

Williams, P. (2003) "Using Microsimulation to Create Synthetic Small-Area Estimates from Australia's 2001 Census" NATSEM working paper.

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

Wilson, A.G. and C.E. Pownall (1976) "A new representation of the urban system for modelling and for the study of micro-level interdependence" *Area*, Vol. 8, pp.246-254.

Wolpert, J. (1965) "Behavioural aspects of the decision to migrate" *Papers and Proceedings, Regional Science Association*, Vol. 15, pp. 159-169.

Wolpert, J. (1966) "Migration as an adjustment to environmental stress" *Journal of Social Issues*, Vol. 22, pp. 92-102.

Wong, D. (1992) "The Reliability of Using the Iterative Proportional Fitting Procedure" *Professional Geographer*, Vol. 44, No. 3, pp.340-348.

Zhang, Y., and A. Mohammadian (2007) "Microsimulation of Household Travel Survey Data" Paper presented at the 87th annual Transportation Research Board Meeting, January 2008, Washington D.C.

Appendices

Appendix I – Main Body of Code for Linking Individuals to Households

```
# open indivs file
individuals = open('output.csv', 'r')

# determine the number of indivs in the indiv file
first_line_indvs = individuals.readline()
indivs_list01 = individuals.readlines()
num_of_indivs = len(indivs_list01)
print 'num_of_indivs:', num_of_indivs

# put the indivs file into a list of lists
indivs_list = []
for bb1 in indivs_list01:
    temp1 = bb1.split('\t')
    temp1.pop() # this gets rid of the newline character in last pos
    indivs_list.append(temp1)

# close the ind txt file
individuals.close()
                -



# open hhlds file
# note: file won't have dwelling ids yet
households = open('DunFlam_h.txt', 'r')

# determine the number of hhlds in the hhld file
first_line_hhlds = households.readline()
hhlds_list01 = households.readlines()
num_of_hhlds = len(hhlds_list01)
print 'num_of_hhlds:', num_of_hhlds

# put hhlds file into a list of lists
hhlds_list = []
for bb2 in hhlds_list01:
    temp2 = bb2.split('\t')

    #find hhld size
    if temp2[5] == '"1 person"':
        hhld_size0 = 1
    elif temp2[5] == '"2 persons"':
        hhld_size0 = 2
    elif temp2[5] == '"3 persons"':
```

```
43              hhld_size0 = 3
44          elif temp2[5] == "'4 - 5 persons'":
45              hhld_size0 = 4
46          elif temp2[5] == "'6 or more persons'":
47              hhld_size0 = 6
48          else:
49              print 'WTF? hhld size'
50
51          temp2.append(hhld_size0) # so position 7 now contains hhld_size integer!!
52
53          #determine hhld income min and max
54          if temp2[4] == "'Under $10000'":
55              inc_min = 0
56              inc_max = 9999
57          elif temp2[4] == "'$10000 - $19999'":
58              inc_min = 10000
59              inc_max = 19999
60          elif temp2[4] == "'$20000 - $29999'":
61              inc_min = 20000
62              inc_max = 29999
63          elif temp2[4] == "'$30000 - $39999'":
64              inc_min = 30000
65              inc_max = 39999
66          elif temp2[4] == "'$40000 - $49999'":
67              inc_min = 40000
68              inc_max = 49999
69          elif temp2[4] == "'$50000 - $59999'":
70              inc_min = 50000
71              inc_max = 59999
72          elif temp2[4] == "'$60000 - $69999'":
73              inc_min = 60000
74              inc_max = 69999
75          elif temp2[4] == "'$70000 - $79999'":
76              inc_min = 70000
77              inc_max = 79999
78          elif temp2[4] == "'$80000 - $89999'":
79              inc_min = 80000
80              inc_max = 89999
81          elif temp2[4] == "'$90000 - $99999'":
82              inc_min = 90000
83              inc_max = 99999
84          elif temp2[4] == "'$100000 and over'":
85              inc_min = 100000
86              inc_max = 10000000
87          else:
88              print 'something wrong hhld income cats'
89
90          temp2.append(inc_min)  # so position 8 now contains inc_min!!
```

```
91              temp2.append(inc_max)  # so position 9 now contains inc_max!!!
92
93
94              hhlds_list.append(temp2)
95
96
97          # close the hhld txt file
98          households.close()
99
100
101         # sort the hhlds_list in increasing order of tract:
102         def get_first1(itemo):
103             shaggy = itemo[1] # census tract
104             return shaggy
105
106         hhlds_list.sort(key = get_first1)
107
108
109
110         # find the number of CTs
111         ct_list = []
112
113         for thang in hhlds_list:
114             the_ct = round(float(thang[1]),2)
115             if (the_ct in ct_list) == False:
116                 ct_list.append(the_ct)
117                 print the_ct
118
119         cts_num = len(ct_list)
120
121         print 'num of CTs:', cts_num
122         print 'ct_list:', ct_list
123
124
125
126         # create the hhld ranges per CT
127         # these will be the indices in the hhlds list from the bottom of the
128         # census tract, to (the top of it if there are more p1's than hhlds) OR
129         # to (the num of p1's if there are fewer p1's than hhlds)
130         # here p1's refers to the amount of indivs in the CT who are p1's
131         # note: will proceed in 4 steps
132
133          # step1
134         nums_hhlds_in_ct = []  # how many hhlds are there in each CT
135
136         temp_list_of_ct_ids = [] # will list all of the CT ids as they occur
137         for thinger in hhlds_list:
138             temp_list_of_ct_ids.append(float(thinger[1]))
```

```
139
140         for thing in ct_list:
141             how_many = temp_list_of_ct_ids.count(thing)
142             nums_hhlds_in_ct.append(how_many)
143
144         print 'hhlds in cts',nums_hhlds_in_ct
145
146          # step2 (find num of indivs in each CT, assists step3)
147         nums_indivs_in_ct = []
148
149         temp_list_of_indiv_ct_ids = []  # will list CTs from indivs_list as they occur
150         for thinger in indivs_list:
151             temp_list_of_indiv_ct_ids.append(float(thinger[0]))
152
153         for thing in ct_list:
154             how_many = temp_list_of_indiv_ct_ids.count(float(thing))
155             nums_indivs_in_ct.append(how_many)
156
157         print 'indivs in cts', nums_indivs_in_ct
158
159          # step3
160         nums_p1s_in_ct = [] # how many p1s are there in each CT
161
162         temp_list_of_rel2p1s = [] # will list 1 for p1, 0 otherwise (indivs_list is ordered by CT)
163         for thinger in indivs_list:
164             if ('R10' in thinger[2]) == True:
165                 yes_p1 = 0
166             elif ('R1' in thinger[2]) == True:
167                 yes_p1 = 1
168             else:
169                 yes_p1 = 0
170
171             temp_list_of_rel2p1s.append(yes_p1)
172
173
174         k = 0
175         for cola in nums_indivs_in_ct:
176             per1_num = 0
177             for i in range(k, k + int(cola)):
