Synthesizing Agents and Relationships for Land Use / Transportation Modelling

by

David R. Pritchard

A thesis submitted in conformity with the requirements
for the degree of Masters of Applied Science
Graduate Department of Civil Engineering
University of Toronto

Abstract

Synthesizing Agents and Relationships for Land Use / Transportation Modelling

David R. Pritchard

Masters of Applied Science

Graduate Department of Civil Engineering

University of Toronto

2008

Agent-based microsimulation models of socioeconomic processes require an initial synthetic population derived from census data. This thesis builds upon the Iterative Proportional Fitting (IPF) synthesis procedure, which has well-understood statistical properties and close links with log-linear models. Typical applications of IPF are limited in the number of attributes that can be synthesized per agent. A new method is introduced, implementing IPF with a sparse list-based data structure that allows many more attributes per agent. Additionally, a new approach is used to synthesize the relationships between agents, allowing the formation of household and family agents in addition to individual person agents. Using these methods, a complete population of persons, families, households and dwellings was synthesized for the Greater Toronto Area and Hamilton.

Acknowledgments

First and foremost, my supervisor Eric Miller has been a source of invaluable training, wisdom, discussion and insight throughout this thesis. Matt Roorda also provided useful suggestions for tackling the real-world issues encountered in this work. I owe Laine Ruus a great debt for digging up the historical data that was required for the thesis. This research builds on the foundations laid by others, particularly Paul Salvini and Juan Carrasco. Finally, I am grateful for fruitful conversations with fellow students, particularly Bilal Farooq and David McElroy.

To Josie, an especially warm thanks for technical discussions, interdisciplinary asides and moral support through an intense year of research.

Finally, my degree at the University of Toronto was supported by scholarships from the Ontario Graduate Scholarship program and the Transportation Association of Canada.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traditional efforts to model transportation in large city regions operated at an aggregate level, splitting the urban area into a small number of zones and forecasting trips between these zones. The classic four-stage Urban Transportation Modelling System (UTMS) is a common example, including a gravity model to distribute trips between zones.

Aggregate models suffer from limited sensitivity to interesting policy questions [51]. While aggregate approaches can be suitable for projecting a continuation of current trends, they are unable to anticipate the effects of many major policy changes. For example, it would be difficult to model the effects of introducing road pricing or urban growth boundaries, or to project the response to major structural changes in the economics of transportation.

Disaggregate models may prove more suitable for tackling such questions, by modelling the behaviour of individual persons and households. While it is hard to understand the behaviour of a large group of persons with only aggregate statistics about these persons, behaviour is easier to grasp at the level of the individual person or household. Disaggregate models do not aim to *predict* the behaviour of individuals, but to understand behaviour at that level and use it to make accurate projections at

the aggregate level.

Agent-based microsimulation models represent the finest level of disaggregation in current practice. These models forecast the future state of an aggregate system by simulating the behaviour of a number of individual *agents* over time. In travel demand modelling, the system is usually the spatial arrangement of travel patterns (including the mode of travel used), and the agents are usually persons, families or households. The execution of such a model can be divided into two steps: the creation of an initial set of agents, describing each agent and the system's state at some initial time; and a series of subsequent steps forward, where the state of each agent and the system as a whole is advanced by a timestep (for example, one year per step).

The construction of the initial set of agents is often known as *population synthesis*, since a "population" of agents must be created. Data is typically not available for the true persons and their attributes at the initial time; hence the initial population is synthetic. A good representation is critical to support a good microsimulation model; "Garbage In, Garbage Out," is a common phrase in computer science, implying that a good method will still produce bad results if its input is poor.

When analyzing behaviour at the level of individual persons, it is possible to observe and model interesting connections between persons. For example, members of a family do not act entirely independently; they share resources and may choose to travel together in a single vehicle, to adjust their travel patterns to suit each others' schedules, or to make decisions about home ownership based on all family members' needs. However, to represent both individual behaviour and family-level behaviour in an agent-based framework, the relationships between individual persons must be known to form family units.

This thesis focuses on these problems, examining the methods necessary to construct a complete population of persons, families and households for the Integrated Land Use, Transportation and Environment (ILUTE) modelling effort at the University

of Toronto. In particular, much of the thesis is concerned with the Iterative Proportional Fitting (IPF) method, a data fusion technique that underlies most population synthesis procedures. While the ILUTE model is the specific context for this thesis, the methods and discussion are relevant to a broader audience. It should be useful to anyone performing agent-based simulation using census data, and may provide new insights to anyone using Iterative Proportional Fitting procedure for data fusion.

The remainder of this thesis is structured as follows. First, a review of the previous work is conducted, covering the ILUTE model, a discussion of the mathematics and notation used for fitting contingency tables, and earlier population synthesis procedures. In the following chapter, the data used for synthesis here is reviewed, including definitions of the agents, attributes, and population universes. Chapter 4 takes a "brainstorming" approach to some of the problems with existing population synthesis procedures, and discusses some potential improvements to established method. This carries directly into the following chapter, which covers the implementation of the ILUTE population synthesizer, including a detailed application of many of the new ideas. Subsequently, the next chapter uses this implementation to conduct a series of experiments to evaluate the new methodological ideas. The final chapter looks at the results of the final synthesis, and summarizes the results of the thesis.

Chapter 2

Previous Work

The research described in this thesis draws on a broad body of knowledge. This literature review begins with a section on the context for this population synthesis effort, the Integrated Land Use Transportation, Environment (ILUTE) model.

The following section describes the mathematics and algorithms used in population synthesis, starting with a discussion of notation for contingency tables. The properties and history of the Iterative Proportional Fitting (IPF) procedure are reviewed, and some generalizations of the method are discussed in the following section. The discussion then shifts to log-linear modelling for contingency tables, and then looks briefly at the literature connecting IPF and log-linear modelling. The final section reviews the literature dealing with zeros in contingency tables.

The review then shifts to the methods used for population synthesis. Two broad classes of method are included: those using IPF and those using the Combinatorial Optimization method.

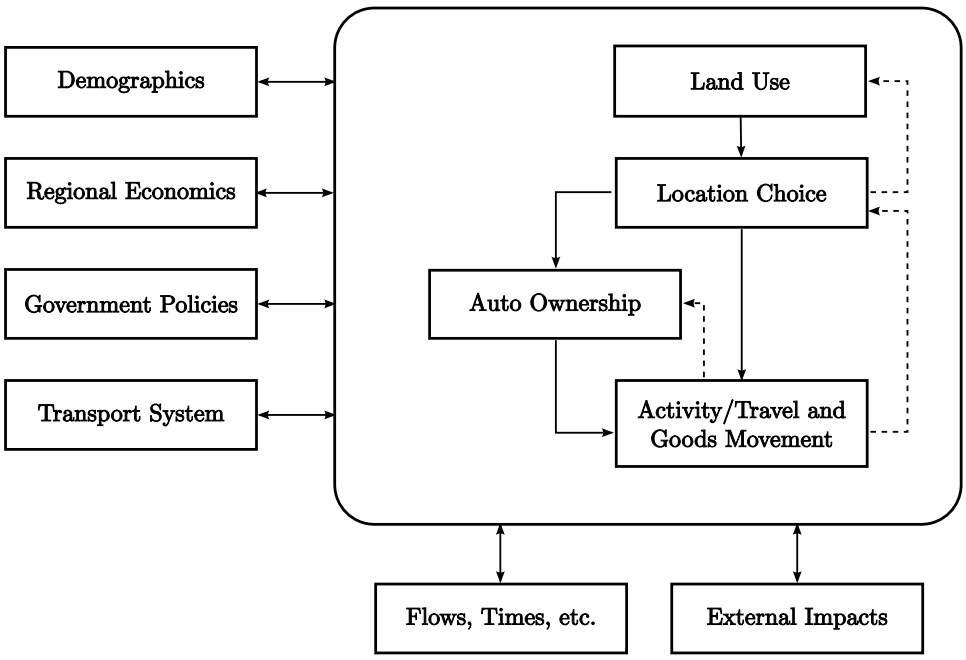

Figure 2.1: The idealized integrated urban modelling system envisioned by Miller, Kriger & Hunt [34].

2.1 The ILUTE Model

The ILUTE research program aims to develop next generation models of urban land use, travel and environmental impacts. The project's ultimate goal is the "ideal model" described by Miller et al. [34]. As shown in Figure 2.1, the behavioural core of an ideal model would include land development, residential and business location decision, activity/travel patterns and automobile ownership. The boxes on the left show the main influences on the urban system: demographic shifts in the population, the regional economy, government policy and the transport system itself. Some of these may be exogenous inputs to the model, but Miller et al. suggest that both demographics and regional economics need to be at least partially endogenous.

The ILUTE model is intended to operate in concert with an activity-based travel demand model. The Travel/Activity Scheduler for Household Agents (TASHA) is an activity-based model designed on disaggregate principles similar to ILUTE, and

connects personal decision making with household-level resources and activities to form travel chains and tours [35, 36].

The operational prototype of the ILUTE system was described in detail by Salvini & Miller [40, 39]. To validate the model, it is intended to be run using historical data, allowing comparison against the known behaviour of the urban system over recent years. The baseline year of 1986 was ultimately chosen as a starting point, since the Transportation Tomorrow Survey of travel behaviour in the Greater Toronto Area was first conducted in that year.

The prototype defines a wide range of agents and objects: persons, households, dwellings, buildings, business establishments and vehicles. It also defines various relationships between these agents and objects: in particular, family relationships between persons in households, occupancy relationships between households and their dwellings, ownership of dwellings/vehicles by households or persons, containment of dwellings within buildings, and employment of persons by business establishments.

These represent the full spectrum of possible agents and relationships that need to be synthesized as inputs to the ILUTE model. In earlier work within the ILUTE framework, Guan synthesized persons, families and households and a set of relationships between them [25]. In this thesis, the same agents and relationships are considered (in addition to dwelling units), with the goal of improving the method and quality of the synthetic populations.

The remaining agents and relationships are also important to the ILUTE model, but the focus here is on the demographic and dwelling attributes since these are central to both the ILUTE and TASHA models, and because rich data from the Canadian census is available to support the synthesis. In this research, families are proposed as a new class of agent for the ILUTE modelling framework. While the family is a central theme in both the ILUTE and TASHA models, it was only modelled distinct from the

household in Guan's work. Furthermore, the original ILUTE prototype did not allow for multifamily households.

2.2 Mathematics for Fitting Contingency Tables

Almost all population synthesis procedures rely on data stored in multiway contingency tables. To help understand and explain this type of data, a consistent notation is first defined, and then the mathematical properties of contingency tables and the Iterative Proportional Fitting procedure are described.

2.2.1 Notation

Throughout this document, scalar values and single cells in contingency tables will be represented using a regular weight typeface (e.g., n or n_{ijk}). Multiway contingency tables and their margins will be represented with boldface (e.g., $\mathbf{n}$ or $\mathbf{n}_{ijk}$) to indicate that they contain more than one cell. Contingency tables may be one-way, two-way or multiway; the number of subscripts indicates the dimension of the table (e.g., $\mathbf{n}_{ijk}$).

Suppose three variables X, Y and Z vary simultaneously, and are classified into I, J and K categories respectively. The variables may be either inherently discrete or continuous, but continuous variables are grouped into a finite set of discrete categories. The variable i denotes a category of X, and the categories are labelled as $\{1, 2, \ldots, I\}$, and likewise for Y and Z. (For example, suppose that these variables represent the attributes of a person, such as age, education and location.) Then, there is a probability π_{ijk} that a random observation will be classified in category i of the first variable, category j of the second variable and category k of the third variable. There are $C = I \times J \times K$ cells in the table, each of which consists of a count n_{ijk} of the number of observations with the appropriate categories. Since the table consists of counts, the cells are Poisson distributed; these counts are observations of the under-

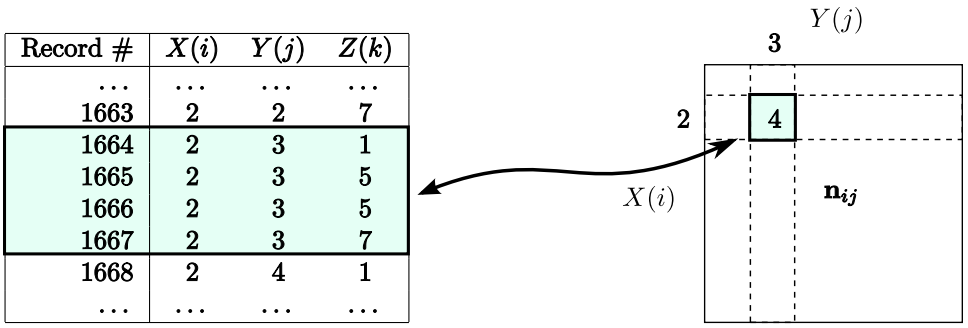

Figure 2.2: The link between list-based and contingency table representations of multivariate categorical data. Left: a list of observations, where each row represents a single observation. Variables X, Y and Z are observed to fall into different categories. Right: a cross-tabulation of the observations using only variables X and Y. Each cell n_{ij} in the table is a count of observations with a given value $X = i$ and $Y = j$. It corresponds to a specific set of n_{ij} observations from the list-based representation.

lying multinomial probability mass function $\boldsymbol{\pi}_{ijk}$. The contingency table has a direct relationship to the list of observations; Figure 2.2 shows an example where a list of observations of three variables is used to form a two-variable contingency table.

Any contingency table can be *collapsed* to a lower-dimensional table by summing along one or more dimensions; a collapsed table is called a *marginal table* or margin. The notation $\mathbf{n}_{i++}$ is used for the margin where the second and third variables are collapsed, leaving only the breakdown of the sample into the I categories of variable X. The + symbols in the notation indicate that the margin is derived by summing $\mathbf{n}_{ijk}$ over all categories j and k. The total size of the tabulated sample is given by n_{+++}, or more typically by n alone.

In this paper, multiple contingency tables are often considered simultaneously. In a typical application of the Iterative Proportional Fitting (IPF) procedure, a "source" population is sampled and cross-classified to form a multiway table $\mathbf{n}_{ij}$. A similarly structured multiway table $\mathbf{N}_{ij}$ is desired for some target population, but less information is available about the target: typically, some marginal totals $\mathbf{N}_{i+}$ and $\mathbf{N}_{+j}$ are known. (Depending upon the application, the target and source populations may

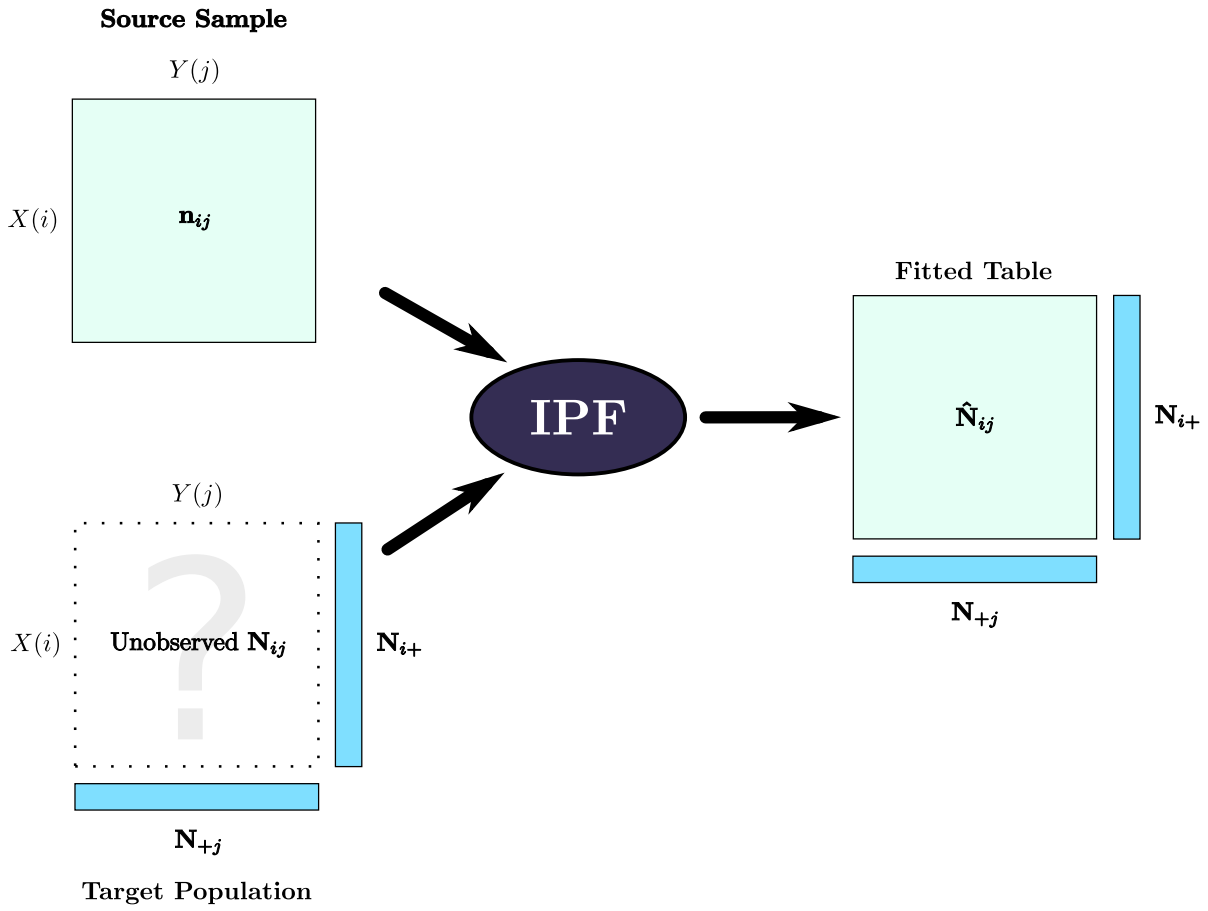

Figure 2.3: An illustration of the Iterative Proportional Fitting procedure with two variables X and Y. The source table $\mathbf{n}_{ij}$ is modified to match the known target marginals $\mathbf{N}_{i+}$ and $\mathbf{N}_{+j}$, producing a fitted table $\hat{\mathbf{N}}_{ij}$ that approximates the unknown target table $\mathbf{N}_{ij}$.

be distinct or identical; in a common example, the populations are identical but the source sample is small (1–5%) while the target margins may be a complete 100% sample of the population.) The complete multiway table $\mathbf{N}_{ij}$ of the target population is never known, but the IPF procedure is used to find an estimate $\hat{\mathbf{N}}_{ij}$. This is achieved through repeated modifications of the table n_{ij}. The entire process and associated notation are shown in Figure 2.3 and Table 2.1.

Note that the source table n_{ij} and target margins N_{i+} are usually integer counts, but the estimated target table $\hat{N}_{ij}$ produced by Iterative Proportional Fitting is real-

Symbol	Description
C	The total number of cells in the contingency table, $C = I \times J \times K$.
I, J, K	The number of categories for variables X, Y and Z respectively, the three dimensions of the multiway tables.
$\mathbf{n}_{ijk}$ or $\mathbf{n}$	A multiway contingency table of the source sample, of size $I \times J \times K$.
n	The size of the source sample. i.e., $\sum_{i,j,k} n_{ijk}$
n_{ijk}	A single cell of $\mathbf{n}$, containing the count of observations in the source sample where variable X was in category i, Y was in category j and Z was in category k.
$\mathbf{N}_{ijk}$ or $\mathbf{N}$	A multiway contingency table of the target population, of the same size as $\mathbf{n}$; never observed.
N	The size of the target population.
N_{ijk}	A single cell in target table $\mathbf{N}$; never observed.
$\mathbf{N}_{i++}$	A one-way table containing a margin of $\mathbf{N}_{ijk}$ showing the total observations for each category i of variable X. While the full table $\mathbf{N}_{ijk}$ of the target population is never observed, some margins are known.
N_{i++}	A single entry in $\mathbf{N}_{i++}$. The + symbols indicate a sum over all categories in that dimension; that is, $N_{i++} = \sum_{j,k} N_{ijk}$.
$\mathbf{N}_{ij+}$	The two-way table containing a margin of $\mathbf{N}$, showing the total observations for each category i of variable X and each category j of variable Y.
$\hat{\mathbf{N}}_{ijk}$ or $\hat{\mathbf{N}}$	The IPF estimate of the target multiway table $\mathbf{N}$, using the initial association pattern in source table $\mathbf{n}$ and adjusting it to exactly fit a selected set of margins $\mathbf{N}_{i++}$ (etc.)
$\hat{N}_{ijk}$	A single cell in the IPF estimate $\hat{\mathbf{N}}$.
$\boldsymbol{\pi}_{ijk}$ (or π_{ijk})	Table (or cell) of probabilities instead of counts, $E[n_{ijk}] = n\pi_{ijk}$
$\boldsymbol{\Pi}_{ijk}$ (or Π_{ijk})	Table (or cell) of probabilities instead of counts, $E[N_{ijk}] = N\Pi_{ijk}$
X or $X(i)$	A variable split into I categories, making up the first dimension of each multiway contingency table.
Y or $Y(j)$	A variable split into J categories.
Z or $Z(k)$	A variable split into K categories. In most cases here, Z will specifically refer to geographic *zones*.

Table 2.1: Summary of the notation used for multiway tables and IPF. IPF is used to estimate a multiway contingency table for an unknown target population, by modifying a table of a source sample to match known margins of the target population. The notation shown is for three variables X, Y, Z, but more can be used.

Input: Source table $\mathbf{n}_{ij}$, target margins $\mathbf{N}_{i+}$ and $\mathbf{N}_{+j}$, and tolerance ϵ
Output: Fitted table $\hat{\mathbf{N}}_{ij}^{(\tau)}$

$\tau = 0$;
$\hat{N}_{ij}^{(\tau)} = n_{ij}$;
repeat
  forall *i,j* **do**
    $\hat{N}_{ij}^{(\tau+1)} = \hat{N}_{ij}^{(\tau)} \left(N_{i+} / \hat{N}_{i+}^{(\tau)} \right)$;
  end
  forall *i,j* **do**
    $\hat{N}_{ij}^{(\tau+2)} = \hat{N}_{ij}^{(\tau+1)} \left(N_{+j} / \hat{N}_{+j}^{(\tau+1)} \right)$;
  end
  $\delta = \max \left(\max_i \left| \hat{N}_{i+}^{(\tau+2)} - N_{i+} \right| , \max_j \left| \hat{N}_{+j}^{(\tau+2)} - N_{+j} \right| \right)$;
  $\tau = \tau + 2$
until $\delta < \epsilon$;

Figure 2.4: A simple example algorithm for the Iterative Proportional Fitting procedure using a two-way table and one-way target margins.

valued.