178                 per1_num = per1_num + temp_list_of_rel2p1s[i]
179
180             k = k + int(cola)
181             nums_p1s_in_ct.append(per1_num)
182
183         print 'p1s in cts', nums_p1s_in_ct
184
185          # step4 (to create the hhld ranges per CT)
186         hhld_ranges = []  # will contain [ [begin index, min(p1s,hhlds)index],... ]
```

```
187
188         b_subr = 0
189         count1 = 0
190         for thing in nums_hhlds_in_ct:
191             e_subr = min(int(thing), int(nums_p1s_in_ct[count1])) + b_subr - 1
192             subr = [b_subr, e_subr]
193
194             hhld_ranges.append(subr)
195
196             b_subr = b_subr + int(nums_hhlds_in_ct[count1])
197             count1 = count1 + 1
198
199
200         print 'hhld_ranges:'
201
202         for thing in hhld_ranges:
203             print thing, hhlds_list[thing[0]][1]
204
205
206
207         # create the hhld_use_list.....
208         hhld_use_list = []
209
210         for ccc in hhld_ranges:
211             b_ctr = int(ccc[0])
212             e_ctr = int(ccc[1])
213
214             mid_list = []
215             for hhh in hhlds_list[b_ctr:(e_ctr+1)]:
216                 mid_list.append(hhh)
217
218             # sort mid_list by inc_min (reversed), then by size
219             # should already be sorted by tract, since hhld_ranges are in order
220
221             def get_eigth(itemo8):
222                 shaggy8 = itemo8[8] # inc_min
223                 return shaggy8
224
225             mid_list.sort(key = get_eigth, reverse = True)
226
227
228             def get_seventh(itemo7):
229                 shaggy7 = itemo7[7] # hhld size
230                 return shaggy7
231
232             mid_list.sort(key = get_seventh)
233
234             # get rid of the last hhld (should be of size 6)
```

```
235             mid_list.pop()
236
237             for each_thing in mid_list:
238                 hhld_use_list.append(each_thing)
239
240
241         xxxo = open('use_list.txt','w')
242         for teck in hhld_use_list:
243             xxxo.write(str(teck))
244             xxxo.write('\n')
245         xxxo.close()
246
247
248
249         # create the out_list
250         out_list = []
251
252         # find length of hhld_use_list
253         hhld_use_list_len = len(hhld_use_list)
254
255         zz = 1
256         # for each hhld in the hhld_use_list, do the following..
257         for hhld in hhld_use_list:
258
259             # determine hhld id:
260             hhld_id = int(hhld[0])
261
262             # determine hhld size:
263             hhld_size = int(hhld[7])
264
265             # determine hhld tract:
266             tract = float(hhld[1])
267
268             # determine hhld inc_min and inc_max
269             inc_min = hhld[8]
270             inc_max = hhld[9]
271
272             # determine length of indivs_list (since picked ones are removed!)
273             remaining_indivs_num = len(indivs_list)
274
275             print zz
276             zz = zz + 1
277
278             ### call the function for the hhld size
279             if hhld_size == 1:
280
281                 loco = one_p_hhld()
282
```

```
283
284         elif hhld_size == 2:
285
286             trackr = 0
287             while trackr != 1:
288                 poker_hand = random.randint(1,100)
289
290                 if poker_hand <= 80:
291                     loco = two_p_hhld()
292                     trackr = inc_2plus(loco, 0)
293                 else:
294                     loco = two_p_hhld()
295                     trackr = 1
296
297
298         else:
299
300             trackr = 0
301             while trackr != 1:
302                 poker_hand = random.randint(1,100)
303
304                 if poker_hand <= 80:
305                     loco = three_plus_hhld()
306                     trackr = inc_2plus(loco, 0)
307                 else:
308                     loco = three_plus_hhld()
309                     trackr = 1
310
311         print 'done'
312
313
314         # for each indiv index in loco, copy the indiv to out_list, with hhld id
315         for ppp in range(len(loco)):
316             qqq = loco[ppp]
317             indivs_list[qqq].append(hhld_id)
318             out_list.append(indivs_list[qqq])
319
320
321         # then delete the indivs from the indivs_list
322         loco.sort(reverse=True) # this is so the indexes still work
323         for pepe in range(len(loco)):
324             coco = loco[pepe]
325             indivs_list.pop(coco)
326
327
328
329     # now, assign the remaining indivs to hhlds of 4-5 or 6+ sizes !!!!!!!!
330     #
```

```
331         #
332         # select hhlds randomly from the hhld_use_list until a good 4-5 or 6+ is found....
333
334         done_list = []  # list to contain picked 4-5 hhlds indices
335         # for each indiv left in the indivs_list, do the following..
336         counter = 0
337         for indiv in indivs_list:
338
339            # find some details about the indiv
340            info_list_x = sex_age_rel2p1(counter)
341
342            sex_px = info_list_x[0]
343            age_px = info_list_x[1]
344            rel_px = info_list_x[2]
345
346
347            # if the indiv is per1, then assign them hhld_id = -69
348            # else proceed..
349            if am_i_p1(counter) == 1:
350               indiv.append(-69)
351               out_list.append(indiv)
352
353            # if the indiv is per2, then assign them hhld_id -6969
354            elif rel_px == 2:
355               indiv.append(-6969)
356               out_list.append(indiv)
357
358            else:
359
360               # find tract
361               tract_px = float(indiv[0])
362
363               ## find ranges for that tract
364               #tract_index_in_ct_list = ct_list.index(tract)
365
366               #begin_pos = int(hhld_ranges[tract_index_in_ct_list][0])
367               #end_pos = int(hhld_ranges[tract_index_in_ct_list][1])
368
369
370               # call endgame(), which chooses a good hhld
371               eg = endgame() # eg = [hhld index r from hhlds list, 4 or 6 size]
372               hhld_id1 = int(hhld_use_list[eg[0]][0])
373
374
375               # if hhld is size 4-5, then add its index to the done_list
376               if int(eg[1]) == 4:
377                  done_list.append(eg[0])
378
```

```
379
380             # for each indiv, copy the indiv to out_list, with hhld id
381             indiv.append(hhld_id1)
382             out_list.append(indiv)
383
384         counter = counter + 1 # keeps track of which indiv index we're on
385         print 'x_ind ', counter
386
387
388     # write the out_list to indvs linked.txt
389     outf = open('out_indivs.txt', 'w')
390
391     for cola in out_list:
392         for i in range(8):
393             outf.write(cola[i])
394             outf.write(',')
395
396         outf.write(str(cola[8]))
397         outf.write('\n')
398
399     outf.close()
```