2.2.2 History and Properties of Iterative Proportional Fitting

The Iterative Proportional Fitting (IPF) algorithm is generally attributed to Deming & Stephan [16]. (According to [14], it was preceded by a 1937 German publication applying the method to the telephone industry.) The method goes by many names, depending on the field and the context. Statisticians apply it to contingency tables and use the terms *table standardization* or *raking*. Transportation researchers use it for trip distribution and gravity models, and sometimes reference early papers in that field by Fratar or Furness [15, 23]. Economists apply it to Input-Output models and call the method RAS [32].

The IPF algorithm is a method for adjusting a source contingency table to match known marginal totals for some target population. Figure 2.4 shows a simple appli-

cation of IPF in two dimensions. The table $\hat{\mathbf{N}}^{(\tau)}$ computed in iteration $\tau + 1$ fits the row totals exactly, with some error in the column totals. In iteration $\tau + 2$, an exact fit to the column margins is achieved, but with some loss of fit to the row totals. Successive iteration yields a fit to both target margins within some tolerance ϵ. The procedure extends in a straightforward manner to higher dimensions, and also with higher-dimensional margins.

Deming and Stephan [16] initially proposed the method to account for variations in sampling accuracy. They imagined that the source table and target marginals were measured on the same population, and that the marginal totals were known exactly, but the source table had been measured through a sampling process with some inaccuracy. The IPF method would then adjust the sample-derived cells to match the more accurate marginal totals. They framed this as a fairly general problem with a d-way contingency table, and considered both one-way margins and higher-order margins (up to $d - 1$ ways). They did not consider the effect of zero values in either the initial cell values or the margins.

Deming and Stephan claimed that the IPF algorithm produces a unique solution that meets two criteria. It exactly satisfies the marginal constraints

$$\sum_j \hat{N}_{ij} = N_{i+}, \quad \sum_i \hat{N}_{ij} = N_{+j} \tag{2.1}$$

and they believed that it minimized the weighted least-squares criterion

$$\sum_i \sum_j \frac{(n_{ij}/n - \hat{N}_{ij}/N)^2}{n_{ij}} \tag{2.2}$$

In a later paper, Stephan realized that IPF only approximately minimized that criterion [50]. He proposed a different algorithm that minimized the least-squares criterion. However, Ireland and Kullback [30] returned to the original IPF algorithm and found that it had interesting properties. They showed that the $\hat{N}_{ij}$ estimated by the IPF method minimizes the *discrimination information* criterion (also known as the *Kullback-Leibler divergence*, or *relative entropy*) [33, 14]. This is conventionally defined in terms of

probabilities π_{ij} and $\hat{\Pi}_{ij}$,

$$I(\hat{\mathbf{\Pi}}\|\boldsymbol{\pi}) = \sum_i \sum_j \hat{\Pi}_{ij} \log(\hat{\Pi}_{ij}/\pi_{ij}) \tag{2.3}$$

For the sake of discussion, it can be translated to counts by substituting $\hat{N}_{ij} = N\hat{\Pi}_{ij}$ and $n_{ij} = n\pi_{ij}$

$$\begin{aligned} I(\hat{\mathbf{N}}\|\mathbf{n}) &= I(\hat{\mathbf{\Pi}}\|\boldsymbol{\pi}) \\ &= \log(n/N) + \frac{1}{N} \sum_i \sum_j \hat{N}_{ij} \log(\hat{N}_{ij}/n_{ij}) \qquad (2.4) \\ &= \frac{1}{N} \left(-N \log(N/n) + \sum_i \sum_j \hat{N}_{ij} \log(\hat{N}_{ij}/n_{ij}) \right) \qquad (2.5) \end{aligned}$$

For constant target population size N, this is equivalent to minimizing

$$\sum_i \sum_j \hat{N}_{ij} \log(\hat{N}_{ij}/n_{ij}) \tag{2.6}$$

Note that discrimination information is not symmetric, since $I(\hat{\mathbf{N}}\|\mathbf{n}) \neq I(\mathbf{n}\|\hat{\mathbf{N}})$ in general.

Ireland and Kullback included a proof of convergence. It omitted one step, and was corrected by Csiszár in a 1975 paper [13]. Csiszár's treatment was somewhat more general than previous papers. In particular, he adopted a convention for the treatment of zero values in the initial cells:

$$\log 0 = -\infty, \quad \log \frac{a}{0} = +\infty, \quad 0 \cdot \pm\infty = 0 \tag{2.7}$$

After adopting this convention, he proved convergence with allowance for zeros.

The IPF method is one of several ways of satisfying the marginal constraints (2.1) while minimizing entropy relative to the source table. Alternative algorithms exist for solving this system of equations, including Newton's method. Newton's method offers the advantage of a faster (quadratic) convergence rate, and is also able to estimate the parameters and variance-covariance matrix associated with the system (to be

discussed in the following section). However, Newton's method is considerably less efficient in computational storage and is impractical for the large systems of equations that occur in high-dimensional problems. Using the asymptotic Landau $\mathcal{O}()$ notation conventional in computer science [11], the IPF method requires $\mathcal{O}(C)$ memory to fit a contingency table with C cells, while Newton's method requires $\mathcal{O}(C^2)$ storage [1, chapter 8], [20].

Additionally, the minimum discrimination information of equation (2.6) is not the only possible optimization criterion for comparing the fitted table to the source table. Little & Wu [33] looked at a special case where the source sample and the target margins are drawn from different populations. In their analysis, they compared the performance of four different criteria: minimum discrimination information, minimum least squares, maximum log likelihood and minimum chi-squared. For certain problems, other optimization criteria may offer some advantages over minimum discrimination information.

In summary, the Iterative Proportional Fitting method is a data fusion technique for combining the information from a source multiway contingency table and lower-dimensional marginal tables for a target population. It provides an exact fit to the marginal tables, while minimizing discrimination information relative to the source table.

2.2.3 Generalizations of the IPF Method

Following the basic understanding of the IPF method in the late 1960s and early 1970s, the method received further attention in the statistical and information theory community. As discussed earlier, Csiszár's 1975 paper [13] was in part a correction and generalization of Ireland & Kullback's work. However, it also introduced a different conception of the underlying problem. Csiszár did not represent the multiway probability distribution as a d-way contingency table with C cells. Instead, he conceived of

I-space, a much larger C-dimensional space containing all possible probability distributions for the C cells in the contingency table. He exploited the geometric properties of this space to prove convergence of the algorithm.

The mechanics of his proof and construction of I-space would be a theoretical footnote, except that further extensions and generalizations of the IPF method have been made using the I-space notation and conceptualization. In I-space, the marginal constraints form closed linear sets. The IPF algorithm is described as a series of I-projections onto these linear constraints.

From a cursory reading of a 1985 paper by Dykstra [17], it appears that he generalized Csiszár's theory and proved convergence of the IPF method for a broader class of constraints: namely any closed, convex set in I-space. Dykstra used a complicated example where the cells are not constrained to equal some marginal constraint, but the tables' marginal total vectors were required to satisfy an ordering constraint—for example, requiring that $n_{i2} < n_{i3}$. Dykstra's iterative procedure was broadly the same as the IPF procedure: an iterative projection onto the individual convex constraints. In other words, small extensions to the IPF method can allow it to satisfy a broader class of constraints beyond simple equality.

Further generalizations of the IPF method are discussed by Fienberg & Meyer [20].

2.2.4 Log-Linear Models

Log-linear models provide a means of statistically analyzing patterns of association in a single contingency table. They are commonly used to test for relationships between different variables when data is available in the form of a simple, low-dimensional contingency table. The method itself derives from work on categorical data and contingency tables in the 1960s that culminated in a series of papers by Goodman in the early 1970s [19]. The theory behind log-linear models is well-established and is described in detail elsewhere [54, 1, 37]. In this section, a few examples are used to

provide some simple intuition for their application. The notation used here follows [54], but is fairly similar to most sources.

Consider a single two-dimensional contingency table $\mathbf{n}_{ij}$ containing observations of variables X and Y. The general form of a log-linear model for such a table is

$$\log n_{ij} = \text{constant} + \text{row term} + \text{column term} + \text{association terms} \tag{2.8}$$

The log-linear name comes from this form: the logarithm of the individual cells' counts is the dependent variable (left hand side), and this variable is modelled as a linear sum of the parameters (right hand side). A concrete example is the model of independence,

$$\log n_{ij} = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} \tag{2.9}$$

The subscripts here make clear the idea of a "row term": for a given row i of table $\mathbf{n}$, each of the cells in that row of $\mathbf{n}$ share a single parameter $\lambda_{X(i)}$; a similar effect can be seen for the columns. This is a model of independence, in that it presumes that the counts can be explained without including any association between variables X and Y. The alternative model including association is

$$\log n_{ij} = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)} \tag{2.10}$$

This log-linear model can be used to test for statistically significant association between X and Y in the observations in a given table $\mathbf{n}_{ij}$. The null hypothesis H_0 is that the variables are independent. The hypotheses are defined as:

$$\begin{aligned} H_0 &: \lambda_{XY(ij)} = 0 \quad \text{for all } i, j \\ H_1 &: \lambda_{XY(ij)} \neq 0 \quad \text{for some } i, j \end{aligned} \tag{2.11}$$

If none of the association parameters are statistically different from zero, then the null hypothesis cannot be rejected. If one or more association parameters are statistically different from zero, then this supports the alternative hypothesis that some association

exists. This is a typical application of a log-linear model: to test for the existence of association between variables in a contingency table. The association terms here are reminiscent of *interaction* in Analysis of Variance (ANOVA) although there are important differences. Wickens suggested that a two-way log-linear model is more similar to one-way ANOVA than to two-way ANOVA [54, §3.10].

As more variables are added, higher-order association patterns such as XYZ can be included, and the number of possible models grows. In practise, only *hierarchical* models are used, where an association term XY is only included when lower-order terms X and Y are also included. Hierarchical models are usually summarized using only their highest-order terms. For example, the model (XY, Z) implicitly includes X and Y terms. For a given set of variables, the model that includes all possible association terms is called the *saturated* model.

To ensure uniqueness, constraints are usually applied to the parameters of a log-linear model. Two different conventions are common, and tools for estimating log-linear models may use either convention. The ANOVA-type coding or effect coding using the following constraints for a two-way table [37]:

$$\sum_i \lambda_{X(i)} = \sum_j \lambda_{Y(j)} = \sum_i \lambda_{XY(ij)} = \sum_j \lambda_{XY(ij)} = 0 \tag{2.12}$$

while the dummy-variable coding blocks out one category for each:

$$\lambda_{X(1)} = \lambda_{Y(1)} = \lambda_{XY(1j)} = \lambda_{XY(i1)} = 0 \tag{2.13}$$

Provided that all cells are non-zero, the breakdown of parameters (and hence degrees of freedom) in a log-linear model is quite simple. For example, a saturated model of a two-way table consists of one constant parameter, $I-1$ row parameters, $J-1$ column parameters and $(I-1)(J-1)$ association parameters for a total of IJ parameters. The presence of zeros can complicate the parameter counting substantially, however.

After estimating the parameters of a log-linear model based on observed counts

n_{ij}, the estimated counts $\hat{n}_{ij}$ are obtained. The model fits can be tested on these estimates using either the Pearson statistic

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{2.14}$$

or the likelihood-ratio statistic

$$G^2 = 2 \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}} \tag{2.15}$$

Both of these are χ^2-distributed, and hierarchically related models can be compared in terms of fit provided that the number of degrees of freedom in the models are known. The G^2 statistic is clearly related to the discrimination information of equation (2.3). After noting that $n = \sum_i \sum_j n_{ij} = \sum_i \sum_j \hat{n}_{ij}$, it is clear that $G^2 = 2nI(\mathbf{n}\|\hat{\mathbf{n}})$. This formula is sometimes also known as the minimum discrimination information (MDI) statistic [31].

Finally, the G^2 statistic for the null model ($\log n_{ij} = \lambda$) is related to the *entropy* of a probability distribution. The formula for entropy is

$$H(\boldsymbol{\pi}_{ij}) = -\sum_{i,j} \pi_{ij} \log \pi_{ij} \tag{2.16}$$

and can be translated to a table of counts as

$$H(\mathbf{n}_{ij}) = -\frac{1}{n} \left(-n \log n + \sum_{i,j} n_{ij} \log n_{ij} \right) \tag{2.17}$$

The fitted null model is a uniform probability distribution, $n_{ij} = n/IJ$. Its G^2 statistic can be shown to equal $2n(\log IJ - H(\mathbf{n}_{ij}))$.

2.2.5 IPF and Log-Linear Models

The Iterative Proportional Fitting procedure has long been associated with log-linear analysis. Given a log-linear model and a contingency table $\mathbf{N}_{ijk}$, it is often useful to know what the fitted table of frequencies $\hat{\mathbf{N}}_{ijk}$ would be under the model. (For

intuitive purposes, this is the equivalent of finding the shape of the fitted line under the model of linear regression.)

In this problem, a log-linear model is called *direct* if the fitted table can be expressed as closed formulae, or *indirect* if the fitting can only be achieved using an iterative algorithm. For indirect models such as (XY, XZ, YZ), IPF has long been employed as an efficient way to fit a table to the data. To achieve this, the source table $\mathbf{n}_{ijk}$ is chosen to have no pattern of association whatsoever; typically, this is done by setting $n_{ijk} = 1$. The target margins used by the IPF procedure are the "minimal sufficient statistics" of the log-linear model; for example, to fit the model (XY, Z), the marginal tables $\mathbf{N}_{ij+}$ and $\mathbf{N}_{++k}$ would be applied. The resulting table found by IPF is known to give the maximum likelihood fit.

Each step of the IPF procedure adjusts all cells contributing to a given marginal cell equally. As a result, it does not introduce any new patterns of association that were not present in the source table; and the source table was chosen to include no association whatsoever. The resulting table shows only the modelled patterns of association [54, §5.2].

The IPF procedure is hence an important tool for log-linear modelling. Additionally, log-linear models provide some useful insight into the behaviour of the IPF procedure. Stephan showed that the relationship between the fitted table $\hat{\mathbf{N}}_{ijk}$ and the source table $\mathbf{n}_{ijk}$ could be expressed as

$$\log \hat{N}_{ijk}/n_{ijk} = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} \tag{2.18}$$

when fitting to the $\mathbf{N}_{i++}$, $\mathbf{N}_{+j+}$ and $\mathbf{N}_{++k}$ margins [50, 33, 1] for some choice of λ parameters. This has the exact same form as a log-linear model, but it is not a model; rather, with a suitable choice for the λ parameters, the formula holds exactly for every cell. In other words, this model is sufficient to explain *all* of the variation between $\mathbf{n}$ and the IPF estimate $\hat{\mathbf{N}}$. A similar model can be constructed for any set of margins applied during the IPF procedure, by adding λ terms that correspond to the variables

in the margins.

This view of the IPF procedure is mostly useful for interpreting its behaviour. IPF modifies the source table $\mathbf{n}_{ijk}$ to create a fitted table $\hat{\mathbf{N}}_{ijk}$; that change, represented by the left hand side of equation (2.18), can be expressed using a small number of parameters (the various λ terms). The number of parameters necessary is directly proportional to the size of the marginal tables used in the fitting procedure; in this case, $1 + (I - 1) + (J - 1) + (K - 1)$. This insight is not unique to log-linear models, but it is perhaps easier to understand than the Lagrangian analysis used in early IPF papers.

2.2.6 Zero Cells

The only shortcoming of the preceding discussion of IPF and log-linear models concerns zeros in the source table or target margins. While Csiszár's treatment of zeros allows the IPF procedure to handle zeros elegantly, it remains difficult to determine the correct number of *parameters* used by IPF when either the source or target tables contain zeros.

The zeros can take the form of either *structural* or *sampling zeros*. Structural zeros occur when particular combinations of categories are impossible. For example, if a sample of women is cross-classified by age and by number of children, the cell corresponding to "age 0–10" and "1 child" would be a structural zero (as would all higher number of children for this age group). A sampling zero occurs when there is no *a priori* reason that a particular combination of categories would not occur, but the sample was small enough that no observations were recorded for a particular cell.

Wickens provides a detailed description of the consequences of zero cells for log-linear modelling [54, §5.6]. For high dimensional tables, the number of parameters in a particular model becomes difficult to compute, and this in turn makes it difficult to determine how many degrees of freedom are present. As he notes, however, "The

degrees of freedom are off only for global tests of goodness of fit and for tests of the highest-order interactions." Clogg and Eliason suggested that goodness-of-fit tests are futile when the data becomes truly sparse:

> But there is a sense in which goodness-of-fit is the wrong question to ask when sparse data is analyzed. It is simply unreasonable to expect to be able to test a model where there are many degrees of freedom relative to the sample size. [10]

For a small number of zeros, then, it seems that some log-linear analysis may be possible. A sparse table with a large number of zeros, by contrast, is unlikely to be tested for goodness-of-fit.

2.3 Population Synthesis

Microsimulation and agent-based methods of systems modelling forecast the future state of some aggregate system by simulating the behaviour of a number of disaggregate *agents* over time [9].

In many agent-based models where the agent is a person, family or household the primary source of data for population synthesis is a national census. In many countries, the census provides two types of data about these agents. Large-sample detailed tables of one or two variables across many small geographic areas are the traditional form of census delivery, and are known as Summary Files in the U.S., Profile Tables or Basic Summary Tabulations (BSTs) in Canada, and Small Area Statistics in the U.K. In addition, a small sample of individual census records is now commonly available in most countries. These samples consist of a list of individuals (or families or households) drawn from some large geographic area, and are called Public-Use Microdata Samples (PUMS) in the U.S. and Canada, or a Sample of Anonymized Records in the U.K. The geographic area associated with a PUMS is the Public-Use Microdata Area

(PUMA) in the U.S., and the Census Metropolitan Area (CMA) in Canada. The population synthesis procedure can use either or both of these data sources, and must produce a list of agents and their attributes, preferably at a relatively fine level of geographic detail.

In this document, the terms Summary Tables, PUMS and PUMA will be used. For small areas inside a PUMA, the Census Tract (CT) will often be used (or more generically a "zone"), but any fine geographic unit that subdivides the PUMA could be used. Further details about the data and definitions are presented in Chapter 3.

In most population synthesis procedures, geography receives special attention. This is not because geography is inherently different than any other agent attribute: like the other attributes, location is treated as a categorical variable, with a fine categorization system (small zones like census tracts) that can be collapsed to a coarse set of categories (larger zones, like the Canadian Census Subdivisions), or even collapsed completely (to the full PUMA) to remove any geographic variation. There are two reasons why geography receives special attention: first, because census data is structured to treat geography specially. One data set (the PUMS) provides data on the association between almost all attributes except geography; the other (Summary Tables) includes geography in every table, and gives its association with one or two other variables. Secondly, geography is one of the most important variables for analysis and often has a large number of categories; while an attribute like age can often be reduced to 15–20 categories, reducing geography to such a small number of categories would lose a substantial amount of variation that is not captured in other attributes. For transportation analysis, a fine geographic zone system is essential for obtaining a reasonable representation of travel distances, access to transit systems, and accurate travel demand. As a result, geography is usually broken up into hundreds of zones, sometimes more than a thousand.

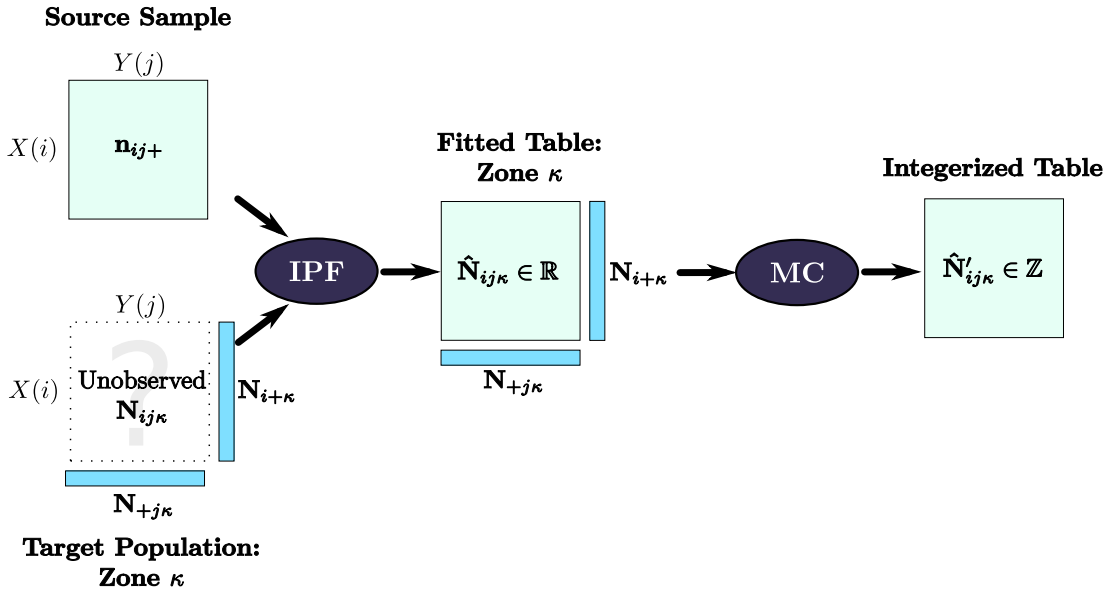

Figure 2.5: A zone-by-zone application of IPF for population synthesis, including a Monte Carlo integerization stage. The source table can either be constant (X independent of Y), or created by cross-classifying a PUMS (shown here). Zone κ is synthesized without consideration of other zones, and under the assumption that its association pattern is the same as the pattern in the PUMS. After integerization, the table no longer exactly fits the margins.

2.3.1 Zone-by-Zone IPF

The Iterative Proportional Fitting method is the most popular means for synthesizing a population. The simplest approach is to consider each small geographic zone independently. Suppose that the geographic zones are census tracts contained in some PUMA. Further, suppose that each agent needs two attributes $X(i)$ and $Y(j)$, in addition to a zone identifier $Z(k)$. An overview of the process is shown in Figure 2.5.

The synthesis is conducted one zone at a time, and the symbol κ denotes the zone of interest. In the simplest approach, the variables X and Y are assumed to have no association pattern (i.e., they are assumed to vary independently), and hence the initial table $\mathbf{n}_{ij\kappa}$ is set to a constant value of one. The summary tables provide the known information about zone κ: the number of individuals in each category of variable X can be tabulated to form $\mathbf{N}_{i+\kappa}$, and likewise with variable Y to give $\mathbf{N}_{+j\kappa}$. These are

used as the target margins for the IPF procedure, giving a population estimate for zone κ.

However, the variables X and Y are unlikely to be independent. The PUMS provides information about the association between the variables, but for a different population: a small sample in the geographically larger PUMA. As discussed by Beckman, Baggerly & McKay [4], under the assumption that the association between X and Y in zone κ is the same as the association in the PUMA, the initial table $\mathbf{n}_{ij\kappa}$ can be set to a cross-classification of the PUMS over variables $X(i)$ and $Y(j)$. IPF is then applied, yielding a different result.

The IPF process produces a multiway contingency table for zone κ, where each cell contains a real-valued "count" $\hat{N}_{ij\kappa}$ of the number of agents with a particular set of attributes $X = i$ and $Y = j$. However, to define a discrete set of agents integer counts are required. Rounding the counts is not a satisfactory "integerization" procedure for three reasons: the rounded table may not be the best solution in terms of discrimination information; the rounded table may not offer as good a fit to margins as other integerization procedures; and rounding may bias the estimates.

Beckman et al. handled this problem by treating the fitted table as a joint probability mass function (PMF), and then used N Monte Carlo draws [24] from this PMF to select N individual cells. These draws can be tabulated to give an integerized approximation $\hat{\mathbf{N}}'$ of $\hat{\mathbf{N}}$. This is an effective way to avoid biasing problems, but at the expense of introducing a nondeterministic step into the synthesis.

Finally, given an integer table of counts, individual agents can be synthesized using lookups from the original PUMS list. (See Figure 2.2 for an illustration of the link between the list and tabular representations.) Beckman et al. observed an important aspect of this process: if n_{ij} is zero (i.e., no records in the PUMS for a particular combination of variables), then the fitted count $\hat{N}_{ij\kappa}$ will be zero, and this carries through to the integerized count $\hat{N}'_{ij\kappa}$. Consequently, any cell in $\hat{\mathbf{N}}'$ that has a non-zero count

is guaranteed to have corresponding individual(s) in the PUMS.

2.3.2 Multizone IPF

Beckman, Baggerly & McKay [4] discussed the simple zone-by-zone technique and also extended it to define a multizone estimation procedure; their approach has been widely cited [5, 21, 2] and is described in great detail by [27]. They described a method for using IPF with a PUMS to synthesize a set of households. Their approach addresses a weakness of the zone-by-zone method: the PUMS describes the association pattern for a large geographic area, and the pattern within small zones inside that area may not be identical. Consequently, Beckman et al. made no assumptions about the association pattern of the individual zones, but instead required the sum over all zones to match the PUMS association pattern. This approach is illustrated graphically in Figure 2.6; their paper includes a more detailed numerical example. (The Monte Carlo integerization step is omitted for clarity.)

This multizone approach offers an important advantage over the zone-by-zone approach. The zone-by-zone approach uses the PUMS association pattern for the initial table, but it is overruled by the marginal constraints, and its influence on the final result is limited. By applying the PUMS association pattern as a *marginal constraint* on the IPF procedure, the multizone method guarantees that the known association pattern in the PUMS is satisfied exactly.

2.3.3 Synthesis Examples Using IPF

Many projects have applied Beckman et al.'s methods. Most microsimulation projects seem to use a zone-by-zone fitting procedure with a PUMS, followed by Monte Carlo draws as described by Beckman. Few seem to have adopted Beckman's multizone fitting procedure, however. This may be due to storage limitations: synthesizing all

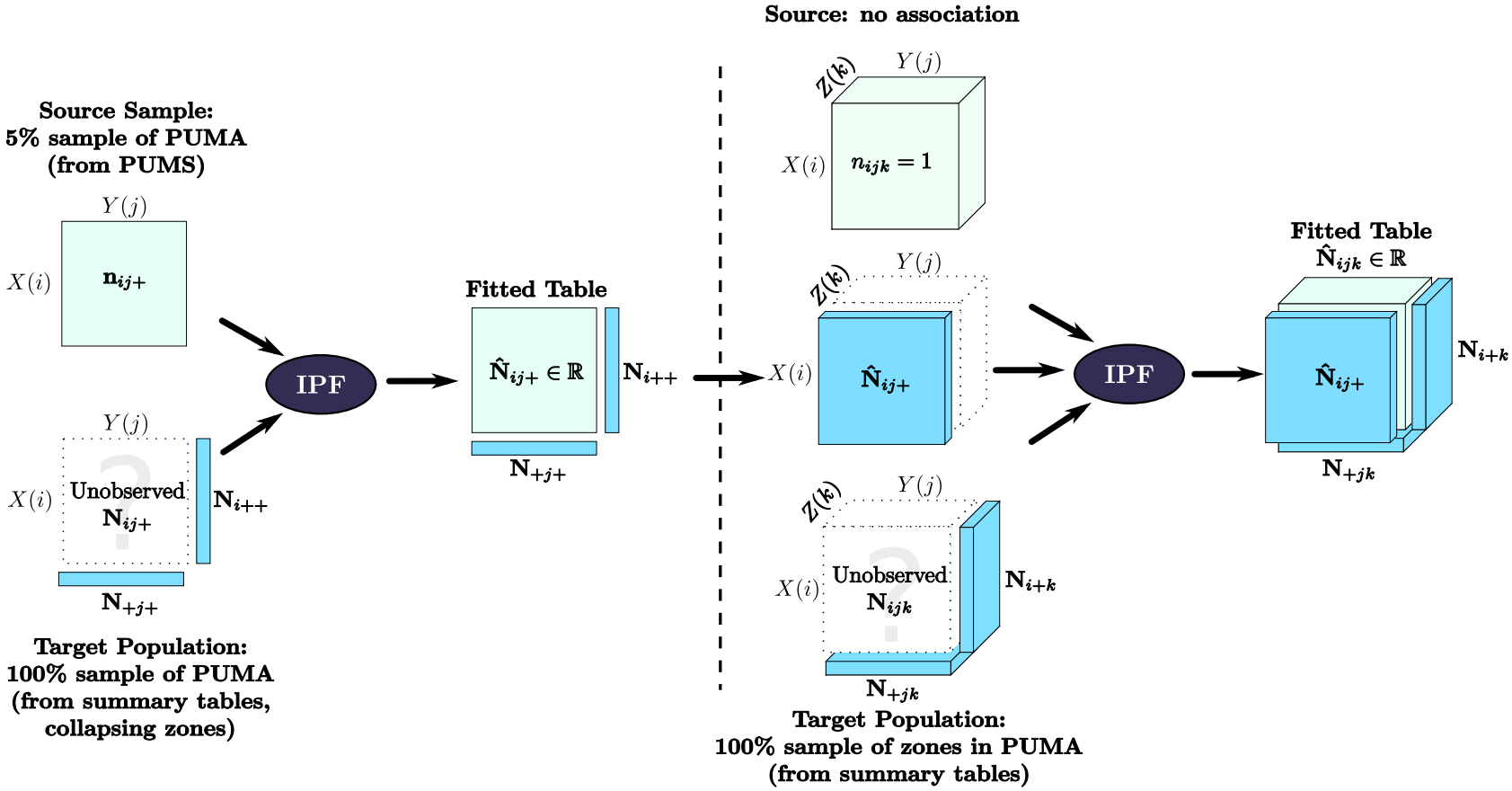

Figure 2.6: An illustration of Beckman et al.'s fitting procedure using two attributes X and Y, plus a variable Z representing the census tract zone within the PUMA. In the left half, Z is ignored and the PUMS is adjusted to match the summary tables; they differ because the PUMS is derived from a smaller sample than the summary tables. In the right half, the variable $Z(k)$ is added to represent the K zones that make up the PUMA. A constant initial table filled with ones is used for a second IPF, which is fitted to the summary tables and the adjusted PUMS. The summary tables now show variation of X by zone Z (and likewise $Y \times Z$), while the adjusted PUMS provides information about the association between X and Y.

zones simultaneously using Beckman's multizone method requires substantially more computer memory, and would consequently limit the number of other attributes that could be synthesized.

In terms of different agent types, Beckman et al. considered households, families and individuals in a single unit. They segmented households into single-family, non-family and group quarters. They then synthesized family households, including a few individual characteristics (age of one family member) and "presence-of-children" variables. They did not synthesize person agents explicitly, did not associate families with dwellings, and did not synthesize the connection between families in multifamily households. Later work on their model (TRANSIMS) did synthesize persons from the households [27]. Their largest table was for family households with $d = 6$ variables and $C = 11,760$ cells before including geography, of which 609 cells were nonzero. Their sample was a 5% sample of a population of roughly 100,000 individuals.

Guo & Bhat [26] applied Beckman's procedure to a population of households in the Dallas-Fort Worth area in Texas. (It is not clear whether they used zone-by-zone synthesis or applied Beckman's multizone approach.) They modified Beckman's integerization procedure by making Monte Carlo draws without replacement, with some flexibility built into the replacement procedure in the form of a user-defined threshold. Further, they proposed a procedure for simultaneously synthesizing households and individuals. In their procedure, the household synthesis includes some information about the individuals within the household: the number of persons and the family structure. A series of individuals are synthesized to attach to the household, using the limited known information from the synthesized household. If any of the synthetic individuals are "improbable" given the number of demographically similar individuals already synthesized, then the entire household is rejected and resynthesized.

The linkage between households and individuals in this method remains weak, and the procedure is fairly *ad hoc*. Guo & Bhat showed that the method did improve

fit in the individual table fits during a single trial, but the results were not adequate for any conclusive claims.

Huang & Williamson [28] implemented an IPF-based method for comparison against the Combinatorial Optimization method. They used a novel zone-by-zone synthesis procedure. In their approach, large zones ("wards") within a PUMA were first synthesized one-at-a-time, under the assumption that each ward has the same association pattern as the larger PUMA. The ward results were then used as the initial table for the synthesis of finer zones (enumeration districts) within each ward. This multilevel approach improves on the conventional zone-by-zone method.

Huang & Williamson also used an incremental approach to attribute synthesis which they call "synthetic reconstruction" where after an initial fitting to produce four attributes, additional attributes are added one-at-a-time. The motivation for this approach is apparently to avoid large sparse tables, and includes collapsing variables to coarser categorization schemes to reduce storage requirements and sparsity. However, their method is complex and requires substantial judgment to construct a viable sequencing of new attributes, by leveraging a series of conditional probabilities. Furthermore, the connection to the PUMS agents is lost: the resulting population is truly synthetic, with some new agents created that do not exist in the initial PUMS. Since only a subset of the attributes are considered at a time, it is possible that some attributes in the synthetic agents may be contradictory. Nevertheless, the analysis is interesting, and reveals the effort sometimes expended when trying to apply IPF rigorously to obtain a population with a large number of attributes.

Finally, Huang & Williamson proposed a modification to the Monte Carlo procedure, by separating the integer and fractional components of each cell in the multiway table. The integer part is used directly to synthesize discrete agents, and the table of remaining fractional counts is then used for Monte Carlo synthesis of the final agents.

2.4 Reweighting and Combinatorial Optimization

The primary alternative to the Iterative Proportional Fitting algorithm is the reweighting approach advocated by Williamson. In a 1998 paper [57], Williamson et al. proposed a zone-by-zone method with a different parameterization of the problem: instead of using a contingency table of real-valued counts, they chose a list representation with an integer weight for each row in the PUMS. (As illustrated in Figure 2.2, there is a direct link between tabular and list-based representations.)

For each small zone κ within a PUMA, the zone population is much smaller than the observations in the PUMS; that is, $N_{++\kappa} < n$. This made it possible for Williamson et al. to to choose a subset of the PUMS to represent the population of the zone, with no duplication of PUMS agents in the zone. In other words, the weight attached to each agent is either zero or one (for a single zone).

To estimate the weights, they used various optimization procedures to find the set of 0/1 weights yielding the best fit to the Summary Tables for a single zone. They considered several different measures of fit, and compared different optimization procedures including hill-climbing, simulated annealing and genetic algorithms. By solving directly for integer weights, Williamson obtained a better fit to the Summary Tables than Beckman et al., whose Monte Carlo integerization step harmed the fit.

Williamson et al. [57] proposed three primary reasons motivating their approach:

1. Efficiency: a list-based representation uses considerably less storage than a tabular representation, particularly for a large number of attributes.

2. Flexible aggregation: due to their storage limitations, array-based approaches often collapse finely-categorized attributes to a coarse categorization scheme. The list-based representation allows fine categorizations which can be flexibly aggregated into simple schemes as required. This can be done during the fitting procedure (to align with the categorization of a constraint), or after synthesis.

3. Linkage: after synthesis, the list of individuals can be expanded to include new attributes from other data sources easily; cross-tabulations are more difficult to disaggregate in a similar manner.

The last two claims are somewhat weak. It is true that many IPF procedure require coarse categorization during fitting, in order to conserve limited memory. However, after completion, Beckman's approach does produce a list of PUMS records (and can be linked to other data sources easily). Even if a coarse categorization was used during fitting, it is still possible to use the fine categorization in the PUMS after synthesis. Nevertheless, IPF does require a carefully constructed category system to make fitting possible, and this can be time-consuming to design and implement.

The reweighting approach has three primary weaknesses. First, the attribute association observed in the PUMS ($\mathbf{n}_{ij}$) is not preserved by the algorithm. The IPF method has an explicit formula defining the relationship between the fitted table $\hat{\mathbf{N}}_{ij}$ and the PUMS table $\mathbf{n}_{ij}$ in equation (2.18). Beckman et al.'s multizone approach also treats the PUMS association pattern for the entire PUMA as a constraint, and ensures that the full population matches that association pattern. The reweighting method does operate on the PUMS, and an initial random choice of weights will match the association pattern of the PUMS. However, the reweighting procedure does not make any effort to preserve that association pattern. While the reweighting method has been evaluated in many ways [52, 28, 56, 38], it does not appear that the fit to the PUMS at the PUMA level has been tested.

Secondly, the reweighting method is very computationally expensive. When solving for a single zone κ, there are n 0/1 weights, one for each PUMS entry. However, this gives rise to $\binom{n}{N_{++\kappa}}$ possible combinations; "incredibly large," in the authors' words [57]. Of course, the optimization procedures are intelligent enough to explore this space selectively and avoid an exhaustive search; nevertheless, the authors reported a runtime of 33 hours using an 800MHz processor [28]. Since the number of

permutations grows factorially with the number of individuals in the zone, it is not surprising that the authors chose to work with the smallest zones possible (1440 zones containing an average of 150 households each); it is possible that larger zones would not be feasible.

Finally, the reweighting method uses $n \times K$ weights to represent a K-zone area. This parameter space is quite large; larger, in fact, than the population itself. It is not surprising that good fits can be achieved with a large number of parameters, but the method is not particularly parsimonious and may *overfit* the Summary Tables. It is likely that a simpler model with fewer parameters could achieve as good a fit, and would *generalize* better from the 2% PUMS sample to the full population.

Chapter 3

Data Sources and Definitions

The data for this project came largely from the Canadian Census administered by Statistics Canada. The census has been conducted every five years since 1981, and Toronto's travel activity survey (the Transportation Tomorrow Survey or TTS) is timed to coincide with census years. The TTS was first conducted in 1986, and this was therefore chosen as the baseline year for the ILUTE model and population synthesis.

The Canadian Census has been conducted as a mail-back self-administered survey since 1971. Eighty percent of households receive a short survey known as the 2A form, while twenty percent receive an expanded version called the 2B form. In 1986, the census was conducted on the first Tuesday of June, which fell on the third day of the month. The 1986 Census was Canada's first full mid-decade census, and was very nearly cancelled due to reduced federal government expenditure in the early 1980s. It was reinstated, but with limited resources. As a result, some useful information was never fully coded or tabulated, such as the place-of-work. However, the provincial government in Ontario did pay for geocoding place-of-work for the entire province, and some tables with geographic distributions of employment do exist, although they can be difficult to obtain [47].

Census data is aggregated by persons, census families, or households and is re-

Figure 3.1: The major groups within the Canadian census' person universe. The numbers in parentheses show the size of each grouping (thousands of persons) within the Toronto Census Metropolitan Area (CMA) in the 1986 census. Adapted from [42].

leased in three distinct forms. Profile tables are assembled for each question from the census, showing the breakdown of responses to a single question within a geographic area. Basic Summary Tabulations (BSTs) are cross-tabulations of two to four questions from the census, also including geographic variation. Profile table and cross-tabulations may be derived from questions from the 2A or 2B forms, and may represent either a 100% sample or a 20% sample that has been expanded to a 100% basis. Finally, Statistics Canada also releases Public Use Microdata (PUMS), a 2% sample of all responses made by a person (and likewise a 1% sample of family responses and a 1–4% sample of household responses). Each PUMS data file is associated with a single Census Metropolitan Area (CMA), a large geographic area of more than 100,000 persons that acts as the equivalent of the Public Use Micro Area (PUMA) in the U.S.; the data contains no information about spatial variation within the CMA.

The PUMS data and different summary tables may be drawn from different samples. The population of persons can be broken down into many subgroups, some of

Source	Geography	Universe	Sample %	Sample size
Persons				
PUMS	CMA	Non-inst. persons	2%	67,992
BST DM86A01	CMA	All persons	100%	3,427,165
BST SC86B01	CMA	Non-inst. persons, age 15+	20%	546,470
BST LF86B04	CMA	Labour force	20%	395,965
BST DM86A01	CT 59.00	All persons	100%	3,745
BST SC86B01	CT 59.00	Non-inst. persons, age 15+	20%	653
BST LF86B04	CT 59.00	Labour force	20%	482
Census Families				
PUMS	CMA	Families in priv. dwellings	1%	9,061
BST CF86A02	CMA	Families in priv. dwellings	100%	906,385
BST CF86A02	CT 59.00	Families in priv. dwellings	100%	800
Dwellings / Households				
PUMS	CMA	Occupied private dwellings	1%	11,998
BST DW86A01	CMA	Occupied private dwellings	100%	1,119,800
BST DW86B02	CMA	Occupied private dwellings	20%	239,960
BST DW86A01	CT 59.00	Occupied private dwellings	100%	1,130
BST DW86B02	CT 59.00	Occupied private dwellings	20%	226

Table 3.1: Sample sizes of some data sources used for synthesis, at different levels of geography This gives a sense of the sample size in PUMS and Summary Table data, both at a broad geographical scale (the Toronto CMA), and at the finer scale of Census Tracts (CT). The example CT 59.00 is a downtown zone neighbouring the University of Toronto. The BSTs that include an "A" are drawn from the Census 2A form and have a 100% sample, while the "B" tables have a 20% sample.

which are shown in Figure 3.1. The 2A census form (100% sample) is collected for the full population, while the 2B form (20% sample) is collected only for the non-institutional population 15 years of age and over. Some summary tables are defined on the 2A universe, where exact population counts are available. Others are defined on the 2B universe, by expanding the 20% sample to an estimate of the complete 2B universe. Combining data from tables derived from the 2A and 2B samples can be challenging, because of their differing universes and errors in the 2B estimates. The PUMS uses a different sample again; it is defined on a 2% sample of the full population excluding institutional residents (and residents of incompletely enumerated Indian reserves, which are not an issue in the Toronto CMA). The sample sizes associated with different universes and tables are summarized in Table 3.1.

The universe of persons is slightly complicated. The 1986 census excluded *non-permanent residents* from all tables, which includes foreign persons present on student authorization, employment authorization, Minister's permits and refugee claimants. These were included in 1991 and subsequent censuses, and do account for a sizeable fraction of the Toronto population. In 1991, there were 98,105 non-permanent residents in the Toronto CMA (2.5% of total); assuming a similar growth rate to the CMA as a whole, this would give approximately 89,000 in 1986. There is no data on this population, however. *Institutional* collective dwellings are defined as hospitals, orphanages, correctional/penal institutions and religious institutions, and the residents of these institutions are excluded from many tables (but not the staff). *Non-institutional collective dwellings* are defined as hotels, motels, tourist homes, lodging- and rooming-houses, work camps, military camps and Hutterite colonies. *Temporary residents* are persons with a usual dwelling elsewhere in Canada living temporarily in another dwelling; they are usually treated as part of their "permanent" household. However, some dwellings are occupied only by temporary residents, and are a separate category from both occupied and unoccupied dwellings. Finally, *foreign residents*

are foreign diplomats or military personnel stationed in Canada. Temporary, foreign and collective (non-institutional) residents are included in most person-based tables, but not in family, household or dwelling tables.

Statistics Canada makes some modifications to the collected census data before publishing tables. Contradictions in the submitted form are resolved using an *edit and impute* method. Furthermore, to protect the privacy of individual persons and households, Statistics Canada applies two disclosure control techniques. In any released table, all numbers are *randomly rounded* (up or down) to a multiple of five and in special cases to a multiple of ten. This is a stronger measure than many countries; the UK and New Zealand use a multiple of three, and the American census does not use random rounding [18]. The UK and Australian agencies apply random rounding only to small cells, but the Canadian agency applies it to every cell in every table. In each reported table, the individual cells and the row and column totals are rounded independently using a procedure called Unbiased Random Rounding. The rounding tends toward the closer multiple of five, so a count of 4 has a probability of 80% of being rounded to 5 and a 20% probability of being rounded to 0 [55]. The alternative is called *unrestricted* random rounding, where there is a fixed probability p that a cell is rounded down, regardless of its value; typically, $p = 0.5$ is used.[1]

Finally, in geographic areas with less than forty persons, no data is released; this is called *area suppression*. Additionally, in areas with less than 250 persons, no income data is released.

[1]Statistics Canada is rarely explicit about which rounding technique they use, but Boudreau implies that unbiased random rounding is used [7].

3.1 Family and Household Definitions

The Canadian census family and household definitions are generally intuitive, but some special cases are tricky. As the Census Handbook notes, "it is very difficult to translate complex human relationships into tables" [42].