Appendix II – Code for Selecting Individuals for Households of Size 1

```
1     # function for picking indvs for hhld size = 1 *******
2     def one_p_hhld():
3         picks = []
4         until = 0
5         while until != 4:  # 4 is the number of conditions
6             r = random.randint(0, remaining_indivs_num - 1)
7             indiv = indivs_list[r]
8
9             # tracts much match
10            if float(indiv[0]) == tract:
11                until = until + 1
12
13            # individual doesn't live without spouse
14            if indiv[7] == 'Never married (single)':
15                until = until + 1
16            elif indiv[7] == 'Separated but still legally married':
17                until = until + 1
18            elif indiv[7] == 'Divorced':
19                until = until + 1
20            elif indiv[7] == 'Widowed':
21                until = until + 1
22
```

```
23              # individual is person 1
24              if am_i_p1(r) == 1:
25                  until = until + 1
26
27
28              # income must fit hhld income
29              ind_inc_set = indiv_income_range(r)
30              middle_inc = ind_inc_set[2]
31
32              if inc_min >= 60000:
33                  if middle_inc >= 47500:
34                      until = until + 1
35
36              elif inc_min >= 50000:
37                  if 37500 <= middle_inc <= 55000:
38                      until = until + 1
39
40              elif inc_min >= 30000:
41                  if 22500 <= middle_inc <= 37500:
42                      until = until + 1
43
44              else:
45                  if middle_inc <= 22500:
46                      until = until + 1
47
48
49              if until != 4: #this if statement resets until if it's not 4
50                  until = 0
51
52          picks.append(r)
53
54          return picks
```