The census distinguishes between two types of families: the "census family" defines a relationship between cohabiting adults and children, while the "economic family" defines other types of family relationships within a single dwelling. The details of family definition are complicated, particularly when considering cohabiting multigeneration families. The household definition is straightforward, consisting of all persons sharing a "dwelling unit;" there is a one-to-one relationship between households and occupied dwelling units. The dwelling unit definition is slightly more complicated, and is defined as living quarters *with a private entrance* from the outside or from a common hallway. More formally, Figure 3.2 graphically shows the relationship between the different types of family membership.

> "People living in the same dwelling are considered a census family only if they meet the following conditions: they are spouses or common-law partners, with or without never-married sons or daughters at home, or a lone parent with at least one son or daughter who has never been married. The census family includes all blood, step- or adopted sons and daughters who live in the dwelling and have never married. It is possible for two census families to live in the same dwelling; they may or may not be related to each other" [49] for 1996; essentially the same as 1986 definition [42, 45].

No distinction is made between common-law and legal marriage; both are coded as "married." While homosexual couples are recognized to exist, the census coding does not allow this type of family. Any household that reports a married/common-law couple with the same sex is recoded; either they are cohabiting unmarried individ-

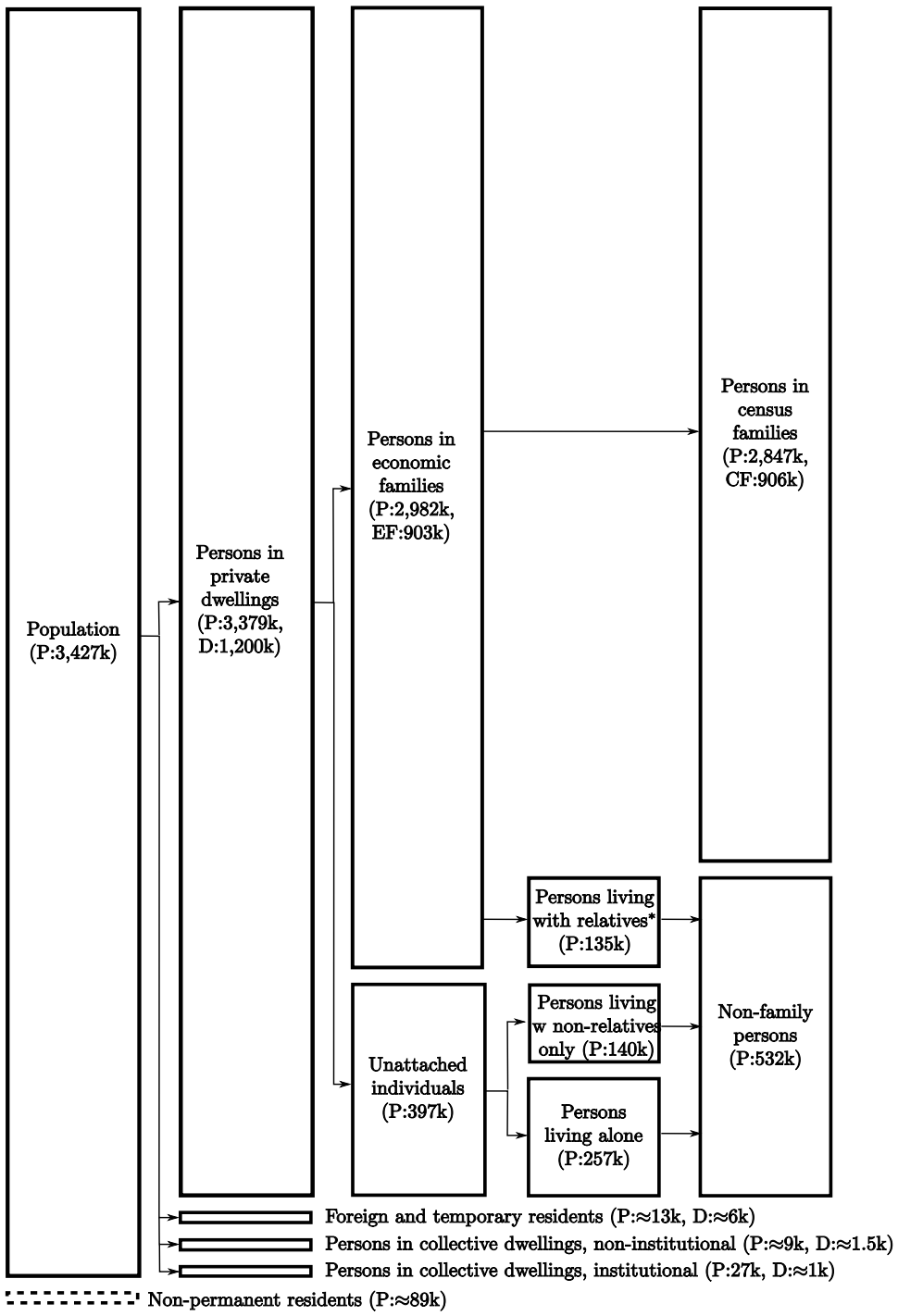

Figure 3.2: A breakdown of the Canadian census' person universe, by family membership. The numbers in parentheses show the size of each grouping in thousands, aggregated into groupings of persons (P), dwellings/households (D), economic (EF) and census families (CF) within the Toronto Census Metropolitan Area in 1986. Not to scale. Adapted from [42]. * Relatives other than spouse, common-law partner or never-married sons and daughters.

Person	Age	Marital status	Relationship	Census family	Economic family
John	63	Now married	Person 1	1	A
Marie	59	Now married	Wife	1	A
Julie	37	Widowed	Daughter	2	A
Robert	12	Single	Grandchild	2	A
Lucie	09	Single	Grandchild	2	A
Marc	25	Separated	Son	-	A
Nicole	12	Single	Niece	-	A
Benjamin	14	Single	Lodger (ward)	-	-
Brian	24	Now married	Lodger	3	B
Janet	21	Now married	Lodger's wife	3	B
Jerry	03	Single	Lodger's son	3	B

Table 3.2: An example household containing unusual family structure. As shown in the census family column, there are three census families here, and three persons who are not in any census family. Marc does not belong to a census-family because he is not a "never-married" child; Nicole is not in a census family because she is not a child of any person in the household; and Benjamin is a foster child and is hence treated as a lodger. The economic family column shows how these same persons can be grouped into two economic families, plus one non-family person (Benjamin). Source: [45].

uals or the gender of one individual is changed, making it an opposite-sex marriage [45]. Finally, foster children are treated as lodgers rather than family members. Table 3.2 details an example household that illustrates several unusual aspects of these family definitions.

The connection between households and families is also illustrated in Figure 3.2. Each "private household" occupies one dwelling, in the language of the census. This one-to-one relationship between private households and "occupied private dwellings" means that the household PUMS can be used as a PUMS for dwellings. Occupied private dwellings are only one part of the dwelling universe, but almost no data is available on other types of dwellings. The missing parts of this universe are collective dwellings, dwellings occupied by foreign/temporary residents, unoccupied

		Data Source (and sample size)							
Attribute	Description	Profile 2B (20%)	CF86A04 (100%)	DM86A01 (100%)	LF86B01 (20%)	LF86B03 (20%)	LF86B04 (20%)	SC86B01 (20%)	Person PUMS (2%)
AGEP	Age		4*	16	6			6	c
CFSTAT	Census Family Status		5						11
HLOSP	Highest Level Of Schooling					7		6	12
LFACT	Labour Force Activity				3	3			15
OCC81P	Occupation						24		17
SEXP	Sex	2	2	2	2	2	2	2	2
TOTINCP	Total Income	11							c
CTCODE	Census Tract	731	731	731	731	731	731	731	

c continuous, discretized to integer; large number of categories

* missing breakdown for a few cells.

Table 3.3: Overview of Person attributes, showing the number of categories for the attributes in each data source. Each column describes a single multiway cross-tabulation derived from the given data source. The rest of the profile tables add no further information, and are not shown.

dwellings, some marginal dwellings (e.g., cottages that are not occupied year-round), and some dwellings under construction or conversion.[2]

3.2 Agent Attributes

The census offers a broad range of attributes that could be used in synthesis. Tables 3.3, 3.4 and 3.5 show the attributes selected for synthesis, and the relevant data sources that include these attributes.

Both the Household PUMS and the Family PUMS lack information on the number

[2]The only data on these dwellings are province-wide, in [41] and [48].

		Data Source (and sample size)				
Attribute	Description	CF86A02 (100%)	CF86A03 (100%)	LF86B08 (20%)	Family PUMS (1%)	Reweighted Person PUMS
AGEF	Age (female)				c	c
AGEM	Age (male)				c	c
CFSIZE	Census Family Size				7†	7
CFSTRUC	Census Family Structure	3	3		16	3†
CHILDA	Number of Children 0-5		2	2	3	
CHILDB	Number of Children 6-14		2	2‡	4	
CHILDC	Number of Children 15-17		2	‡	3	
CHILDDE	Number of Children 18-24, 25+		2	‡	9	
HHSIZE	Household Size					8
HHNUMCF	Number of Families in Household					3
LFACTF	Labour Force Activity (female)			3	13	15
LFACTM	Labour Force Activity (male)				13	15
NUCHILD	Number of Children	6	2	2	9	8†
ROOM	Dwelling # of Rooms				10	10
TENURE	Tenure				2	2
CTCODE	Census Tract	731	731	731		

c continuous, discretized to integer; large number of categories

† inferred from other attributes

‡ 2 categories for "number of children ages 6 and higher".

Table 3.4: Overview of Census Family attributes, showing the number of categories for the attributes in each data source. While HHSIZE and HHNUMCF are not present in any family tables, they are present in the Person PUMS, which can be reweighted to a family universe for synthesis. The profile tables add no information beyond that already in the BSTs, and are not shown.

		Data Source (and sample size)								
Attribute	Description	DW86A01 (100%)	DW86A02 (100%)	DW86B02 (20%)	DW86B04 (20%)	HH86A01 (100%)	HH86A02 (100%)	HH86B01+B02 (20%)	Household PUMS (1–4%)	Reweighted Person PUMS
BUILTH	Dwelling Age			8					7	
DTYPEH	Dwelling Type	4	4	4	4				8	
HHNUEF	# Econ. Fam. in HH								2	2
HHNUMCF	# Cens. Fam. in HH					3	3	3		3
HHSIZE	Household Size		10				10		8	8
PAYH	Monthly Dwell. Cost							5	c	c
PPERROOM	Persons Per Room				5				5†	5†
ROOM	Dwelling # of Rooms								10	10
TENURH	Household Tenure	3				2		2	2	2
CTCODE	Census Tract	731	731	731	731	731	731	731		

c continuous, discretized to integer; large number of categories

† inferred from other attributes

Table 3.5: Overview of Household/Dwelling Unit attributes, showing the number of categories for the attributes in each data source. Each column shows a single data source's coverage of different attributes. Note that HHNUMCF is missing from the Household PUMS, but present in the Person PUMS, where it can be reweighted to a household or economic family universe. The profile tables add no information beyond that already present in the BSTs, and are not shown.

of census families sharing a dwelling, and the Family PUMS also lacks information about the household size. These attributes would be useful, but can fortunately be derived from another source: the Person PUMS. Suppose that we consider only the family persons in the Person PUMS, and treat each person as an observation of a census family. Then, the attributes from the Person PUMS could be used to derive information about census families. A similar procedure could be used to gain additional information about households.

However, persons in large families are over-represented in the person PUMS. For example, consider the complete population of families and persons, ignoring for the moment the small sample in the PUMS itself. A family of eight persons is repeated eight times in the person population, while a family of two persons is repeated twice. Large families are thus overrepresented in the person population, but this can be corrected by weighting each observation in the person population by $1/$CFSIZE, the inverse of the family size. In the PUMS, not every member of an eight-person family will be present in the Person PUMS, but large families will still be observed proportionately more often, and the same reweighting method can be applied to correct this.

3.3 Exploration of a Summary Table

To help understand the census data (and contingency tables in general), a brief examination of a single summary table is useful. This exploration focuses on the SC86B01 table, a summary table that cross-classifies age, sex and education by zone. The study area is the Toronto CMA, and the geography has been simplified to a set of twelve zones. Table 3.6 shows the counts in SC86B01, excluding the geographic breakdown. Figure 3.3 shows the same information graphically.

What are the statistical properties of this table? Is there statistically significant association between these variables? Is there significant geographic variation? A log-

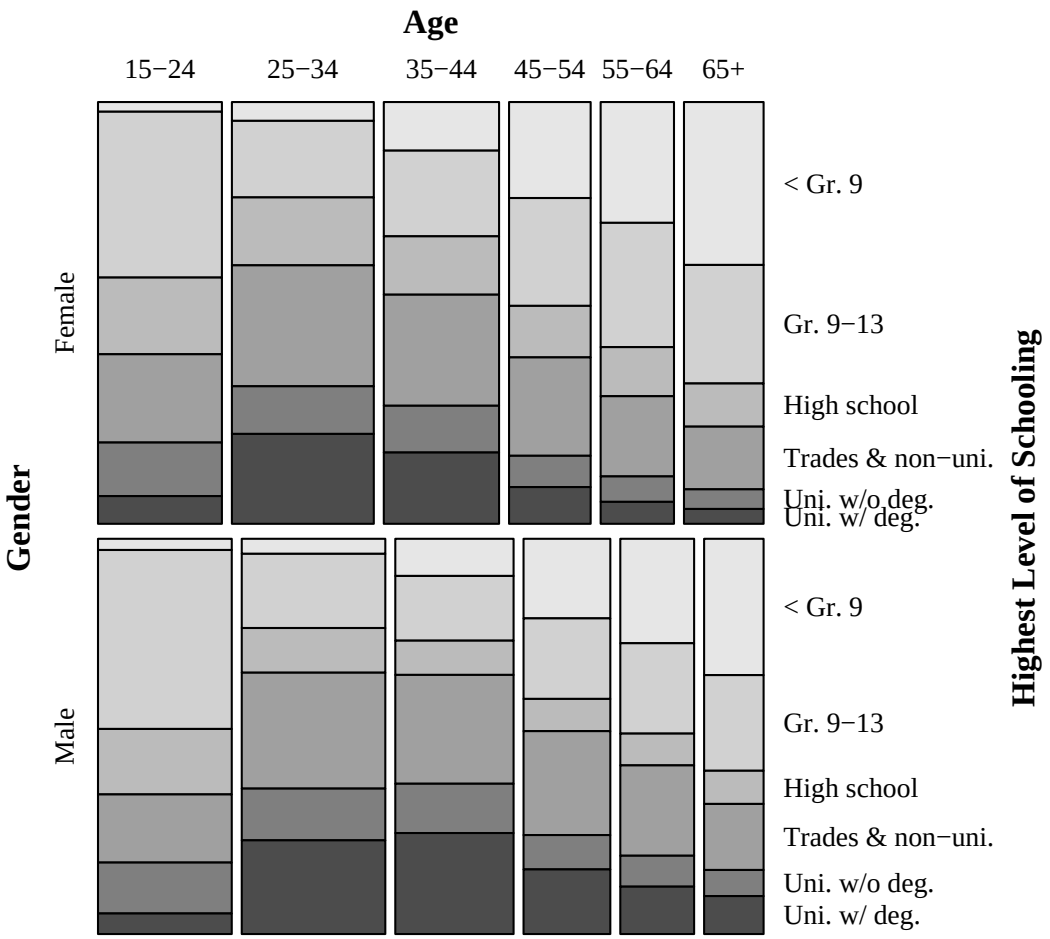

Figure 3.3: A mosaic plot showing the breakdown of the SC86B01 summary tables: population by sex, age and highest level of schooling. Mosaic plots are useful tools for visualizing the breakdown of categories in low-dimensional contingency tables [22]. As usual for these plots, the area of each box represents the number of persons with a given sex, age and schooling. The difference in age breakdown between the two genders can be easily seen, and the differences in the schooling breakdown between each age group can also be seen. Shading has been added to make it easier to see similar schooling levels.

		Age					
Sex	Highest Level of Schooling	15–24	25–34	35–44	45–54	55–64	65+
Female	Less than grade 9	6440	14330	30050	41980	47515	69550
	Grades 9–13	110165	58255	52950	47170	48870	50600
	High school	50930	51645	36085	22540	19300	18425
	Trades and non-uni	58650	92025	68655	43035	31550	26760
	University w/o degree	35570	36250	28900	13685	10005	8380
	University w/ degree	18410	68395	44060	16060	8625	6340
Male	Less than grade 9	8035	11575	23565	37025	41335	43465
	Grades 9–13	128325	57110	41030	37470	35780	30505
	High school	46955	34400	21725	14985	12580	10575
	Trades and non-uni	48870	89200	69015	48350	35765	21055
	University w/o degree	36505	39805	31245	15990	12240	8325
	University w/ degree	14735	72130	64040	30060	18755	12105

Table 3.6: The contents of the SC86B01 summary tables: population by sex, age and highest level of schooling. Since this table is derived from a 20% sample, these counts have been expanded by a factor of five from the original sample.

linear model can be used to answer these questions. In the following, the variables $W(h)$, $X(i)$, $Y(j)$ and $Z(k)$ will be used to represent gender, age, level of schooling and zone respectively.

First, to consider statistically significant association between the variables (excluding geography), a hierarchy of models can be constructed. The final model in this hierarchy (WXY) defines all-way association between the non-geographic variables, and is given by

$$\log \hat{N}_{hijk}/5 = \lambda + \lambda_W + \lambda_X + \lambda_Y + \lambda_{WX} + \lambda_{WY} + \lambda_{XY} + \lambda_{WXY} \tag{3.1}$$

To test for this three-way association, the G^2 statistics of model (WXY) and the restricted model (WX, WY, XY) are compared and tested using a chi-squared distribution. Because SC86B01 is derived from a 20% sample that was expanded to 100%, the counts must be deflated by a factor of five before estimating the models. Table 3.8 shows the complete series of models leading up to (WXY). Each model in the series

Model	New term	Df	Deviance (ΔG^2)	Residual Df	Residual Deviance (G^2)
NULL				863	628862
(W)	SEXP	1	568	862	628294
(W, X)	AGEP	5	40637	857	587657
(W, X, Y)	HLOSP	5	63315	852	524342
(WX, Y)	SEXP $\times$ AGEP	5	1537	847	522806
(WX, WY)	SEXP $\times$ HLOSP	5	4338	842	518467
(WX, WY, XY)	AGEP $\times$ HLOSP	25	102424	817	416043
(WXY)	SEXP $\times$ AGEP $\times$ HLOSP	25	3525	792	412517

Table 3.7: Series of log-linear models to test for association between gender, age and highest level of schooling in SC86B01 table. Each row shows a model that adds one term to the model in the previous row. (The complete model is shown using symbols W, X and Y for compactness, but these correspond to SEXP, AGEP and HLOSP.) The statistical significance of each model is tested using the chi-square statistic between adjacent rows; all models are significant at the 99% level.

exhibits statistically significant improvement in fit over the previous model. Consequently, we can reject the hypothesis that there is no three-way association between gender, age and highest level of schooling.

In a second series of models, the influence of geography is included. (In this analysis, the simplified 12-zone representation of geography is used; the full 731-zone system cannot be analyzed with a log-linear model, due to the memory requirements of generalized linear model estimation.) Table 3.8 shows the series of log-linear models leading up to the saturated model $(WXYZ)$. As shown, every model is statistically significant with respect to the next simplest model; we can therefore conclude that there is significant four-way association in this dataset. Furthermore, the (WX, WY, XY, Z) model describes 95% of the deviance in the data; while the higher-order geographic associations are statistically significant, they are responsible for only a small part of the total deviance.

Model	New term	Df	Deviance (ΔG^2)	Residual Df	Residual Deviance (G^2)	$P(>\|X\|)$
NULL				863	628862	0
(W)	SEXP	1	568	862	628294	0
(W, X)	AGEP	5	40637	857	587657	0
(W, X, Y)	HLOSP	5	63315	852	524342	0
(W, X, Y, Z)	ZONE	11	381164	841	143178	0
(WX, Y, Z)	SEXP $\times$ AGEP	5	1537	836	141641	0
(WX, WY, Z)	SEXP $\times$ HLOSP	5	4338	831	137303	0
(WX, WY, XY, Z)	AGEP $\times$ HLOSP	25	102424	806	34878	0
(WX, WY, WZ, XY)	SEXP $\times$ ZONE	11	207	795	34671	0
(WX, WY, WZ, XY, XZ)	AGEP $\times$ ZONE	55	11130	740	23541	0
(WX, WY, WZ, XY, XZ, YZ)	HLOSP $\times$ ZONE	55	15950	685	7591	0
(WXY, WZ, XZ, YZ)	SEXP $\times$ AGEP $\times$ HLOSP	25	3520	660	4071	0
(WXY, WXZ, YZ)	SEXP $\times$ AGEP $\times$ ZONE	55	304	605	3767	0
(WXY, WXZ, WYZ)	SEXP $\times$ HLOSP $\times$ ZONE	55	733	550	3034	0
(WXY, WXZ, WYZ, XYZ)	AGEP $\times$ HLOSP $\times$ ZONE	275	2573	275	461	0
$(WXYZ)$	SEXP $\times$ AGEP $\times$ HLOSP $\times$ ZONE	275	461	0	0	0

Table 3.8: Series of log-linear models testing association in SC86B01, including geography. Each row shows a model that adds one term to the model in the previous row. The statistical significance of each model is tested using the chi-square statistic between adjacent rows; all models are significant at the 99% level.

Model	New term	Df	Deviance (ΔG^2)	Residual Df	Residual Deviance (G^2)	$P(> \|X\|)$
NULL				863	413094	
(W)	SEXP	1	0.07	862	413094	1
(W, X)	AGEP	5	7	857	413088	0.23
(W, X, Y)	HLOSP	5	11	852	413077	0.05
(W, X, Y, Z)	ZONE	11	381164	841	31912	0
(WX, Y, Z)	SEXP $\times$ AGEP	5	2	836	31910	1
(WX, WY, Z)	SEXP $\times$ HLOSP	5	72	831	31838	0
(WX, WY, XY, Z)	AGEP $\times$ HLOSP	25	280	806	31558	0
(WX, WY, WZ, XY)	SEXP $\times$ ZONE	11	207	795	31351	0
(WX, WY, WZ, XY, XZ)	AGEP $\times$ ZONE	55	11130	740	20221	0
(WX, WY, WZ, XY, XZ, YZ)	HLOSP $\times$ ZONE	55	15945	685	4276	0
(WXY, WZ, XZ, YZ)	SEXP $\times$ AGEP $\times$ HLOSP	25	205	660	4071	0
(WXY, WXZ, YZ)	SEXP $\times$ AGEP $\times$ ZONE	55	304	605	3767	0
(WXY, WXZ, WYZ)	SEXP $\times$ HLOSP $\times$ ZONE	55	733	550	3034	0
(WXY, WXZ, WYZ, XYZ)	AGEP $\times$ HLOSP $\times$ ZONE	275	2573	275	461	0
$(WXYZ)$	SEXP $\times$ AGEP $\times$ HLOSP $\times$ ZONE	275	461	0	0	0

Table 3.9: Series of log-linear models testing association in SC86B01 relative to PUMS. The left hand side of the model is the ratio of the SC86B01 count to the PUMS count for the same cell.