Appendix III – Code for Selecting Individuals for Households of Size 2

```
1       # function for picking indvs for hhld size = 2 *******
2       def two_p_hhld():
3           picks = []
4
5           # pick first indiv
6           until = 0
7           while until != 2:
8               r1 = random.randint(0, remaining_indivs_num - 1)
9               indiv = indivs_list[r1]
10
```

```
11              # tracts much match
12              if float(indiv[0]) == tract:
13                  until = until + 1
14
15              # individual is person 1
16              if am_i_p1(r1) == 1:
17                  until = until + 1
18
19              if until != 2:
20                  until = 0
21
22          picks.append(r1)
23
24
25          # pick second indiv.....
26
27          # first get info on p1 (age, sex)
28          info_list1 = sex_age_rel2p1(r1)
29          sex_p1 = info_list1[0]
30          age_p1 = info_list1[1]
31
32          mar_p1 = indivs_list[r1][7]
33
34
35          until = 0
36          while until != 6:
37              r2 = random.randint(0, remaining_indivs_num - 1)
38              indiv = indivs_list[r2]
39
40              # get info on sex, age, rel2p1
41              info_list2 = sex_age_rel2p1(r2)
42              sex_p2 = info_list2[0]
43              age_p2 = info_list2[1]
44              rel_p2 = info_list2[2]
45
46
47              # tracts much match
48              if float(indiv[0]) == tract:
49                  until = until + 1
50
51              # individual can't be person 1
52              if am_i_p1(r2) == 0:
53                  until = until + 1
54
55              # individual should sometimes be spouse (rel# = 2)
56              d_roll = random.randint(1,100)
57
58              if mar_p1 == 'Legally married (and not separated)':
```

```
59
60                  if d_roll <= 20:
61                      until = until + 1
62                  else:
63                      if rel_p2 == 2:
64                          until = until + 1
65
66              else:
67
68                  if d_roll <= 20:
69                      until = until + 1
70                  else:
71                      if rel_p2 != 2:
72                          until = until + 1
73
74
75              # sex can't be the same (in most cases)
76              d_roll2 = random.randint(1,100)
77
78              if rel_p2 != 2:
79                  until = until + 1
80              else:
81                  if d_roll2 <= 30:
82                      until = until + 1
83                  else:
84                      if sex_p1 != sex_p2:
85                          until = until + 1
86
87
88              # their ages must be fairly close
89              d_roll3 = random.randint(1,100)
90
91              if rel_p2 != 2:
92                  until = until + 1
93              else:
94                  if d_roll3 <= 10:
95                      until = until + 1
96                  else:
97                      if (age_p1 -3) <= age_p2 <= (age_p1 + 3):
98                          until = until + 1
99
100
101             # if individual is son/daughter (rel 3), age constraints
102             d_roll4 = random.randint(1,100)
103
104             if d_roll4 <= 6:
105                 until = until + 1
106             else:
```

```
107                 if rel_p2 != 3:
108                     until = until + 1
109                 else:
110                     if age_p1 - 8 <= age_p2 <= age_p1 - 3:
111                         until = until + 1
112
113
114             if until != 6:
115                 until = 0
116
117         picks.append(r2)
118
119
120         return picks
```

Appendix IV – Selecting Additional Individuals for Households of Size 3+ (addit-mems)

```
1       # function for picking additional members (3, 4 etc) *******
2         # edad_p1 is age of person1
3         # ya_en_fam = a list of indiv id's that have been picked for this family
4         # but haven't been added to the bigger picked list yet
5
6       def addit_mems(ya_en_fam, edad_p1):
7         # just returns the index of picked indv: r
8
9         until = 0
10        while until != 5:
11          r = random.randint(0, remaining_indivs_num - 1)
12          indiv = indivs_list[r]
13
14          # get info on sex, age, rel2p1
15          info_list_n = sex_age_rel2p1(r)
16          sex_pn = info_list_n[0]
17          age_pn = info_list_n[1]
18          rel_pn = info_list_n[2]
19
20
21          # tracts much match
22          if float(indiv[0]) == tract:
23            until = until + 1
24
25          # individual can't be person 1
26          if am_i_p1(r) == 0:
27            until = until + 1
28
```

```
29              # indiv can't be person 2
30              dice_roll1 = random.randint(1,100)
31              if dice_roll1 <= 15:
32                 until = until + 1
33              else:
34                 if rel_pn != 2:
35                    until = until + 1
36
37              # if individual is son/daughter (rel 3), must be younger
38              dice_roll2 = random.randint(1,100)
39              if dice_roll2 <= 10:
40                 until = until + 1
41              else:
42                 if rel_pn != 3:
43                    until = until + 1
44                 else:
45                    if edad_p1 - 9 <= age_pn <= edad_p1 - 3:
46                          until = until + 1
47
48
49              # individual can't have already been picked
50              donkey = 0
51              for thing in ya_en_fam:
52                 if r == thing:
53                    donkey = 1
54
55              if donkey == 0:
56                 until = until + 1
57
58
59              if until != 5:
60                 until = 0
61
62           return r
```