The final set of models shown in Table 3.9 simulate the effect of using IPF with a particular set of margins from SC86B01. The left hand side follows equation (2.18), dividing the fitted margins of SC86B01 ($\hat{N}_{hijk}/5$) by the PUMS (n_{hij}):

$$\log \hat{N}_{hijk}/5n_{hij} = \lambda + \lambda_W + \lambda_X + \lambda_Y + \lambda_Z + \cdots \tag{3.2}$$

(In practice, the PUMS term n_{hij} is used as an *offset* to the generalized linear model.) This series of models also shows statistically significant improvements, except for the first few terms. Effectively, the series shows the amount of information that SC86B01 adds beyond what is already available in the PUMS. The low deviance associated with the one-way models indicates that the 20% SC86B01 sample adds little information to the 2% PUMS sample of these variables. Terms involving ZONE, by contrast, add a lot of information, since the PUMS includes no geographic variation. The main difference between Tables 3.8 and 3.9 is that the deviance associated with terms that do *not* involve ZONE drops by 92% or more when the PUMS is included, and usually drops by more than 99%. Much of the non-geographic information is already present in the PUMS.

Furthermore, the inclusion of higher-order interactions shows diminishing returns in terms of the explained deviance. The one-way model (W, X, Y, Z) explains 92.3% of the deviance in the NULL model. Of the remaining 7.7% of total deviance, the two-way model (WX, WY, WZ, XY, XZ, YZ) explains 86.6%. Of the final 1.0% of total deviance, the three-way model (WXY, WXZ, XYZ) explains 89.2% and the four-way model $(WXYZ)$ explains the final 10.8%. The total deviance does depend on the choice of variables and the fineness of the categories in the table, but this trend of diminishing returns is interesting. It suggests that the available census data—largely describing lower-order interactions, with only a few higher-order interactions, apart from the 2%-sample PUMS—may capture the bulk of the actual information about the population. However, this single table is clearly not sufficient to say anything conclusive.

In closing, this analysis has focused on a single contingency table, SC86B01. No attempt was made to find the *best* model for SC86B01, particularly in terms of model parsimony; instead, the analysis demonstrated that statistically significant higher-order interactions are present in the data. Furthermore, it is not possible to apply this type of log-linear analysis to multiple contingency tables, although if multiple tables are combined using a fitting procedure the result could be analyzed. The largest limitation, however, is one of software: log-linear analysis is not generally feasible with high-dimensional tables. Nevertheless, the analysis provides valuable insight about the utility of information recorded in contingency tables.

Chapter 4

Method Improvements

The existing population synthesis procedures have many limitations. In this chapter, the following concerns are discussed:

- High-dimensional contingency tables are very sparse, and the many-way association patterns in them have little statistical significance
- The number of attributes that can be synthesized is quite limited, for the IPF method in particular, and also for reweighting/combinatorial optimization to a lesser extent
- Random rounding of reported tables may influence the quality of results
- Synthesizing populations of related agents is challenging

The chapter is structured largely as a discussion of these issues, together with an attempt to brainstorm solutions to these limitations, focusing particularly on the IPF-based methods; not all ideas are entirely successful. The new methods are intended to be largely independent, but most can be combined if desired. In the following chapters, several of these new methods are implemented and evaluated.

# of			Cells Equal To		
Attributes	# Cells	Median	0	1	2+
1	3	956	0.0%	0.0%	100.0%
2	6	478	0.0%	0.0%	100.0%
3	54	30	5.6%	1.9%	92.6%
4	486	0	57.8%	7.8%	34.4%
5	4374	0	84.7%	4.0%	11.3%
6	21870	0	93.9%	2.4%	3.7%
7	196830	0	98.9%	0.5%	0.6%
8	984150	0	99.7%	0.1%	0.1%

Table 4.1: Illustration of relationship between sparsity and number of dimensions/cross-classification attributes. Reading from the top, the table shows increasing sparsity as a sample is cross-classified using a larger number of attributes. Alternatively, the table can be read from the bottom: starting with an 8-way contingency table, one dimension is collapsed at a time, giving a progressively denser table. By the time the 8-way table has been collapsed to a 3-way table, the majority of cells are non-zero. Data: 1986 Census PUMS of $n = 9061$ families in the Toronto CMA, cross-classified using dimensions CFSTRUC (3 categories), TENURE (2), ROOM (10), NUCHILD (9), AGEF (9), LFACTF (5), AGEM (9) and LFACTM (5).

4.1 Simplifying the PUMS

The IPF method has been applied to tables with a large number of dimensions, as many as eight. Such high dimensional tables are almost always sparse, with the vast majority of the cells in the table containing zeros. (High dimensional tables are generally notorious for their difficulties and differences from two-dimensional tables; see [12] for "ubiquitous" ill-behaved examples.) In these sparse tables, a large fraction of the non-empty cells contain only one observation, and the number of cells is often much larger than the number of observations. In other words, the sample does not provide a statistically meaningful estimate of the probability distribution for such a high dimensional table. However, the high-dimensional table can be collapsed to produce 2D or 3D tables, each of which is adequately sampled and gives a statistically

valid distribution of counts. This is illustrated in Table 4.1, where 99.7% of the cells are zero in the 8-way table. The two-way and three-way margins of the table shown (CFSTRUC × TENURE and CFSTRUC × TENURE × ROOM) have 0% and 5.6% zeros, and median cell counts of 90 and 30 respectively. Of course, an 8-way table has many other lower-dimensional margins, and only one such choice is illustrated in Table 4.1. (When collapsing a d-way table to d' dimensions, there are $\binom{d}{d'}$ possible choices of variables to keep in the low dimensional table.) Consequently, while a sparse 8-way table does not provide statistically testable 8-variable interaction, it could be viewed as a means of linking the many 2- or 3-way tables formed by its margins.

The cell counts in the high dimensional table say very little statistically; in fact, half of the non-zero cells contain just one observation. The most important information in an 8-way table is not necessarily the counts in the cells, but rather the *co-ordinates* of the non-zero cells, which determine how each cell influences the various low-dimensional margins. In this manner, the importance of the high-dimensional table is in many ways its *sparsity pattern*, which provides the link between the many 2- and 3-way tables. However, the sparsity pattern in the high dimensional table is in some ways an artefact of sampling. Some of the zeros in the table represent *structural zeros*: for example, it is essentially impossible for a family to exist where the wife's age is 15, her education is university-level and the family has eight children. Other zeros may be *sampling zeros* where the sample did not observe a particular combination of variables, but it does exist in the population.

All of the population synthesis strategies reviewed in Chapter 2 preserved the sparsity pattern of the high-dimensional PUMS. It would be useful if the synthesis procedure could account for the uncertainty surrounding the sparsity pattern. One strategy for resolving this would be to *simplify* the PUMS, building a variant that fits the low-dimensional margins but makes no assumptions about the associations or sparsity pattern in high dimensions where sampling is inadequate. The primary goal

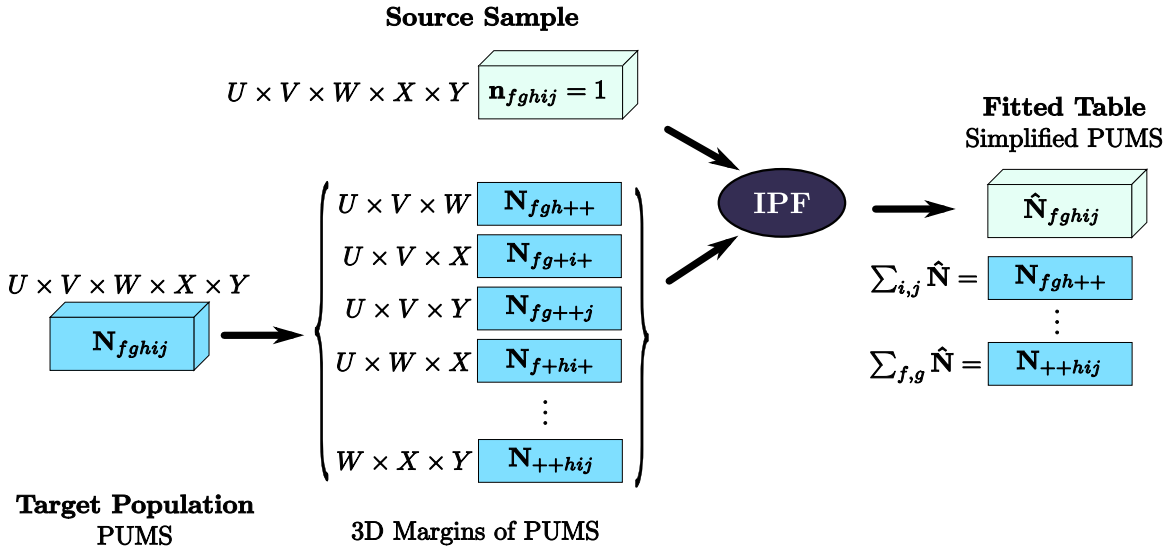

Figure 4.1: Simplifying the PUMS by removing high-dimensional associations. A five-dimensional PUMS is simplified by removing four- and five-way associations. This is achieved by fitting a uniform probability table to the complete set of 3-way margins of the PUMS. The resulting table has the same 1-, 2- and 3-way margins as the original PUMS, but has no 4- or 5-way association.

would be to correct sampling zeros in the high-dimensional PUMS, replacing them with small positive counts.

This strategy is illustrated in Figure 4.1. IPF is applied on a high-dimensional table, with the source sample set to a uniform distribution. The constraints applied to IPF are the full set of (say) 3-way cross-tabulations of the PUMS. For a 5-dimensional problem, this is equal to $\binom{5}{3} = 10$ separate constraints. After IPF, the simplified PUMS matches all of the 3-way margins from the PUMS, but includes no 4-way or 5-way interaction between variables. In this example, a 5-way table is shown for illustration, but the intention is for higher-dimensional tables to be simplified. Also, 3-way margins of the PUMS are used as constraints for illustrative purposes, but the actual margins chosen could be selected using tests of statistically significant interaction; If a log-linear model shows some significant 4-way interactions in the PUMS, then these could be included as margins.

This method does indeed reduce the number of zeros in the high-dimensional table, correcting for sampling zeros and capturing some unusual individuals that were not included in the original PUMS. However, this strategy has significant downsides. In particular, it makes no distinction between structural and sampling zeros at high dimensions: there may be structural zeros involving four or more variables that do not appear in 3-way tables, but they will be treated the same as sampling zeros and be filled in with small positive counts.

Furthermore, it makes the entire population synthesis procedure more difficult, since agents can be synthesized who did not exist in the original PUMS. This makes the technique difficult to combine with some of the other methodological improvements discussed here.

4.2 Sparse List-Based Data Structure

For a microsimulation model spanning many aspects of society and the economy it is useful to be able to associate a range of attributes with each agent. Different attributes are useful for different aspects of the agent's behaviour. For a person agent, labour force activity, occupation, industry and income attributes are useful for understanding his/her participation in the labour force. Meanwhile, age, marital status, gender and education attributes might be useful for predicting demographic behaviour.

However, as more attributes are associated with an agent, the number of cells in the corresponding multiway contingency table grows exponentially. A multiway contingency table representing the association pattern between attributes has $I \times J \times K \times \ldots$ cells. If a new attribute with L categories is added, then L times more cells are needed. Asymptotically, the storage space is exponential in the number of attributes. As a result, fitting more than eight attributes with a multizone IPF procedure typically requires more memory than available on a desktop computer. However, the table itself

is cross-classifying a fixed number of observations (i.e., a PUMS), and is extremely sparse when a large number of attributes are included as shown in Table 4.1. Is there a way that this sparsity can be exploited to allow the synthesis to create a large number of attributes?

Sparsity is a familiar problem in numerical methods. Many branches of science store large sparse 2D matrices using special data structures that hold only the non-zero sections of the matrix, instead of using a complete array that includes cells for every zero. In the data warehousing field, for example, multiway tables were initially stored in complete arrays accessed through a procedure called MOLAP (Multidimensional On-line Analytical Processing). More recently, that field has shifted to ROLAP (Relational OLAP) methods that allow data to be stored in its natural microdata form while retaining the ability to answer queries efficiently [8].

The IPF method itself has little impact on the sparsity pattern of a table. After the first iteration of fits to the constraints, the final sparsity pattern is essentially known; some cells may eventually converge to near-zero values, but few other changes occur. Once a cell is set to zero, it remains zero for the remainder of the procedure. Also, the IPF algorithm does not require a complete representation of the underlying table. All that it needs are three operations (described later), which could be implemented using either a complete or sparse representation of the data.

The benefits of using a sparse data structure are substantial. Williamson et al. presented many of the arguments when describing their Combinatorial Optimization method: efficiency in space, flexibility of aggregation and easier linking of data sources [57]. In terms of efficiency, the method described here allows the IPF algorithm to be implemented using storage proportional to the number of non-zero cells in the initial table. For agent synthesis with the zone-by-zone method, this is proportional to n (the number of observations in the PUMS) multiplied by d (the number of attributes to fit). The multizone method combines several IPF stages, and requires

# PUMS Attributes	Complete Storage		Sparse List	
d	Zone-by-Zone	Multizone	Zone-by-Zone	Multizone
1	0.00 MB	0.0 MB	0.05 MB	26.5 MB
2	0.00 MB	0.0 MB	0.05 MB	26.5 MB
3	0.00 MB	0.2 MB	0.06 MB	26.5 MB
4	0.00 MB	1.6 MB	0.07 MB	26.5 MB
5	0.02 MB	14.2 MB	0.08 MB	26.5 MB
6	0.10 MB	71.1 MB	0.09 MB	26.5 MB
7	0.87 MB	639.5 MB	0.10 MB	26.6 MB
8	4.37 MB	3,197.4 MB	0.11 MB	26.6 MB
9	13.12 MB	9,592.2 MB	0.12 MB	26.6 MB
10	52.49 MB	38,368.7 MB	0.13 MB	26.6 MB
11	157.46 MB	115,106.2 MB	0.14 MB	26.6 MB
12	472.39 MB	345,318.6 MB	0.14 MB	26.6 MB
13	1,417.18 MB	1,035,955.7 MB	0.15 MB	26.6 MB

Table 4.2: Comparison of memory requirements for implementations of an agent synthesis procedure using a complete array or a sparse list. The Zone-by-Zone columns show the memory requirements when synthesizing d attributes, all present in the PUMS; the Multizone columns show the storage requirements for d PUMS attributes plus one non-PUMS attribute, a zone number within the PUMA. Data: PUMS of $n = 9061$ census families in the 1986 Toronto CMA, cross-classified using variables CFSTRUC (3 categories), TENURE (2), ROOM (10), NUCHILD (9), AGEF (9), LFACTF (5), AGEM (9), LFACTM (5), CHILDA (3), CHILDB (4), CHILDC (3), CHILDD (3) and CHILDE (3). The geography variable CTCODE is divided into $K = 731$ zones.

considerably more memory: a similar $\mathcal{O}(nd)$ in the first stage, but $\mathcal{O}(n(d+K))$ in the second stage, where K is the number of zones. This is illustrated in Table 4.2.

There are many types of data structures that could be used to represent a sparse high dimensional contingency table. The data structure proposed here is not the most efficient, but is conceptually simple. It borrows directly from Williamson's Combinatorial Optimization method: the data is represented as a list of the PUMS microdata entries, with a weight attached to each. The weight is an expansion factor, representing the number of times to replicate that record to form a complete population. Williamson's representation includes only integer weights, and operates on a zone-by-zone basis. The new approach described here behaves exactly like IPF and hence allows fractional weights. With a small extension, it can also supports multiple zones: instead of attaching one weight to each PUMS entry, K weights are attached with one for each zone. An illustration of the format of the data structure is shown in Table 4.3.

As Williamson et al. pointed out, flexible aggregation is a real advantage of a list-based representation. Complete array storage is proportional to the number of categories used for each attributes, while the sparse storage scheme is not affected by the categorization of the attributes. Many applications of IPF that used complete arrays were forced to abandon detailed categorization schemes to conserve space and allow more attributes to be synthesized (e.g., [2]). This in turn makes it difficult to apply several margins, since different margins may categorize a single attribute differently. When a large number of categories are possible, however, the attribute can be represented with a fine categorization and collapsed to different coarse categorizations as required during the fitting procedure.

4.2.1 Algorithmic Details

The operation of a typical IPF algorithm was presented earlier in Figure 2.4. To implement this procedure with the sparse list structure, the following operations are

(a)

	Co-ordinates						
Index	CFSTRUC	ROOM	TENURE	NUCHILD	...	CHILDE	Weight
1	Husband-wife	7	Owned	3	...	1	81.8
2	Lone female parent	4	Rented	0	...	0	70.9
3	Husband-wife	9	Rented	0	...	0	54.8
4	Husband-wife	9	Owned	0	...	0	86.2
...	...	...	...	...	...	...	...
9060	Husband-wife	9	Rented	0	...	0	64.8
9061	Husband-wife	6	Rented	0	...	0	100.3

(b)

	Co-ordinates						Weight			
Index	CFSTRUC	ROOM	TENURE	NUCHILD	...	CHILDE	CTCODE1	CTCODE2	...	CTCODE731
1	Husband-wife	7	Owned	3	...	1	0.000	0.121	...	0.021
2	Lone female parent	4	Rented	0	...	0	0.000	0.212	...	0.020
3	Husband-wife	9	Rented	0	...	0	0.000	0.244	...	0.143
4	Husband-wife	9	Owned	0	...	0	0.002	0.037	...	0.019
...	...	...	...	...	...	...	...	...	...	
9060	Husband-wife	9	Rented	0	...	0	0.000	0.349	...	0.011
9061	Husband-wife	6	Rented	0	...	0	0.004	0.213	...	0.074

Table 4.3: Format of a sparse list-based data structure for Iterative Proportional Fitting As shown, each row corresponds to a PUMS entry. The columns give the co-ordinates of each PUMS entry within the high-dimensional array. Each row also stores (a) a single weight when synthesizing only PUMS attributes (e.g., a zone-by-zone IPF); or (b) a set of weights, corresponding to the categories of a non-PUMS attribute (e.g., a multizone IPF where the non-PUMS attribute defines a zone (census tract) code within the PUMA).

necessary:

- Set the initial weights (line 2).
- Collapse to the dimensions of a target margin (e.g., $\hat{N}_{i+}^{(\tau)}$ in line 5).
- Update a weight according to its location within a target margin (e.g., $\hat{N}_{ij}^{(\tau+1)}$ in line 5).

If the target population margins remain stored as a complete array, these operations are relatively straightforward. The collapse operation can be done in a single pass over the list, using the category numbers in each list row as co-ordinates into the complete array that stores the collapsed table. The update operation can likewise be done in a single pass over the list. All of these operations are fast, with complexity equal to the storage cost, $\mathcal{O}(nd)$. (The method is basically the same when non-PUMS attributes are included and multiple weights are stored per row in the list. The non-PUMS attribute must be handled as a special case, since it is stored slightly differently. Since there are K times more weights, the computation cost grows proportionally to $\mathcal{O}(n(d+K))$, again matching the storage cost.)

The only tricky part of the procedure is setting the initial weights. When using a complete array representation for zone-by-zone methods, the initial table needs to follow the distribution of the PUMS, while for Beckman et al.'s multizone method, the initial table needs to be a uniform distribution, usually by setting all cells to one. In the sparse representation described here, the situation is reversed because there is one row per PUMS entry, instead of one cell grouping many PUMS entries. As a result, the initial weights need to be 1.0 for the zone-by-zone method. For the multizone method, the sum of the weights in the r rows that contribute to a single "cell" in the complete table needs to add to one. Therefore, the individual weights should be $1/r$. The tricky part is finding how many rows share a cell, since the data is not structured for this purpose. To achieve this, the list is sorted by the table dimensions, which has the

effect of grouping rows contributing to a single cell. In this manner, the initial weights can be set in $\mathcal{O}(n \log n)$ time.

Additionally, the multizone procedure requires a fit to the distribution of all non-geographic variables simultaneously. (See the $\hat{\mathbf{N}}_{ij+}$ margin on the right half of Figure 2.6.) This margin is high-dimensional—it includes all variables except for the geographic variable. However, it is also sparse, and is in fact computed through an IPF procedure. As a result, it is already stored as a sparse list with a single weight per row. This weight can be treated as a constraint on the total for the weights in each row, and suitable collapse/update procedures are then easily defined. The computation cost is still $\mathcal{O}(n(d + K))$ for this type of constraint.

Finally, the Monte Carlo integerization for this sparse structure is quite simple, and little changed. The list of weights (or 2D array of weights for the multizone procedure) is normalized and treated as a probability mass function, and individual rows (cells for multizone) are synthesized using Monte Carlo draws.

4.2.2 Discussion

This sparse data structure removes a substantial limitation from the IPF algorithm, but also raises new questions. Is there a limit to the number of attributes that can be synthesized? If there is a limit, how is it related to the size n of the PUMS sample?

The answers to these questions remain elusive. Ultimately, the addition of a new attribute is a decision to increase the dimension of the multiway contingency table. As discussed earlier, the behaviour of this high-dimensional table and its relationship to lower-dimensional margins remains poorly understood.

4.3 Fitting to Randomly Rounded Margins

Many census agencies apply random rounding procedures to published tables, including the agencies in Canada, the United Kingdom and New Zealand. Each agency has a base b that it uses, and then modifies a cell count N_{i+} by rounded up to the nearest multiple of b with a probability p, or down with a probability $1-p$. In most applications, a procedure called *unbiased* random rounding is used, where $p = (N_{i+} \bmod b)/b$. The alternative is called *unrestricted* random rounding, where p is constant and independent of the cell values; for example, with $p = 0.5$ it is equally likely that a cell will be rounded up or down.