Appendix V – Code for Selecting Households for Remaining Individuals (end-game)

```
1        # function for picking hhlds of 4-5 or 6+ (for remaining indivs)
2        # should pick these hhlds from the hhld_use_list
3
4        def endgame():
5           # returns [index of picked house, size(4 or 6)]
6           game = []
7           until = 0
8
9           while until != len(done_list) + 2:
```

```
10              r = random.randint(0, hhld_use_list_len - 1)
11
12              # get some info on the family already chosen
13              tract_h = float(hhld_use_list[r][1])  # hhld tract
14
15
16              # check if this hhld has already been chosen (and is 4-5)
17              for clack in done_list:
18                  if r != clack:
19                      until = until + 1
20
21              # check if hhld is a 4 or 6
22              if (int(hhld_use_list[r][7]) in [4,6]) == True:
23                  until = until + 1
24
25              # check if tract matches
26              if tract_px == tract_h:
27                  until = until + 1
28
29
30              # if relation to per1 is 3 (son), check age against per1
31              #if possible
32
33
34
35              if until != len(done_list) + 2:
36                  until = 0
37
38
39          game.append(r)
40
41          lll = hhld_use_list[r][7]
42          game.append(lll)
43
44          return game
```

Appendix VI – Additional Tables from Chapter Four

Table 4.2a: Distribution of Hamilton Individuals by 'Citizenship'

'Citizenship' categories	Count of Individuals
Canadian citizenship	431210
Citizenship other than Canadian	26115
Grand Total	457325

Table 4.2b: 'Sex by Employment' Distribution of Hamilton Individuals

'Sex by Employment' categories	Count of Individuals
M under 15	47650
M Employed	111135
M Unemployed	11015
M Not in the labour force	53575
F under 15	45030
F Employed	95750
F Unemployed	9920
F Not in the labour force	83250
Grand Total	457325

Table 4.2c: '5 Year Mobility Status' Distribution of Hamilton Individuals

'5 Year Mobility Status' categories	Count of Individuals
Non-movers	280975
Movers Non-migrants	110455
Movers Intraprovincial migrants	48385
Movers Interprovincial migrants	4555
Movers External migrants	12955
Grand Total	457325

Table 4.2d: 'Sex by Income' Distribution of Hamilton Individuals, 38 Categories

'Sex by Income' categories	Count of Individuals
M under 15	47650
M Without income	7630
M Under $1000	8320
M $1000 - $2999	5905
M $3000 - $4999	4750
M $5000 - $6999	6155
M $7000 - $9999	9080
M $10000 - $11999	7210
M $12000 - $14999	9015
M $15000 - $19999	13380
M $20000 - $24999	13855
M $25000 - $29999	12600
M $30000 - $34999	13520
M $35000 - $39999	10940
M $40000 - $44999	11015
M $45000 - $49999	8535

M $50000 - $59999	16015
M $60000 and over	17800
F under 15	45030
F Without income	19295
F Under $1000	9225
F $1000 - $2999	10255
F $3000 - $4999	8770
F $5000 - $6999	10360
F $7000 - $9999	15295
F $10000 - $11999	13755
F $12000 - $14999	16650
F $15000 - $19999	19405
F $20000 - $24999	15980
F $25000 - $29999	13605
F $30000 - $34999	10980
F $35000 - $39999	6765
F $40000 - $44999	5175
F $45000 - $49999	3790
F $50000 - $59999	5120
F $60000 and over	4495
Grand Total	457325

Table 4.2e: 'Marital Status' Distribution of Hamilton Individuals

'Marital Status' categories	Count of Individuals
Person under 15	92680
Never married (single)	89015
Legally married (and not separated)	209720
Separated but still legally married	11655
Divorced	26820
Widowed	27435
Grand Total	457325

Table 4.2f: 'Sex by Age by Relation to Person 1' Distribution of Hamilton Individuals, 180 Categories

'Sex by Age by Relation to Person 1' categories	Count of Individuals (Females)	'Sex by Age by Relation to Person 1' categories	Count of Individuals (Males)
F A1 R1	0	M A1 R1	0
F A1 R2	0	M A1 R2	0
F A1 R3	14290	M A1 R3	14975
F A1 R9	595	M A1 R9	685