For example, cells and marginal totals in Canadian census tables are randomly rounded up or down to a multiple of $b = 5$ using the unbiased procedure. For a cell with a count of $N_{i+} = 34$, there is a 20% probability that it is published as $\tilde{N}_{i+} = 30$ and an 80% probability that it is published as $\tilde{N}_{i+} = 35$. Most importantly, the expected value is equal to that of the unrounded count; it is therefore an unbiased random rounding procedure.

As discussed by Huang & Williamson [28], this can lead to conflicts between tables: two different cross-tabulations of the same variable or set of variables may be randomly rounded to different values. The standard IPF procedure will not converge in this situation. The procedure is also unable to take into account the fact that margins do not need to be fitted exactly, since there is a reasonable chance that the correct count is within ± 4 of the reported count.

4.3.1 Modified Termination Criterion

Using the termination criterion of Figure 2.4 (line 10), the IPF procedure will not necessarily terminate if two randomly rounded margins conflict. The termination criterion

shown requires the fitted table to match all margins simultaneously:

$$\delta = \max\left(\max_i \left|\hat{N}_{i+}^{(\tau+2)} - N_{i+}\right|, \max_j \left|\hat{N}_{+j}^{(\tau+2)} - N_{+j}\right|\right) \tag{4.1}$$

Instead of requiring a fit, the algorithm could terminate when the net effect of one iteration drops below a threshold. That is,

$$\delta = \max\left(\max_i \left|\hat{N}_{i+}^{(\tau+2)} - \hat{N}_{i+}^{(\tau)}\right|, \max_j \left|\hat{N}_{+j}^{(\tau+2)} - \hat{N}_{+j}^{(\tau)}\right|\right) \tag{4.2}$$

or even

$$\delta = \max_{i,j} \left|\hat{N}_{ij}^{(\tau+2)} - \hat{N}_{ij}^{(\tau)}\right| \tag{4.3}$$

The intention here is to terminate the algorithm when the *change* in error in the margins drops below a threshold, instead of the absolute error.

4.3.2 Hierarchical Margins

For each cross-tabulation, statistical agencies publish a hierarchy of margins, and these margins are rounded independently of the cells in the table. For a three-way table N_{ijk} randomly rounded to give $\tilde{N}_{ijk}$, the data release will also include randomly rounded two-way margins $\tilde{N}_{ij+}$, $\tilde{N}_{i+k}$ and $\tilde{N}_{+jk}$, one-way margins $\tilde{N}_{i++}$, $\tilde{N}_{+j+}$ and $\tilde{N}_{++k}$, and a zero-way total $\tilde{N}_{+++}$. The sum of the cells does not necessarily match the marginal total. For example, the sum $\sum_k \tilde{N}_{ijk}$ includes K randomly rounded counts. The expected value of this sum is the true count N_{ij+}, but the variance is large and the sum could be off by as much as $K(b-1)$ in the worst case. By contrast, the reported marginal total $\tilde{N}_{ij+}$ also has the correct expected value, but its error is at most $b-1$.

For this reason, it seems sensible to include the hierarchical margins in the fitting procedure, in addition to the detailed cross-tabulation itself.

4.3.3 Projecting onto Feasible Range

As described in equations (2.6) and (2.1), the IPF procedure minimizes the Kullback-Leibler divergence $I(\hat{\mathbf{N}}\|\mathbf{n})$,

$$\sum_i \sum_j \hat{N}_{ij} \log(\hat{N}_{ij}/n_{ij})$$

while satisfying the marginal constraints

$$\sum_j \hat{N}_{ij} = \tilde{N}_{i+}, \quad \sum_i \hat{N}_{ij} = \tilde{N}_{+j}$$

To handle random rounding, the marginal constraints could instead be treated as inequalities,

$$\left| \sum_j \hat{N}_{ij} - \tilde{N}_{i+} \right| \leq b-1, \quad \left| \sum_i \hat{N}_{ij} - \tilde{N}_{+j} \right| \leq b-1 \tag{4.4}$$

That is, any value within the range $\tilde{N}_{i+} \pm (b-1)$ is an acceptable solution, with no preference for any single value within that range.

Dykstra's generalization of IPF [17] provides some fruitful ideas for handling this type of constraint. Csiszár [13] described the IPF procedure as a series of *projections* onto the subspace defined by each constraint. Csiszár was not working in a d-dimensional space (where d is the number of attributes being fitted), but in a C-dimensional space (where C is the number of *cells* in the table) representing all possible probability distributions, which has since been called I-space.

Nevertheless, the idea of *projection* is still useful: each iteration of IPF is a modification of the probability distribution to fit a margin. It is a "projection" in that it finds the "closest" probability distribution in terms of Kullback-Leibler divergence, just as the projection of a point onto a plane finds the closest point on the plane in terms of Euclidean distance. (Note, however, that Kullback-Leibler divergence is *not* a true distance metric.)

Csiszár only considered equality constraints. Dykstra extended Csiszár's method to include a broader range of constraints: any closed convex set in I-space. This

Δ	$P(N_{i+} = \tilde{N}_{i+} + \Delta \mid \tilde{N}_{i+})$
-5	0%
-4	4%
-3	8%
-2	12%
-1	16%
0	20%
1	16%
2	12%
3	8%
4	4%
5	0%

Table 4.4: Relationship between unknown true count and the randomly rounded count published by the statistical agency. The table shows the probability distribution for unrounded count N_{i+} given published randomly rounded count $\tilde{N}_{i+}$, assuming base $b = 5$.

class of constraints appears to include the desired inequality constraints defined by Equation 4.4. Dykstra's method for applying these constraints is also a projection procedure, finding the set of counts that satisfy the constraint while minimizing the Kullback-Leibler divergence.

To give an example, consider the algorithm of Figure 2.4. Line 5 would be replaced with

$$\hat{N}_{ij}^{(\tau+1)} = \begin{cases} \hat{N}_{ij}^{(\tau)} \frac{\tilde{N}_{i+}-(b-1)}{\hat{N}_{i+}^{(\tau)}} & \hat{N}_{i+}^{(\tau)} < \tilde{N}_{i+} - (b-1) \\ \hat{N}_{ij}^{(\tau)} \frac{\tilde{N}_{i+}+(b-1)}{\hat{N}_{i+}^{(\tau)}} & \hat{N}_{i+}^{(\tau)} > \tilde{N}_{i+} + (b-1) \\ \hat{N}_{ij}^{(\tau)} & \text{otherwise} \end{cases} \tag{4.5}$$

(and likewise for line 8).

However, this projection procedure has its own problems. The standard IPF procedure ignores the probability distribution associated with each marginal value and uses only the published cell count. The projection procedure described here suffers from

the opposite problem: it focuses on the range of possible values, without acknowledging that one outcome is known to be more likely than the others. To see this, consider the probability distribution for the unrounded value N_{i+} given the published value $\tilde{N}_{i+}$ shown in Table 4.4. The distribution is triangular, with a strong central peak. The projection algorithm forces the fit to match the range $\pm(b-1)$ of this distribution, but it treats all values inside this range as equally probable. This would be suitable if the census used *unrestricted* random rounding, but not for the more typical case where *unbiased* rounding is used.

4.4 Synthesizing Agent Relationships

Suppose that a population of person agents has been synthesized, with a limited amount of information about their relationships in families (such as a CFSTRUC, which classifies a person as married, a lone parent, a child living with parent(s), or a non-family person). In the absence of any information about how families form, the persons could be formed into families in a naïve manner: randomly select male married persons and attach them to female married persons, and randomly attach children to couples or lone parents. Immediately, problems would emerge: some persons would be associated in implausible manners, such as marriages with age differences over 50 years, marriages between persons living at opposite ends of the city, or parents who are younger than their children.

A well-designed relationship synthesis procedure should carefully avoid such problems. A good choice of relationships satisfies certain *constraints* between agents' attributes, such as the mother being older than her child, or the married couple living in the same zone. It also follows known *probability distributions*, so that marriages with age differences over 50 years have a low but non-zero incidence.

Most constraints and probability distributions are observed in microdata samples

of aggregate agents, such as families and households. A complete Family PUMS includes the ages of mothers and children, and none of the records includes a mother who is younger than her children.[1] Similarly, only a small fraction of the records include marriages between couples with ages differing by more than 50 years. The question, however, is one of method: how can relationships between agents be formed to ensure that the desired constraints are satisfied?

Guo & Bhat [26] used a top-down approach, synthesizing a household first and then synthesizing individuals to connect to the household. The attributes used to link the two universes were gender and age: the gender of the husband/wife or lone parent are known, and coarse constraints on the age of the household head (15–64 or 65+) and children (some 0–18 or all 18+). These constraints are quite loose, and no constraint is enforced between the husband/wife's ages or parent/child ages.

Guan [25] used a bottom-up approach to build families, with slightly stronger constraints. The persons were synthesized first, and then assembled to form families. Children are grouped together (and constrained to have similar ages), then attached to parents. Constraints between parent/child ages and husband/wife ages were included, although there are some drawbacks to the method used for enforcement. Guan likewise used a bottom-up approach to combine families and non-family persons into households.

Arentze & Timmermans [2] only synthesized a single type of agent, the household. Their synthesis included the age and labour force activity of both husband and wife, and the linkage to the number of children in the household. They did not connect this to a separate synthesis of persons with detailed individual attributes, but by synthesizing at an aggregate level, they guaranteed that the population was consistent and

[1] Of course, according to the census definition of family, a "mother" could in fact be a *stepmother*, and there is a small but non-zero probability that she could be younger than her "children." This is not evident anywhere in the Canada-wide PUMS, but there are two other baffling families: one with a 27 year old father, a 24 year old mother, and a child of 25 years or older; the other has a 17 year-old single mother and a child of 18 years or older.

satisfied key constraints between family members.

Both Guo & Bhat and Guan's procedures suffer from inconsistencies between the aggregate and disaggregate populations. The family population may contain 50 husband-wife families in zone k where the husband has age i, while the person population contains only 46 married males of age i in zone k. In the face of such inconsistencies, either families or persons must be changed: a family could be attached to a male of age $i' \neq i$, or a person could be modified to fit the family. In both cases, either the family or person population is deemed "incorrect" and modified. The editing procedures are difficult to perform, and inherently *ad hoc*. Furthermore, as the number of overlapping attributes between the two populations grows, inconsistencies become quite prevalent.

What are the sources of these inconsistencies? They come from two places: first, the fitting procedure used to estimate the population distribution $\hat{\mathbf{N}}^P$ for persons and $\hat{\mathbf{N}}^F$ for families may not give the same totals for a given set of common attributes. Second, even if $\hat{\mathbf{N}}^P$ and $\hat{\mathbf{N}}^F$ agree on all shared attributes, the populations produced by Monte Carlo synthesis may not agree, since the Monte Carlo procedure is non-deterministic. In the following sections, a method is proposed to resolve these two issues.

4.4.1 Fitting Populations Together

For the purposes of discussion, consider a simple synthesis example: synthesizing husband-wife families. Suppose that the universe of persons includes all persons, with attributes for gender SEXP(g), family status CFSTRUC(h), age AGEP(i), education HLOSP(j) and zone CTCODE(k). The universe of families includes only husband-wife couples, with attributes for the age of husband AGEM(i_m) and wife AGEF(i_f), and zone CTCODE(k). IPF has already been used to estimate the contingency table cross-classifying persons ($\hat{\mathbf{N}}^P_{ghijk}$) and likewise for the table of families ($\hat{\mathbf{N}}^F_{i_m i_f k}$). The shared

attributes between the two populations are age and zone, and implicitly gender. The two universes do not overlap directly, since only a fraction of the persons belong to husband-wife families; the others may be lone parents, children, or non-family persons, and are categorized as such using the CFSTRUC attribute.

In order for consistency between $\hat{\mathbf{N}}^P$ and $\hat{\mathbf{N}}^F$, the following must be met for $h =$ *husband-wife* and any choice of i, k:

$$\hat{N}^P_{ghi+k} = \begin{cases} \hat{N}^F_{i+k} & \text{for } g\text{=}\textit{male} \\ \hat{N}^F_{+ik} & \text{for } g\text{=}\textit{female} \end{cases} \tag{4.6}$$

That is, the number of married males of age i in zone k must be the same as the number of husband-wife families with husband of age i in zone k. While this might appear simple, it is often not possible with the available data. A margin $\mathbf{N}^P_{g+i+k}$ giving the SEXP $\times$ AGEP $\times$ CTCODE distribution is probably available to apply to the person population. However, a similar margin for just *married* males is not likely to exist for the family population; instead, the age breakdown for married males in the family usually comes from the PUMS alone. As a result, equation (4.6) is not satisfied.

One suggestion immediately leaps to mind: if the person population is fitted with IPF first and $\hat{\mathbf{N}}^P$ is known, the slice of $\hat{N}^P_{ghi+k}$ where $g =$ *male* and $h =$ *husband-wife* could be applied as a margin to the family fitting procedure, and likewise for $g =$ *female*. This is entirely feasible, and does indeed guarantee matching totals between the populations. The approach can be used for the full set of attributes shared between the individual and family populations. There is one downside, however: it can only be performed in one direction. The family table can be fitted to the person table or vice versa, but they cannot be fitted simultaneously.[2]

Finally, there remains one wrinkle: it is possible that the family population will

[2]It is conceivable that an IPF procedure could be devised where the two populations are fitted in parallel and could be constrained against each other; however, the convergence and discrimination information-minimizing properties of such a process are unknown.

still not be able to fit the total margin from the individual population, due to a different sparsity pattern. For example, if the family PUMS includes no families where the male is (say) 15–19 years old but the individual PUMS does include a married male of that age, then the fit cannot be achieved. This is rarely an issue when a small number of attributes are shared, but when a large number of attributes are shared between the two populations it is readily observed. The simplest solution is to minimize the number of shared attributes, or to use a coarse categorization for the purposes of linking the two sets of attributes.

Alternatively, the two PUMS could be cross-classified using the shared attributes and forced to agree. For example, for $g =$ *male* and $h =$ *husband-wife*, then the pattern of zeros in $\mathbf{n}^P_{ghi++}$ and $\mathbf{n}^F_{i++}$ could be forced to agree by setting cells to zero in one or both tables. (In the earlier example, this would remove the married male of age 15–19 from the Person PUMS.) The person population is then fitted using this modified PUMS, and the family population is then fitted to the margin of the person population.

4.4.2 Conditioned Monte Carlo

The second problem with IPF-based synthesis stems from the independent Monte Carlo draws used to synthesize persons and families. For example, suppose that mutually fitted tables $\hat{\mathbf{N}}^P$ and $\hat{\mathbf{N}}^F$ are used with Monte Carlo to produce a complete population of persons and families $\hat{\mathbf{N}}'^P \in \mathbb{Z}$ and $\hat{\mathbf{N}}'^F \in \mathbb{Z}$. If it can be guaranteed for $g =$ *male* and $h =$ *husband-wife* that

$$\hat{\mathbf{N}}'^P_{ghi+k} = \hat{\mathbf{N}}'^F_{i+k} \tag{4.7}$$

(and likewise for $g =$ *female*), then a perfectly consistent set of connections between persons and families is possible. How can equation (4.7) be satisfied?

A simplistic solution would be a stratified sampling scheme: for each combination of i and k, select a number of individuals to synthesize and make exactly that many

draws from the subtables $\hat{\mathbf{N}}^P_{++i+k}$ and $\hat{\mathbf{N}}^F_{i+k}$. This approach breaks down when the number of strata grows large, as it inevitably does when more than one attribute is shared between persons and families.

The problem becomes clearer once the reason for mismatches is recognized. Suppose a Monte Carlo draw selects a family with husband age i in zone k. This random draw is not synchronized with the draws from the person population, requiring a person of age i in zone k to be drawn; the two draws are independent. Instead, synchronization could be achieved by *conditioning* the person population draws on the family population draws. Instead of selecting a random value from the joint distribution

$$P(\text{SEXP}, \text{CFSTRUC}, \text{AGEP}, \text{HLOSP}, \text{CTCODE})$$

of the person population, a draw from the conditional distribution

$$P(\text{HLOSP} \mid \text{SEXP} = \textit{male}, \text{CFSTRUC} = \textit{husband-wife}, \text{AGEP} = i, \text{CTCODE} = k)$$

could be used, and a similar draw for the wife. Converting the joint distribution generated by IPF to a conditional distribution is an extremely easy operation.

This reversal of the problem guarantees that equation (4.7) is satisfied, and allows consistent relationships to be built between agents. While it has been described here in a top-down manner (from family to person), it can be applied in either direction. The two approaches are contrasted in Figures 4.2 and 4.3.

4.4.3 Summary

As demonstrated in the preceding sections, it is possible to synthesize persons and relate them together to form families, while still guaranteeing that the resulting populations of persons and families approximately satisfy the fitted tables $\hat{\mathbf{N}}^P$ and $\hat{\mathbf{N}}^F$. By carefully choosing a set of shared attributes between the person and family agents and using conditional synthesis, a limited number of constraints can be applied to

```
1  for 1 ... N^F do
2      Synthesize a husband-wife family using a Monte Carlo draw ;
3      Synthesize a person, conditioning on AGEM, CFSTRUC, CTCODE and
       SEXP = male;
4      Synthesize a person, conditioning on AGEF, CFSTRUC, CTCODE and
       SEXP = female;
5  end
6  for 1 ... (N^P − 2N^F) do
7      Synthesize a person, conditioning on CFSTRUC ≠ husband-wife;
8  end
```

Figure 4.2: A top-down algorithm for synthesizing persons and husband-wife families.

```
1  for 1 ... N^P do
2      Synthesize a person using a Monte Carlo draw ;
3      if CFSTRUC = husband-wife then
4          if SEXP = male then
5              Synthesize a husband-wife family, conditioning on AGEM and
               CTCODE;
6              Synthesize a person, conditioning on AGEF, CFSTRUC, CTCODE and
               SEXP = female;
7          else
8              Synthesize a husband-wife family, conditioning on AGEF and
               CTCODE;
9              Synthesize a person, conditioning on AGEM, CFSTRUC, CTCODE and
               SEXP = male;
10         end
11     end
12 end
```

Figure 4.3: A bottom-up algorithm for synthesizing persons and husband-wife families.

the relationship formation process. In the example discussed earlier, the ages of husband/wife were constrained; in a more realistic example, the labour force activity of husband/wife, the number of children and the ages of children might also be constrained. Furthermore, multiple levels of agent aggregation could be defined: families and persons could be further grouped into households and attached to dwelling units.

The synthesis order for the different levels of aggregation can be varied as required, using either a top-down or bottom-up approach. However, the method is still limited in the types of relationships it can synthesize: it can only represent nesting relationships. Each individual person can only belong to one family, which belongs to one household. Other types of relationships cannot be synthesized using this method, such as a person's membership in another group (e.g., a job with an employer).

Chapter 5

Implementation

For the purposes of the ILUTE land use/transportation model, most of the improvements described in Chapter 4 seemed promising for the synthesis of a population of persons, families, households and dwelling units. A sparse data structure was used, a hierarchy of margins were used to help with random rounding, and conditional synthesis was used to link the different types of agents. The PUMS simplification procedure would increase the memory requirements of the sparse data structure, and was not employed. The projection method for dealing with random rounding was not deemed a significant improvement over the conventional IPF procedure, and was also not used.

A complete overview of the population synthesis procedure is shown in Figure 5.1. The numbered steps shown in the figure are:

1. a. Fit households/dwellings using PUMS and Summary Tables (using Beckman's multizone IPF approach).

   b. Fit persons using PUMS and Summary Tables.

2. Fit families using PUMS and Summary Tables; also fit to distributions of attributes shared with households/dwellings and persons.

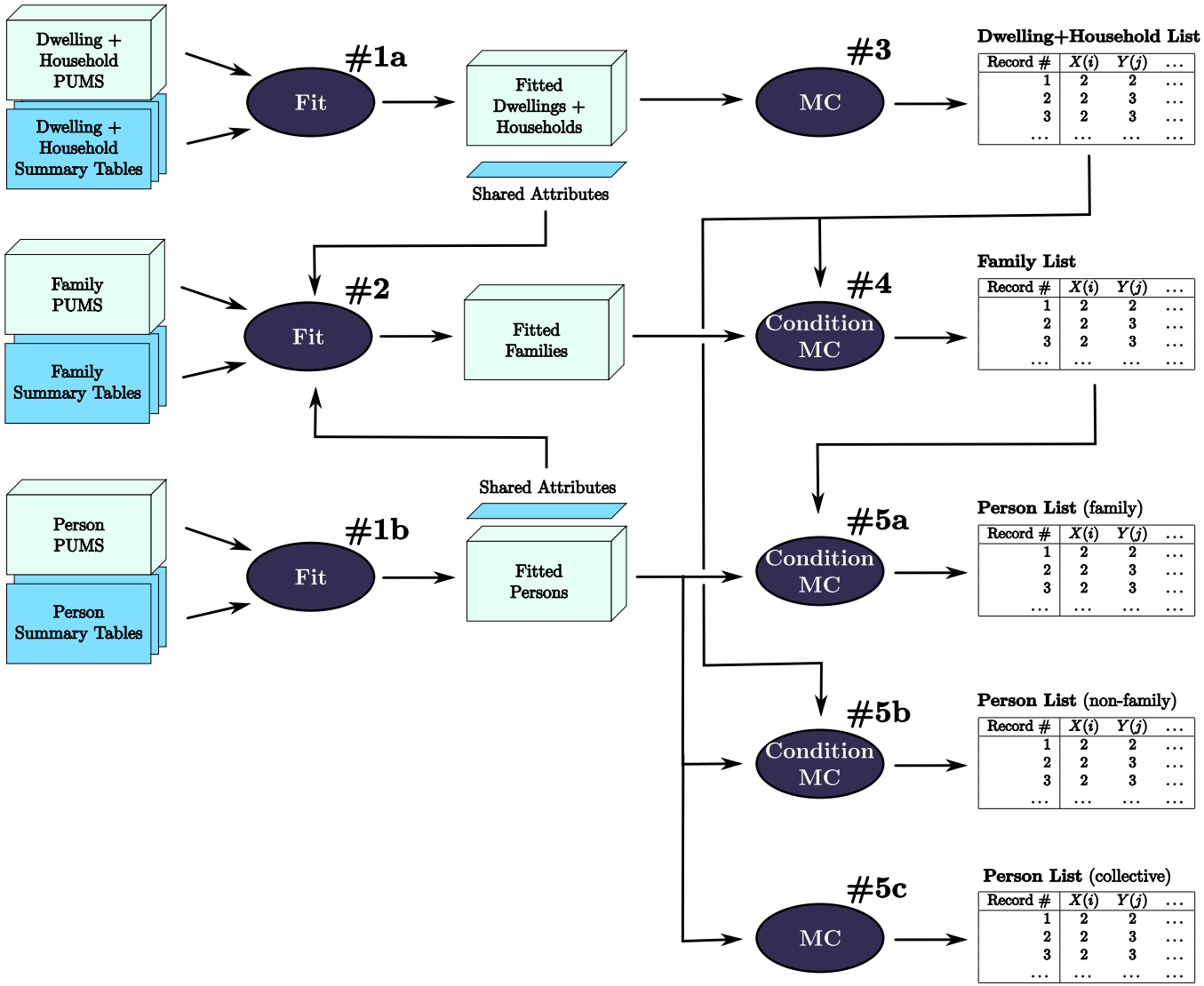

Figure 5.1: Overview of complete synthesis procedure. Numbers show the order of steps. On the left, PUMS and Summary Table data are combined using a fitting procedure (Beckman et al.'s multizone IPF). On the right, Monte Carlo is used to synthesize a list of individual agents from the fitted tables.