F A1 R10	60	M A1 R10	95
F A2 R1	0	M A2 R1	0
F A2 R2	0	M A2 R2	0
F A2 R3	14825	M A2 R3	15645
F A2 R9	415	M A2 R9	450
F A2 R10	85	M A2 R10	40
F A3 R1	0	M A3 R1	0
F A3 R2	0	M A3 R2	0
F A3 R3	14405	M A3 R3	15250
F A3 R9	225	M A3 R9	365
F A3 R10	130	M A3 R10	145
F A4 R1	395	M A4 R1	280
F A4 R2	260	M A4 R2	30
F A4 R3	12435	M A4 R3	13790
F A4 R9	425	M A4 R9	290
F A4 R10	420	M A4 R10	425
F A5 R1	3160	M A5 R1	2425
F A5 R2	2145	M A5 R2	540
F A5 R3	8685	M A5 R3	10370
F A5 R9	455	M A5 R9	645
F A5 R10	985	M A5 R10	1060
F A6 R1	5845	M A6 R1	7395
F A6 R2	6545	M A6 R2	1575
F A6 R3	3260	M A6 R3	5555
F A6 R9	415	M A6 R9	895
F A6 R10	535	M A6 R10	955
F A7 R1	6425	M A7 R1	12635
F A7 R2	11165	M A7 R2	2645
F A7 R3	1480	M A7 R3	2585
F A7 R9	330	M A7 R9	370
F A7 R10	390	M A7 R10	695
F A8 R1	6595	M A8 R1	14335
F A8 R2	11630	M A8 R2	2460
F A8 R3	865	M A8 R3	1405
F A8 R9	310	M A8 R9	470
F A8 R10	240	M A8 R10	485
F A9 R1	5690	M A9 R1	13300
F A9 R2	11490	M A9 R2	1935
F A9 R3	445	M A9 R3	920
F A9 R9	210	M A9 R9	245
F A9 R10	260	M A9 R10	495
F A10 R1	5180	M A10 R1	12705

F A10 R2	10875	M A10 R2	1685
F A10 R3	295	M A10 R3	550
F A10 R9	120	M A10 R9	160
F A10 R10	225	M A10 R10	300
F A11 R1	3500	M A11 R1	10760
F A11 R2	8370	M A11 R2	1155
F A11 R3	295	M A11 R3	255
F A11 R9	185	M A11 R9	140
F A11 R10	135	M A11 R10	215
F A12 R1	2950	M A12 R1	9240
F A12 R2	6995	M A12 R2	885
F A12 R3	120	M A12 R3	65
F A12 R9	470	M A12 R9	155
F A12 R10	130	M A12 R10	140
F A13 R1	3465	M A13 R1	8805
F A13 R2	6390	M A13 R2	745
F A13 R3	60	M A13 R3	50
F A13 R9	365	M A13 R9	245
F A13 R10	120	M A13 R10	175
F A14 R1	4280	M A14 R1	8520
F A14 R2	5760	M A14 R2	585
F A14 R3	30	M A14 R3	0
F A14 R9	465	M A14 R9	240
F A14 R10	70	M A14 R10	125
F A15 R1	4990	M A15 R1	7135
F A15 R2	4755	M A15 R2	460
F A15 R3	30	M A15 R3	10
F A15 R9	585	M A15 R9	235
F A15 R10	70	M A15 R10	70
F A16 R1	4255	M A16 R1	4255
F A16 R2	1950	M A16 R2	375
F A16 R3	0	M A16 R3	0
F A16 R9	470	M A16 R9	165
F A16 R10	30	M A16 R10	40
F A17 R1	3180	M A17 R1	2315
F A17 R2	950	M A17 R2	160
F A17 R3	0	M A17 R3	0
F A17 R9	560	M A17 R9	120
F A17 R10	20	M A17 R10	40
F A18 R1	1955	M A18 R1	1035
F A18 R2	290	M A18 R2	55
F A18 R3	0	M A18 R3	0

F A18 R9	455	M A18 R9	135
F A18 R10	35	M A18 R10	10
Grand Total: 457325			

Table 4.3a: 'Relation to Person 1' Distribution of Hamilton Individuals, Original Classification Scheme

'Relation to Person 1' categories, original	Count of Individuals
R1, Person 1	177005
R2, Person 1's spouse or common-law partner	104860
R3, Person 1's son or daughter	152945
R4, Person 1's father or mother	2230
R5, Person 1's brother or sister	2365
R6, Person 1's son or daughter in-law	1210
R7, Person 1's father or mother in-law	1635
R8, Person 1's brother or sister in-law	975
R9, Other relatives of Person 1	4650
R10, Persons not related to Person 1	9450
Grand Total	457325

Table 4.3b: 'Relation to Person 1' Distribution of Hamilton Individuals, Regrouped Classification Scheme

'Relation to Person 1' categories, original	Count of Individuals
R1, Person 1	177005
R2, Person 1's spouse or common-law partner	104860
R3, Person 1's son or daughter	152945
R9, Other relatives of Person 1	13065
R10, Persons not related to Person 1	9450
Grand Total	457325