3. Use Monte Carlo to synthesize a list of households/dwellings.

4. For each household/dwelling with one or more families, synthesize family/families conditioned on household/dwelling characteristics.

5. a. For each family, synthesize persons conditioned on family characteristics.

   b. For each household/dwelling, synthesize non-family persons conditioned on household/dwelling characteristics.

   c. Use Monte Carlo to synthesize a list of foreign/temporary/collective (non-institutional) residents (not associated with a household/dwelling).

The method was implemented using special-purpose software written for the R/S+ statistical computing platform [29] with a few routines in C for additional speed. The following sections discuss the population universe, relationship model, population attributes, selection of shared attributes and software implementation.

5.1 Population Universe

The person, family and household universes are slightly reduced to match available data. No data is available on unoccupied dwellings, so only occupied dwellings are synthesized. This simplifies the dwelling/household relationship to a one-to-one mapping, allowing dwellings and households to be synthesized simultaneously. Almost no data is available on persons in institutions, so they are excluded from synthesis. Temporary, foreign and collective residents are included in most tables and are included in the synthesis for the purposes of accounting, but are not associated with any household, family or dwelling. For the fitting procedure, only persons 15 years of age and older are included, since most tables exclude younger persons. The conditional synthesis procedure does create persons under 15 years of age, but their only attributes are age and sex, since nothing further is available.

Finally, it is difficult to combine data from the 20% and 100% samples of the person universe. Most tables are on the 20% sample and exclude institutional residents, but the few that are defined on the 100% sample include the institutional residents. There is very little data on the institutional population, and they cannot always be removed from the 100% sample to match the 20% universe. Since more data is available on the 20% sample, it was used for synthesis, and the only 100% table used was CF86A04 (CFSTAT × AGEP × SEXP × CTCODE); DM86A01 was not used. The CF86A04 table was fitted to the 20% totals for AGEP × SEXP × CTCODE

For the family and household/dwelling synthesis, the 20% and 100% samples are defined on the same universe and are easier to combine. The 100% samples were used for both of these universes, which required a few 20% household table to be fitted to the 100% universe.

5.2 Relationship Model

The relationships synthesized between the different agents/objects are shown in Figure 5.2. Each household consists of zero or more census families, and zero or more non-family persons. There are approximately 28,000 multifamily households in the Toronto CMA, accounting for 2.3% of all households and 4.7% of the population. Multifamily households are not particularly desirable from a modelling standpoint; they were not contemplated as part of the original ILUTE prototype, and their behaviour would be challenging to model. Nevertheless, to properly account for persons and families during the synthesis of the dwellings, families and persons, multifamily households must be included. There is no data on exactly how many households contain more than two families, but it can be estimated as approximately 1,000 of the

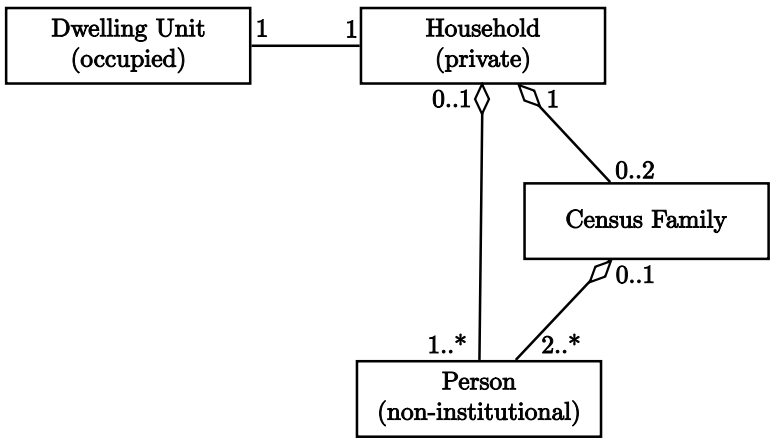

Figure 5.2: Diagram of the relationships synthesized between agents and objects, using the Unified Modelling Language (UML) notation [6]. Each line indicates a relationship, and the numbers at each end of the line show the "multiplicity", the number of agents/objects involved in the relationship. Edges with a diamond represent an aggregation relationship, where the diamond end is a "whole" and the other end is a "part." Thus, each household is composed of zero to two families, and conversely each family is a part of exactly one household.

28,000 multifamily households[1]. For the purposes of synthesis, these are treated as two-family households.

Some of the non-family persons in a household may still form an economic family, and be related to other household members; as described in Chapter 3, 3.9% of the Toronto CMA population are non-family persons living with relatives. However, there is very little data on these persons and on economic families in general, although a patchwork of information can be gleaned from the Person PUMS and the Household PUMS. Furthermore, the economic family is not a particularly useful unit to synthesize from a behavioural perspective. While census families make many decisions as a unit (e.g., moving home or buying/selling vehicles), economic families are less unified in their behaviour. Elderly parents or married children living with relatives may

[1]From the HH86A01 table, there are 849,950 one-family households and 27,720 multifamily households. Assuming 1,000 of these are three-family households, this gives 906,390 census families in total, quite close to the 906,385 total family count found in various family tables.

choose to change homes or vehicle ownership independent of the other members of their economic family. In light of its limited usefulness and importance for the rest of synthesis, economic families were excluded from synthesis. Persons living with relatives are treated the same as other non-family persons.

Finally, each census family contains two or more persons (at a minimum, either a husband and wife or a lone parent and child). These relationships between agents can also be examined in the reverse direction. Each person is a member of zero or one census family, and is a member of zero or one household; each family belongs to a single household. (Persons in collective dwellings and institutions are the only persons who do not belong to a household.) Each household occupies a single dwelling unit.

The relationships (and universes) used for synthesis may not be ideal for the actual microsimulation model. The existing ILUTE and TASHA models do not define families as an explicit agent, but instead include family relationships as part of the household agent; they also did not allow for multifamily households. It is admittedly difficult to build behavioural models at the family level; the definitions of family relationships are sufficiently complex that few data sources are collected on the family universe. Even if more data was available, it is unlikely that the family definitions would be sufficiently consistent to be useful. Similarly, multifamily households are rare enough (and complex enough) that activity diary data is not always adequate to model their behaviour.

The synthesis here only accounts for some of the agents needed for the ILUTE microsimulation. Some of the other agents, objects and relationships can easily leverage this initial synthesis: household-level vehicle ownership, for example, can be readily modelled once the household composition is known. The combined synthesis of household vehicle ownership and location of work for multiple-worker households remains an important challenge, however, given the limitations of available data.

Dwelling + Household	Census Family	Person
BUILTH (7)	AGEF (9)	AGEP (8)
DTYPEH (6)	AGEM (9)	CFSTAT (7)
HHNUEF (2)	CFSIZE (7)	HLOSP (9)
HHNUMCF (3)	CFSTRUC (3)	LFACT (4)
HHSIZE (8)	CHILDA (3)	OCC81P (16)
PAYH (5)	CHILDB (4)	SEXP (2)
PPERROOM (5)	CHILDC (3)	TOTINCP (13)
ROOM (9)	CHILDDE (9)	CTCODE (731)
TENURH (2)	HHNUMCF (2)	
CTCODE (731)	LFACTF (5)	
	LFACTM (5)	
	NUCHILD (9)	
	ROOM (9)	
	TENURE (2)	
	CTCODE (731)	

Table 5.1: Attributes and number of categories used during IPF fitting of three agent types. See Chapter 3 for comparison to categorization in source data, and see Appendix A for descriptions and further details.

5.3 Attributes

The attributes attached to each agent were largely selected based on the needs of the ILUTE model, plus a few additional attributes to help with linking agents to form relationships. As discussed in Chapter 3 these attributes are taken from both PUMS and Summary Table data. All summary tables discussed in Tables 3.3–3.5 were included in the synthesis except for the DM86A01 table (due to its inclusion of the institutional population) and the LF86B08 table. All margins of these summary tables were included to help with random rounding. For example, in the SC86B01 table, the four-way table AGEP $\times$ HLOSP $\times$ SEXP $\times$ CTCODE was applied as a margin, and all of its three-way, two-way and one-way margins were also applied as margins.

The categorization schemes in these data sources are often different, and some effort must be taken to establish suitable categorizations. A relatively fine categorization scheme was chosen for the source table during the IPF procedure, although not quite as fine as the PUMS categorization. The marginal tables generally had a coarser categorization for their attributes. To connect the two, mappings were constructed defining how the fine categories in the high-dimensional table could be collapsed to produce the coarser categorization in the marginal tables.

The final set of attributes synthesized during the IPF stage are shown in Table 5.1, along with the number of categories used in synthesis. Further details are shown in Appendix A.

5.4 Shared Attribute Selection

For any group of agents linked through a relationship, the agents' attributes need to satisfy certain constraints, precluding impossible agent relationships such as a mother who is younger than her child. The method described in Chapter 4 was used to ensure that a selected set of agent attributes are consistent and follow an observed probability distribution. In brief, the stages of the method are:

1. Select a set of attributes that are shared between two types of agents. Typically, attributes are selected to allow enforcement of behaviourally important constraints between agents.

2. Ensure that agents agree on the distribution of the shared attributes, possibly by fitting one population's contingency table against a margin of the other. As shown in Figure 5.1, the household/dwelling and person populations were fit first in this implementation. Margins for certain shared attributes were then taken from these tables, and applied as constraints when fitting the family population.

#	Agent	Attribute	Agent	Attribute	Notes
1	Household + Dwelling	CTCODE HHNUMCF HHSIZE ROOM TENURH	Family	CTCODE HHNUMCF CFSIZE ROOM TENURE	For family households where HHNUMCF > 0. Linkage between sizes is indirect.
2	Family	CTCODE CFSTRUC AGEF LFACTF	Person	CTCODE CFSTAT AGEP LFACT SEXP	For husband-wife or lone female parent families.
3	Family	CTCODE CFSTRUC AGEM LFACTM	Person	CTCODE CFSTAT AGEP LFACT SEXP	For husband-wife or lone male parent families.
4	Family	CTCODE CFSTRUC	Person	CTCODE CFSTAT AGEP	For children 15–17 in families where CHILDC > 0.
5	Family	CTCODE CFSTRUC	Person	CTCODE CFSTAT AGEP	For children 18+ in families where CHILDDE > 0.
6	Household + Dwelling	CTCODE	Person	CTCODE CFSTAT	For non-family persons, where HHSIZE $- \sum$ CFSIZE > 0.

Table 5.2: Summary of all attributes that are shared between agents to define and constrain relationships. The left agent and attributes are used to conditionally synthesize the right agent and attributes. For this to work, the distributions of these attributes must match in the fitted tables for both agents. Published tables are available for both agents for #4–6, but not for #1–3. Not shown: there are similar shared attributes for children under age 15 using CHILDA and CHILDB, but these persons are not part of the core person population.

3. Synthesize related agents by conditioning on shared attributes. As shown in Figure 5.1, this was done in a top-down manner in this implementation, starting with households/dwellings, conditionally synthesizing families from household/dwelling attributes, and then conditionally synthesizing family persons from family attributes.

This section focuses on the first step; the last two steps are described in detail in Chapter 4. The full set of shared attributes are shown in Table 5.2, and explained in the remainder of this section.

5.4.1 Households and Dwellings

The household/dwelling linkage was easy and automatic, thanks to the one-to-one relationship between occupied dwellings and households and the existence of a single PUMS combining both sets of attributes. Consistency between related household attributes (e.g., HHSIZE), dwelling attributes (ROOM) and combined attributes (PPERROOM) was automatic, since all data in the Household PUMS is consistent.

5.4.2 Families and Persons

The family/person linkage was fairly straightforward to select and construct. There are clear constraints between the family members that need to be preserved: for example, the age of the parents relative to the children and similarity in the parents' ages. To enforce such an age constraint, an age attribute must be present on both family and person agents, and the agents must agree on the distribution of ages. On the family agent, the attribute can be explicit like AGEF and AGEM (the husband/wife ages) or implicit like CHILDA (the number of children in the family of age 0–5).

The second obvious candidate for a constraint within the family is the labour force activity attribute. The presence of young children has a strong effect on the parents'

labour force activity, and the two parents' activity is correlated. As a result, AGEP, LFACT, SEXP and CFSTAT are the obvious candidates for linkage attributes, and are included (directly or indirectly) on both the family and person agents. This matches the set of constraints applied by Arentze & Timmermans [2] in their synthesis of households.

Other person attributes such as highest level of schooling (HLOSP) or occupation (OCC81P) are also likely to exhibit correlation between husband and wife, but are not deemed critical for the ILUTE model. For a transportation model, the travel to work associated with labour force activity is more critical. Because HLOSP and OCC81P are not treated as shared attributes, the association pattern between the husband and wife may not be accurate for these attributes.

5.4.3 Households/Dwellings and Families

The household/family linkage was the most challenging in this dataset. There were three primary options for performing the linkage, which could be used independently or combined:

1. **Household maintainer** demographics. The Household PUMS includes demographic information about a person self-designated as the maintainer, and the demographics of his/her spouse.

2. **Dwelling** characteristics such as the number of rooms and tenure. Data on rooms is present in both the Household and Family PUMS, and is in fact the only data in the Family PUMS related to household size.

3. **Financial** attributes such as the monthly rent/mortgage payments and the family income.

Initially, the household maintainer looked like an appealing link, since it would allow a single set of attributes to be shared between the three types of agents; perhaps

the maintainer's age and labour force activity could be carried throughout. However, the definition of the maintainer is too open-ended to be consistently useful. In 4.9% of households including census families, a child or non-family person is the maintainer; little or no demographic information about these persons is present in the Family PUMS, making linkage difficult. Additionally, in multifamily households the maintainer demographics only give information about one of the families.

Dwelling/household characteristics are more usable for linkage. Given the importance of the housing market to the ILUTE model, it is vital to ensure that families occupy legitimate dwellings, particularly homes that are large enough. The HHSIZE attribute combined with the ROOM attribute in the Household PUMS can ensure that the dwelling has enough rooms to accommodate the persons in the household. The Family PUMS includes a CFSIZE attribute; if it can be guaranteed that CFSIZE $\leq$ HHSIZE, then the family can fit in the dwelling. However, families can share rooms in a dwelling in a different manner from unrelated persons. The ROOM attribute is one of the few household/dwelling attributes present in the Family PUMS, and is the only data available showing how families use dwelling space differently from non-family households. Finally, the tenure TENURH also provides an important link with parents' ages. These two attributes were ultimately chosen to define the dwelling/family link, with an additional special constraint between ROOM, family size CFSIZE, HHSIZE and the number of families HHNUMCF.[2]

Financial attributes are also a possible link and a useful constraint, but were not pursued in this work. From a modelling standpoint, it would be valuable to be able to ensure that the members of a household have an income sufficient to pay the rent/mortgage for the dwelling they occupy. However, due to the large num-

[2]The details are a little complicated. After synthesizing a dwelling, a special conditional probability table is used to add a CFSIZE attribute using a Monte Carlo draw. The conditional probability is $P(\text{CFSIZE} \mid \text{ROOM}, \text{HHSIZE}, \text{HHNUMCF})$, and is calculated by reweighting the Person PUMS for family persons to the family universe. Finally, the dwelling with this additional attribute is used to synthesize the family, conditioning on the shared attributes ROOM, CFSIZE, TENURH, HHNUMCF and CTCODE.

ber of persons (both family and non-family) potentially involved in this relationship, it would likely be tricky to implement.

5.4.4 Households and Non-Family Persons

The final linkage is between household and non-family persons, and it is trivial: only the family status attribute on the person is used to link these two levels. Non-family persons are assumed to be independent of each other, and are hence synthesized independently and attached to the household.

There are a few constraints that would be useful to apply to non-family persons. Non-family persons under 15 years of age are more likely to live in a household that has at least one family, rather than living in a household of unrelated adults. Additionally, as discussed in Chapter 3, the census codes many same-sex couples as cohabiting non-family persons. The underlying data does not provide any information about the distribution of genders and ages of non-family persons sharing a dwelling, however, so no constraints can be applied.

5.5 Software

The population synthesis procedure was implemented in the R language [29]. R is a statistical computing platform whose syntax closely resembles S [3], but with an underlying implementation borrowed from the Scheme and Lisp languages. It was selected largely because of good performance, concise syntax, a good set of built-in routines for analyzing and visualizing categorical data and multiway contingency tables, and built-in log-linear and generalized linear models. While it was suitable for prototyping and experimenting with new methods, its data storage is not efficient for large amounts of data, and its performance is poorer than low-level languages like C.

The central components of the software are a sparse list-based implementation of

the Iterative Proportional Fitting algorithm, and a sparse list-based conditional Monte Carlo procedure.

5.5.1 IPF Implementation

The implementation of the Iterative Proportional Fitting procedure largely followed the description in Chapter 4. Its inputs include a list-based representation of a PUMS (in the R environment, this is called a *data frame*), a list of marginal constraints, a termination tolerance ϵ and an iteration limit. The marginal constraints are complete multiway contingency tables, which are associated with columns in the PUMS through the use of standardized variable names. Each constraint can also include a category mapping scheme, defining how the PUMS categories need to be collapsed in order to match the category system used by the margin.

Marginal constraints are applied in series, in the conventional manner for IPF. This does mean that the result is slightly dependent on the order that the constraints are applied; typically, the final constraint achieves perfect fit while earlier constraints do less well. Dykstra's suggestion of a parallel update procedure [17] is worth considering as an alternative.

A small part of the IPF procedure was implemented in C for performance reasons: collapsing the sparse list down to the marginal dimensions, and applying the marginal update back to the weights in the sparse list. The R language provided adequate performance for the other parts of the procedure.

5.5.2 Random Rounding and Area Suppression

To deal with random rounding, the modified IPF termination criterion described in Chapter 4 was employed. Additionally, the full hierarchy of margins was used to reduce rounding error in aggregate tables.

The data did include some area suppression, but a small amount of data was available to estimate the bare minimum information for these zones: the total population. The suppressed areas were assumed to follow the PUMA average distribution for each margin, scaled to the appropriate total population.

5.5.3 Conditional Monte Carlo

As discussed in Chapter 4, ordinary Monte Carlo synthesis can easily be implemented using a sparse data structure, and conditional synthesis is only slightly more complicated. Suppose attributes X and Z are given, and attribute Y needs to be synthesized using a joint probability distribution $P(X, Y, Z)$. Then, the formula for conditional probability is

$$P(Y \mid X, Z) = \frac{P(X, Y, Z)}{P(X, Z)}. \tag{5.1}$$

In order to make a draw from $P(Y \mid X, Z)$, it must be possible to find the contributing cells of $P(X, Y, Z)$ efficiently. This is not automatic when using a list-based data structure, since random access to the rows associated with a particular cell (i, j, k) is not efficient. To deal with this, the list was sorted by the given attributes. This makes it easy to find the rows associated with a particular cell, with asymptotic performance of $\mathcal{O}(\log n)$.

The rest of the algorithm was simple to implement, and the complete details are shown as pseudocode in Figure 5.3. The overall performance is $\mathcal{O}(N \log n)$, and the operation was also implemented in C to improve performance.

Some authors have used other versions of Monte Carlo, such as drawing without replacement [26, 28]. In such approaches, after making draw a particular agent from a table of counts, the corresponding cell is decremented by 1 to prevent synthesis of too large a number of persons of any particular type.

These techniques have little or no value for this dataset, because the number of cells

Step	Description	Time (min.)
Multizone IPF		
1a	Households/dwellings	30.4
1b	Persons	58.9
2	Families	10.3
Subtotal		*1:45.5*
Monte Carlo		
3	Households/dwellings	0.9
4	Families	3.6
5a	Persons (family)	10.9
5b	Persons (non-family)	3.2
5c	Persons (collective)	0.0
Subtotal		*21.8*
Overhead		9.2
Total		2:07.3

Table 5.3: Computation time for the different stages of the synthesis procedure on a 1.5GHz computer for the Toronto Census Metropolitan Area. Step numbers refer to the stages shown in Figure 5.1.

with counts greater than or equal to 1.0 is very small; almost all cells have fractional counts less than 1. For example, in the population of 2.7 million persons, only 20,090 persons are synthesized from cells with counts greater than or equal to 1.0.

5.6 Results

The final population was synthesized for the Toronto Census Metropolitan Area using the associated PUMS datasets. The compute times for population synthesis are substantial, but not extravagant. As shown in Figure 5.3, the synthesis required two hours and seven minutes to complete on an older 1.5 GHz computer with 2GB of memory. Synthesis of this duration is not a major issue since it can be performed once before a set of ILUTE model runs (or once per run, if different populations are desired), and

the ILUTE model itself is considerably more compute-intensive.

Finally, the process was repeated for other CMAs using their own PUMS data: the Hamilton CMA was synthesized together with the Kitchener and Niagara-St. Catharines CMAs (since these three CMAs had a single shared PUMS in 1986), and the Oshawa CMA was also synthesized. Oshawa did not have its own PUMS in 1986, so the Toronto PUMS was used instead. Together, these three CMAs form the Greater Toronto/Hamilton Area, the urban region that the ILUTE project aims to study.