Table 4.3c: 'Age' Distribution of Hamilton Individuals

'Age' categories	Count of Individuals
A1, 0-4	30700
A2, 5-9	31460
A3, 10-14	30520
A4, 15-19	28750
A5, 20-24	30470
A6, 25-29	32975
A7, 30-34	38720
A8, 35-39	38795

A9, 40-44	34990
A10, 45-49	32095
A11, 50-54	25010
A12, 55-59	21150
A13, 60-64	20420
A14, 65-69	20075
A15, 70-74	18340
A16, 75-79	11540
A17, 80-84	7345
A18, 85+	3970
Grand Total	457325

Table 4.3d: 'Sex' Distribution of Hamilton Individuals

'Sex' categories	Count of Individuals
F, Female	233950
M, Male	223375
Grand Total	457325

Table 4.4a: Level of Schooling categories

'HLOSP' categories	Description
99	Not applicable (Person under 15 yrs old)
1	Less than Grade 5
2	Grades 5 to 8
3	Grades 9 to 13
4	Secondary (high) school graduation certificate
5	Trades certificate or diploma
6	Non-university: Without trades or other non-university certificate or diploma
7	Non-university: With trades certificate or diploma
8	Non-university: With other non-university certificate or diploma
9	University: Without certificate, diploma or degree
10	University: With university or other non-university certificate or diploma
11	University: With bachelor or first professional degree
12	University: With certificate or diploma above bachelor lever

13	University: With master's degree(s)
14	University: With earned doctorate

Table 4.4b: Standard Industrial Classification categories

'IND80P' categories	Description
99	Person under 15 yrs old & those unemployed since Jan 1st, 1995
1	Agriculture
2	Other primary industries
3	Manufacturing
4	Construction
5	Transportation and storage
6	Communication and other utilities
7	Wholesale trade
8	Retail trade
9	Finance, insurance and real estate
10	Business services
11	Government services: Federal
12	Government services: Other
13	Educational services
14	Health and social services
15	Accommodations, food and beverage services
16	Other services

Table 4.4c: Occupation categories

'OCC91P' categories	Description
99	Person under 15 yrs old & those unemployed since Jan 1st, 1995
1	Senior managers
2	Middle and other managers
3	Professionals
4	Semi-professionals and technicians
5	Supervisors
6	Supervisors: crafts and trades
7	Administrative and senior clerical personnel
8	Skilled sales and service personnel
9	Skilled crafts and trades workers
10	Clerical personnel
11	Intermediate sales and service personnel
12	Semi-skilled manual workers
13	Other sales and service personnel

14	Other manual workers

Table 4.5a: Distribution of Hamilton Households by 'Tenure'

'Tenure' categories	Count of Households
Owned	110845
Rented	66160
Grand Total	177005

Table 4.5b: Distribution of Hamilton Households by 'Income'

'Income' categories	Count of Households
Under $10000	12665
$10000 - $19999	29940
$20000 - $29999	23300
$30000 - $39999	20380
$40000 - $49999	18520
$50000 - $59999	17835
$60000 - $69999	13475
$70000 - $79999	11585
$80000 - $89999	8810
$90000 - $99999	6085
$100000 and over	14410
Grand Total	177005

Table 4.5c: Distribution of Hamilton Households by 'Size'

'Size' categories	Count of Households
1 person	45747
2 persons	55266
3 persons	29451
4 - 5 persons	41319
6 or more persons	5222
Grand Total	177005

Table 4.5d: Distribution of Hamilton Households by 'Structure'

'Structure' categories	Count of Households
Single-detached house	103525
Semi-detached house	5520
Row house	13745
Apartment detached duplex	4955
Apartment building five or more storeys	31935

Apartment building less than five storeys	16470
Other single attached house	480
Movable dwelling	375
Grand Total	177005

Appendix VII – Variable Re-Classifications: Individuals

For URM-Microsim, there are 9 variables of interest for individuals: Sex; Date of Birth; Marital Status; Position in Household; Education; Employment Status; Income; Occupation; Industry.

The individuals synthesized for the City of Hamilton are endowed with many attributes, including: sex by age by relationship to Person 1; sex by employment; sex by income; marital status; highest level of schooling; occupation; industry. Also, individuals can be linked to their original sample record, which gives access to a plethora of attributes.

Note: For reference, the exact classification scheme used by URM-Microsim can be found in Appendix J.

Proceeding through the 9 variables required by URM-Microsim:

Sex – This can be derived from sex by age by relationship to Person 1

Date of Birth – This can be inferred from the numerical age of individuals, through linking with the original sample record. The month and day can be randomly assigned, or drawn from a distribution.

Marital Status – for each Hamilton Classification, assign the corresponding value which corresponds to the needs of URM-Microsim

Hamilton Classification	URM-Microsim value
Person under 15	Never married
Never married (single)	Never married
Legally married (and not separated)	Married
Separated but still legally married	Divorced
Divorced	Divorced
Widowed	Widowed

Position in Household –

Hamilton Classification	URM-Microsim value
Person 1	Person 1
Person 1's spouse or common-law partner	Couple
Person 1's son or daughter	Child

Other relatives of person 1	Single
Persons not related to person 1	Flat-mate

Education -

Hamilton Classification	URM-Microsim value
Not Applicable (under 15 yrs old)	2
Less than Grade 5	2
Grades 5 to 8	3
Grades 9 to 13	4
High School graduation certificate	4
Trades certificate or diploma	5 or 6
Non-university, without diploma	5 or 6
Non-university, with diploma	5 or 6
Non-university, other certificate or diploma	5 or 6
University, without diploma	7
University, certificate (below Bachelor)	7
University, Bachelor degree	7
University, certificate beyond Bachelor	7
University, Master's degree	7
University, with earned doctorate	7

Employment status –

Hamilton Classification	URM-Microsim value
Under 15	11 (school child)
Employed	1
Unemployed	5 (unemployed)
Not in the labour force	7,9,10

Income – This can be taken from sex by income.

Occupation –

Hamilton Classification	URM-Microsim value
Not applicable (less than 15 yrs, or unemployed)	
Senior managers	1
Middle and other managers	1
Professionals	2
Semi-professionals and technicians	3
Supervisors	1
Supervisors: crafts and trades	7
Administrative and senior clerical personnel	4
Skilled sales and service personnel	5
Skilled crafts and trades workers	7

Clerical personnel	4
Intermediate sales and service personnel	5
Semi-skilled manual workers	8 or 9
Other sales and service personnel	5
Other manual workers	9

Industry – This can be easily derived from the industry variable.

Appendix VIII - Variable Re-Classifications: Households

For URM-Microsim, there are 3 variables of interest for households: size; income; tenure.

The households synthesized for Hamilton are endowed with attributes including: size; income range; tenure.

Proceeding through the 3 required variables:

Size – URM-Microsim requires an integer for size, while the Hamilton data has the following categories: 1, 2, 3, 4-5, 6+. In order to make the conversion, the individuals belonging to each household have to be counted, and the new attribute assigned to households.

Income – Although households have an income range, the best practice is to sum the household members' incomes.

Tenure – This corresponds to the Hamilton variable.

Appendix IX - Variable Re-Classifications: Dwellings

For URM-Microsim, there are 5 variables of interest for dwellings: size; number of rooms; market value; rent value; availability.

The dwellings synthesized for Hamilton are endowed with attributes including: dwelling ID; building ID; Census Type (describing building structure); Property Code (describing land use of the parcel); Number of Rooms.

Proceeding through the 5 required variables:

Size – This refers to area, and can be found by dividing the square footage of the dwelling's building by the number of dwellings in the building (both attributes of buildings which can be accessed using the 'building ID' linking dwellings to buildings.

Number of Rooms – Can be taken directly from the corresponding Hamilton variable.

Market Value and Rent Value – For each CT, averages of these values are presented. We weighted these averages by the number of rooms in each dwelling, to obtain a distribution of values.

Availability – This can be found by searching through the household database for Dwelling IDs, if the Dwelling ID can't be found, then the dwelling is vacant and hence available.

Appendix X - Variable re-classifications: Buildings

For URM-Microsim, there are 5 variables of interest for buildings: Type (describing structure); Floor Space; Number of Floors; Number of Dwellings; Year of Construction.

The buildings synthesized for Hamilton are endowed with attributes including: Area; Date of Construction; Property Code; Res1 (a dummy variable indicating if the parcel contains a residential building); Census Type (describing structure); Number of Dwellings; Rooms Per Dwelling.

Proceeding through the 5 required variables:

Type - here is the procedure for obtaining a URM-Microsim 'Type' value from a building in the Hamilton synthetic population:

To begin with, look at Res1 (Hamilton), which is 0 for non-residential, 1 for residential.

If Res1 = 1:

Then we look at Number of Dwellings (Hamilton), and take a value of 1, 2, 3+

(If Number of Dwellings = 0, assign building type as 'Other' (#10))

If Res 1 = 0:

Then we look at Property Code (Hamilton):

If Property Code = 701, Then assign building type as 'Church' (#4)

If Property Code = 621 or 627, Then assign building type 'Hospital' (#5)

If Property Code = 363; 383; 440; 450; 451; 460, Then assign 'Hotel' (#6)

If Property Code = 730; 731; 720; 735, Then assign 'Public' (#7)

If Property Code = 608; 605; 601; 610, Then assign 'School' (#8)
(I'm including universities in here.. 601)

If Property Code = {400; 401; 405; 475; 406; 410; 428; 429; 430; 431; 473; 499; 711; 415; 411; 412; 441; 422; 420; 421; 705; 432; 750; 472}, Then assign 'Shop/ Office' (#9)

For other values of Property Code, assign 'Other' (#10)

Note that the Property Code values are interpreted in publicly available Canadian Census documentation (pumfh).

The remaining variables can be taken directly from the corresponding Hamilton variables.

Appendix XI – The URM-Microsim Database Domains

Source: Svinterikou, 2007.