Using this population, any number of cross-tabulations and maps can be produced. To give a sense of the geography, Figure 5.4 shows a map of the median number of rooms in the dwelling units in each census tract in the Toronto CMA. This data is not available in any existing summary tables, although one table shows household size by zone and another shows persons-per-room by zone. Without any ground truth, the result cannot be verified, but it does match local general knowledge of dense and/or high-rise neighbourhoods. In particular, the zones with the lowest median number of rooms (smallest dwellings) are known to contain a large number of tall apartment buildings (often social housing) or student residences. One surprising zone with a median of 3 rooms per dwelling occurred in rural Niagara, but proved to contain largely "movable dwellings," which are otherwise rare in the Toronto area.

Input: List $\mathbf{W}$ contains a joint distribution of attributes $X(i), Y(j), Z(k)$ in sparse list format. Each row r contains a co-ordinate for X and Y and weights for the K possible values of Z, i.e. $\mathbf{W}_{r\cdot} = \{i, j, w_1, w_2, \ldots, w_K\}$. There is one row for each entry in the PUMS. List $\mathbf{A}$ contains a preliminary population of agents with the given attributes X and Z already defined. Row a contains $\mathbf{A}_{a\cdot} = \{i, k\}$.

Output: List of complete agents $\mathbf{A}'$ equal to $\mathbf{A}$ but with a new column defining j

`// Ensure that identical values of given attribute` $X(i)$ `are in adjacent rows.`

1 Sort rows of $\mathbf{W}$ by attribute i;

2 **foreach** *row* $\mathbf{A}_{a\cdot} = \{i, k\}$ *of* $\mathbf{A}$ **do**

`// The rows between` r_1 `and` r_2 `are the candidates for synthesis given the known attribute value` i.

3 r_1, r_2 = first and last rows in $\mathbf{W}$ containing $X = i$, found using a binary search;

`// Vector w contains the weights associated with each candidate row given` i `and` k.

4 $\mathbf{w}$ = column of $\mathbf{W}$ corresponding to w_k, restricted to rows between r_1 and r_2;

`// Convert to a probability mass function.`

5 $\mathbf{p} = \mathbf{w} / \sum \mathbf{w}$;

6 r = random row in range $[r_1, r_2]$ selected using a Monte Carlo draw from $\mathbf{p}$;

7 $\mathbf{A}'_{a\cdot} = \{i, j, k\}$ where j is taken from row r of $\mathbf{W}$;

8 **end**

Figure 5.3: Algorithm showing conditional Monte Carlo synthesis using a sparse list-based data structure. Attribute $Y(j)$ is synthesized given known attributes $X(i)$ and $Z(k)$. Attributes X and Y are from a PUMS source, while Z is a non-PUMS variable (e.g., geographic zone). The method can be easily generalized to a large number of attributes.

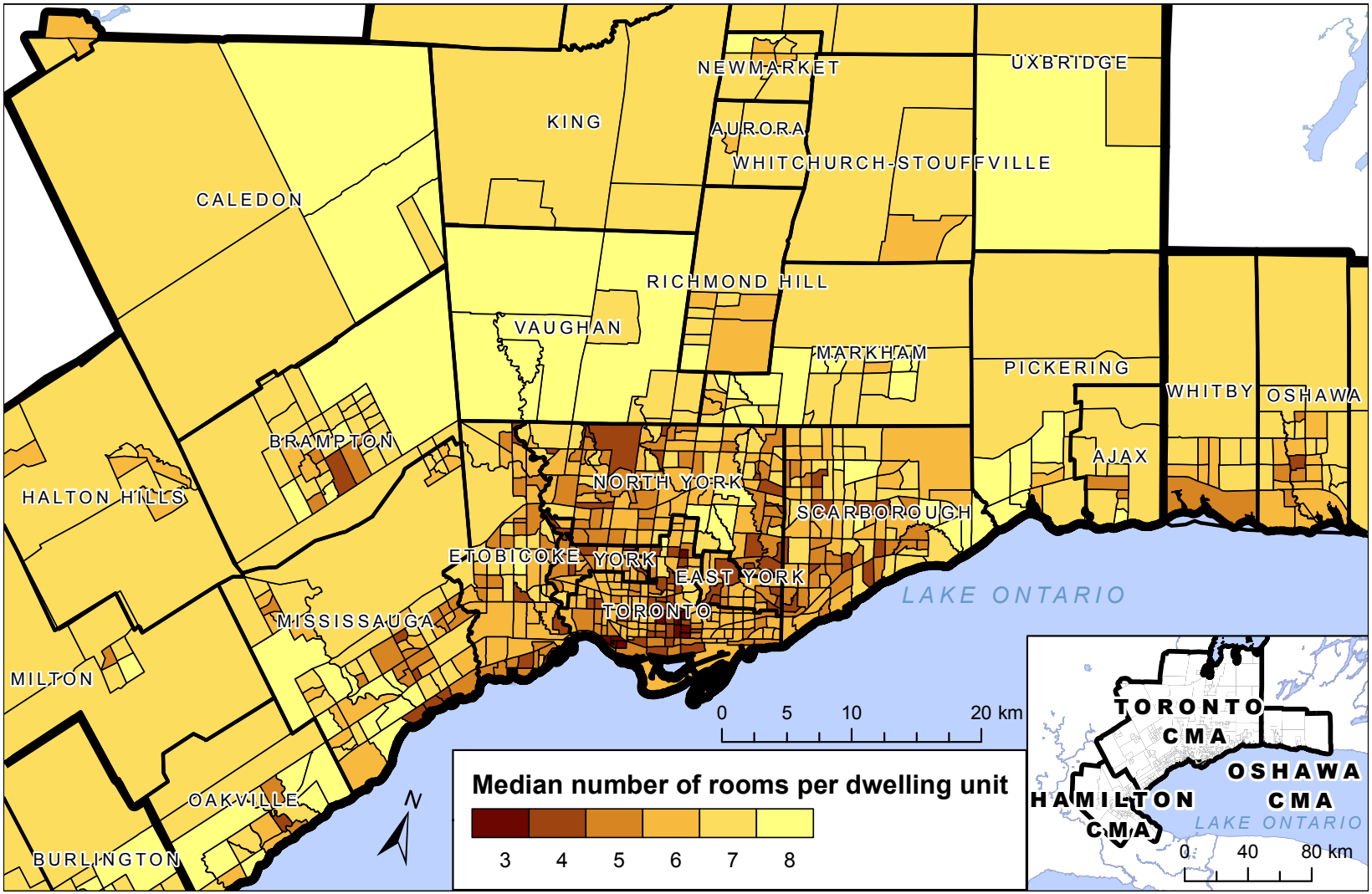

Figure 5.4: Map showing a dwelling attribute from the synthesized population.

Chapter 6

Evaluation

It is challenging to evaluate the results of a data synthesis procedure. If any form of complete "ground truth" were known, the synthetic population could be tested for goodness-of-fit against the true population's characteristics; but instead only partial views of truth are available in smaller, four-way tables.

In theory, IPF-based procedures have many of the qualities necessary for a good synthesis: an exact fit to their margins, while minimizing the changes to the PUMS (using the discrimination information criterion). This does not mean that the full synthesis procedure is ideal: the fit may be harmed by conflicting margins (due to random rounding), and will almost certainly be poorer after Monte Carlo (or conditional Monte Carlo). Furthermore, it still leaves a major question open: how much data is sufficient for a "good" synthesis? Are the PUMS and multidimensional margins both necessary, or could a good population be constructed with one of these two types of data? Does the multizone method offer a significant improvement over the zone-by-zone approach?

To answer these questions, a series of experiments was conducted. In the absence of ground truth, each synthetic population is evaluated in terms of its goodness-of-fit to a large collection of low-dimensional contingency tables. These validation tables

are divided into the following groups:

1. One-dimensional margins for the entire PUMA, for each attribute.
2. One-dimensional margins by zone for each attribute.
3. Higher-dimensional Summary Tables for the entire PUMA.
4. Higher-dimensional Summary Tables by zone.
5. Higher-dimensional margins from PUMS that are unavailable in summary tables. A selection of 2D and 3D margins are taken from the PUMS after fitting each to the 1–3D margins in the Summary Tables.

The complete list of tables in each group is shown in Table B.1. The evaluation was performed using a single PUMA, the Toronto Census Metropolitan Area, and excluded the Hamilton and Oshawa CMAs used for the final ILUTE synthesis.

6.1 Goodness-of-Fit Measures

After cross-classifying the synthetic population to form one table $\hat{\mathbf{N}}_{ijk}$, it can be compared to a validation table $\mathbf{N}_{ijk}$ using various goodness-of-fit statistics. This is repeated for each of the validation tables in turn, and the goodness-of-fit statistics in each group are then averaged together to give an overall goodness-of-fit for that group.

The choice of evaluation statistic is challenging, with many trade-offs. Knudsen & Fotheringham provided a good and even-handed overview of different matrix comparison statistics [31], framed in the context of models of spatial flows, but applicable to many other matrix comparison problems. They reviewed three categories of statistics: information theoretic, generalized distance, and traditional statistics (such as R^2 and χ^2). In a comparison of the statistics, their ideal was "one for which the relationship between the value of the statistic and the level of error is linear," and using this

benchmark they found that the Standardized Root Mean Square Error (SRMSE) and $\bar{\Psi}$ were the "best" statistics. The former is a representative distance-based statistic, while the latter is an unusual information theoretic statistic. As Voas & Williamson noted, $\bar{\Psi}$ is actually very little different from another distance-based statistic, total absolute error [53].

$$\text{SRMSE} = \frac{\sqrt{\frac{1}{IJK}\sum_{i,j,k}(\hat{N}_{ijk} - N_{ijk})^2}}{\frac{1}{IJK}\sum_{i,j,k}N_{ijk}} \tag{6.1}$$

$$\bar{\Psi} = \sum_{i,j} N_{ijk}\left|\log\frac{N_{ijk}}{(N_{ijk}+\hat{N}_{ijk})/2}\right| + \sum_{i,j}\hat{N}_{ijk}\left|\log\frac{\hat{N}_{ijk}}{(N_{ijk}+\hat{N}_{ijk})/2}\right| \tag{6.2}$$

However, Knudsen & Fotheringham's definition of an "ideal" metric is somewhat questionable. True information theoretic measures are supposed to have deep statistical underpinnings, representing the information content of a probability distribution. The Minimum Discrimination Information statistic is equivalent to G^2:

$$\begin{aligned}\text{MDI} = G^2 &= 2NI(\mathbf{N}\|\hat{\mathbf{N}}) \\ &= 2\sum_{ijk} N_{ijk}\log\frac{N_{ijk}}{\hat{N}_{ijk}}\end{aligned}$$

It does not measure goodness-of-fit *per se*, but rather measures the amount of information of a cross-tabulation. Additionally, when testing fit to multiple tables with different sample sizes, the G^2 statistic gives greater weight to large-sample tables. (For example, when comparing the fit to a 100% Summary Table, a 20% Summary Table and a 2% PUMS-only table, the G^2 statistic would be scaled by 1, 0.2 and 0.02 respectively, to account for the lower actual sample size of these tables.) For these reasons, the G^2 statistic does offer compelling advantages over the other statistics. (The other information theoretic statistics—ϕ, Ψ and $\bar{\Psi}$—lack the theoretical underpinnings of G^2.)

An example comparing the two types of statistics is shown in Table 6.1. In the experiment shown, the population was fitted using a zone-by-zone method, with all

	Validation Tables N		Fitted Table $\hat{\mathbf{N}}$	
Group of Validation Tables	Average # of Cells	Average null model G^2	Average G^2	Average SRMSE $\times 100,000$
1. 1D STs (entire PUMA)	97	211819	2	92
2. 1D STs (by zone)	4699	369388	102	146
3. 2–3D STs (entire PUMA)	26	423265	20	297
4. 2–3D STs (by zone)	18777	529606	1599	241
5. 2–3D (only in PUMS)	604	105583	72	5580

Table 6.1: Comparison of G^2 and SRMSE statistics for validation. The left two columns show statistics on the groups of validation tables themselves: the number of cells and the G^2 of the table relative to a null model, averaged over the group. For the right two columns, a zone-by-zone IPF fit was conducted (experiment I8) and two different goodness-of-fit statistics were applied, the information theoretic Minimum Discrimination Information (G^2) statistic and the distance-based Standardized Root Mean Square Error. SRMSE is scaled by 100,000 to allow comparison.

available Summary Tables applied as margins (identical to experiment I8 in the following section). A good fit is expected in the first four groups of validation tables, and a reasonable fit is expected for the final group since the initial table was the complete PUMS. In terms of *fit*, the SRMSE statistic matches expectations. In terms of *information*, the G^2 statistic shows a huge improvement over a null model; in other words, most of the information present in the tables is explained by the fitted population. However, using the G^2 statistic, the poorest group of validation tables is not group five but group four (*2–3D STs by zone*); these tables are where most of the missing information lies.

Nevertheless, distance-based statistics are more widespread in the literature, and have been reported for many other population synthesis applications. For these reasons, the SRMSE statistic is used as the primary evaluation metric here. It is scaled by 1000 throughout, rather than 100,000 as above.

Finally, it would be useful to also be able to apply traditional statistical tests to

compare different models. In particular, tests such as the Akaike Information Criterion (AIC) which reward parsimonious low-parameter models would be interesting to apply. However, because the data is sparse, it is difficult to determine the number of degrees of freedom and the number of free parameters during Iterative Proportional Fitting. Without this information, statistical tests are not possible.

6.2 Tests of IPF Method and Input Margins

In the first series of experiments, the IPF procedure is tested with different inputs to see how the quality of fit is affected. Three questions are tested simultaneously:

- **Source Sample**: How does the initial table in IPF affect the result? Can a good fit be obtained with a constant initial table, or is the PUMS necessary?
- **1D Margins**: Are 1D margins sufficient, or does a better fit result when 2D and 3D margins are applied?
- **Geography**: What is the difference between the zone-by-zone and multizone approach to geographic variation?

To test these hypotheses, a set of ten fits was conducted, labelled I1 through I10. Essentially, the experiments evaluate these three different questions, showing the impact of different source samples, 1D versus 2–3D margins, and three different approaches to geography. The input data included in each experiment are shown together with the output goodness-of-fit in Table 6.2. The first set of experiments (I1–I4) show the results with no geographic input data, and are largely intended as a "base case" to show the effect of better data. Experiments I5–I8 show a zone-by-zone IPF method, where each zone is fitted independent of the others. I6 represents a "typical" application of IPF for population synthesis: a zone-by-zone approach using 1D margins. Finally, I9

	Experiment									
	Almost no geography				Zone-by-zone				Multizone	
	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
Input: Margins										
1. 1D STs (entire PUMA)	✓	✓	✓	✓					✓	✓
2. 1D STs (by zone)					✓	✓	✓	✓	✓	✓
3. 2–3D STs (entire PUMA)			✓	✓						✓
4. 2–3D STs (by zone)							✓	✓		✓
5. PUMS									✓[1]	✓[2]
Input: Source Sample	1	PUMS	1	PUMS	1	PUMS	1	PUMS	1	1
Output: SRMSE × 1000, averaged over group of validation tables										
1. 1D STs (entire PUMA)	1	0	1	1	4	1	2	1	0	0
2. 1D STs (by zone)	285	285	285	285	2	3	2	1	3	1
3. 2–3D STs (entire PUMA)	192	15	2	2	187	19	7	3	15	0
4. 2–3D STs (by zone)	566	522	522	522	252	130	3	3	131	3
5. 2–3D (only in PUMS)	883	38	735	6	849	73	659	56	38	0

[1] PUMS fitted to 1D margins [2] PUMS fitted to 2–3D margins

Table 6.2: Design and results of experiments I1–I10, testing goodness-of-fit of IPF under varying amounts of input data. Each column shows a single experiment, including the input data applied (top); and goodness-of-fit after IPF (bottom). Further details can be found in Appendix B.

and I10 show the multizone approach suggested by Beckman et al., where the PUMS is applied as a marginal constraint.

The combinatorial optimization method is not IPF-based, but it does operate on a zone-by-zone basis. It is not possible to determine which of the experiments I5–I8 is "closest" to combinatorial optimization. While combinatorial optimization starts with a random sample of the PUMS, it does not guarantee that the final result fits the PUMS associations (validation table group 5); it almost certainly has a poorer fit to group 5 than I6 or I8. Furthermore, the method does have some convergence issues when applying a large number of constraints, and it is not clear whether the full set of 2–3D constraints could be applied in a practical implementation. As a result, the combinatorial optimization method might give a fit close to any of I5–I8, or possibly even poorer than I5; without a direct comparison, little can be said.

6.2.1 Source Sample

To evaluate the effect of source sample, compare the use of a constant initial table filled with ones (I1, I3, I5 and I7) to the use of the PUMS as the initial table (I2, I4, I6 and I8).[1] Because a sparse IPF procedure is used, the initial table necessarily has the *sparsity pattern* of the PUMS in all cases. The constant initial table's non-zero cells were all ones, while the PUMS initial table initially had many higher integer cells.

In all cases, the use of the PUMS for the initial table drastically improves the fit to validation group 5 (*2–3D PUMS-only*) by at least an order of magnitude. In experiments I2 and I6 where no 2–3D margins are applied, the use of PUMS similarly improves the fit to validation group 3 (*2–3D STs, entire PUMA*). However, the improvement is considerably smaller on validation group 4 (*2–3D STs by zone*).

All of these are expected results. The only interesting finding is for I6: there are

[1] Experiments I9 and I10 are excluded from this discussion, since they use the PUMS as a margin instead of using it as the starting table.

geographical variations in the 2–3D association pattern that are not explained by the combination of 1D margins with geography and the PUMS.

6.2.2 1D Margins versus 2–3D Margins

To observe the effect of higher-dimensional margins, compare the experiments with 1D margins (I1, I2, I5, I6, I9) against the experiments with 2–3D margins (all others). From a brief glance at the results, it is quickly evident that there is multiway variation in the data that is not explained unless either the PUMS or 2–3D margins are applied. For validation groups 1 and 3 (*1D and 2–3D STs, entire PUMA*), the difference between inputting the PUMS (I2, I6, I9) or the 2–3D margins (I3, I7, I10) is fairly small. The main difference is due to sample size: the PUMS is a 2% sample, while the margins are drawn from 20% samples.

The primary benefit of including 2–3D margins appears to be the ability to capture geographic variation in these 2–3 way relationships. The goodness-of-fit against these validation tables remains poor until the 2–3D Summary Tables by zone are included in I7, I8 and I10.

6.2.3 Zone-by-zone versus Multizone

The difference between the zone-by-zone and multizone methods was surprisingly small. While the zone-by-zone method makes no attempt to explicitly fit validation groups 1 and 3 (*1D and 2–3D STs, entire PUMA*), it still seems to achieve a fairly good fit.

The only improvement recorded by the multizone method is in the fit to validation group 5 (*2–3D, PUMS only*): I9 does better than I6 on this score, and I10 likewise does better than I8. Nevertheless, the difference is fairly small.

The reasons for the difference lie in the contrasting approaches to the PUMS. In

	Experiment			
	R1	R2	R3	R4
Zone-by-Zone (like I8)	✓	✓		
Multizone (like I10)			✓	✓
Hierarchical Margins		✓		✓
Output: SRMSE × 1000, averaged over tables				
1. 1D STs (entire PUMA)	3	1	0	0
2. 1D STs (by zone)	7	2	7	2
3. 2–3D STs (entire PUMA)	6	3	0	0
4. 2–3D STs (by zone)	7	3	7	3
5. 2–3D (only in PUMS)	58	56	0	0

Table 6.3: Design and results of experiments R1–R4, testing goodness-of-fit after using different methods to deal with random rounding. Each column shows a single experiment, including the input data applied (top) and goodness-of-fit after IPF (bottom).

many cases, the fit to the initial table decreases as more margins are included in the IPF. The additional margins show up as more terms in the log-linear model and appear as additional free parameters during the fitting process, giving the fitted table more freedom to vary from the source table. Deterioration of fit to the source table can be clearly seen by comparing I2 to I6 or I4 to I8. The I6 experiment added new tables with geographic variation; however, this does not improve the fit to the PUMA-wide 2–3D PUMS-only tables (validation group 5). Indeed, the fit to these tables deteriorates due to the addition of parameters.

By contrast, the multizone approach forces a fit to the PUMS, treating it as a margin with equal importance to the other constraints. As a result, it achieves a better fit to validation group 5.

6.3 Effects of Random Rounding

In a second series of experiments, the effects of random rounding were tested. Table 6.3 shows the results of four experiments. The first two (R1 and R2) show the effects of hierarchical margins when using a zone-by-zone algorithm, and the second two (R3 and R4) show the effects when using a multizone algorithm.

As shown, there is a small improvement in fit when using hierarchical margins. The improvement of fit at the zonal level is somewhat surprising—after all, when hierarchy is not used, the only input margins are the zonal tables. However, the improvement comes due to conflicts between tables sharing the same attributes. In R1 and R3, the fit to the final zonal table is perfect, but the other zonal tables have a poorer fit than in R2 and R4.

Overall, this suggests that hierarchical margins are useful, but their impact on goodness-of-fit is relatively small.

6.4 Effects of Monte Carlo

In a third series of experiments, the effects of the Monte Carlo integerization procedures are tested. The design and results of these experiments are shown in Table 6.4. The first experiment M0 is the null case: the results of the IPF procedure before any Monte Carlo integerization takes place. Experiment M1 shows the conventional Monte Carlo procedure, where a set of persons are synthesized directly from the IPF-fitted tabulation for persons. Experiment M2 is the conditioned Monte Carlo procedure described in Chapter 5, where households/dwellings are synthesized by Monte Carlo, families are conditionally synthesized on dwellings, and persons are conditionally synthesized on families. The results are evaluated on the person population only, to focus on the effects of the two stages of conditioning prior to generating the persons.

	Experiment		
	M0	M1	M2
Multizone IPF Fit (I10)	✓	✓	✓
Monte Carlo (Person)		✓	
Conditional Monte Carlo (Person | Family | Dwelling)			✓
Output: SRMSE × 1000, averaged over group of Person tables			
1. 1D STs (entire PUMA)	0	3	8
2. 1D STs (by zone)	1	39	58
3. 2–3D STs (entire PUMA)	0	3	7
4. 2–3D STs (by zone)	3	80	99
5. 2–3D (only in PUMS)	0	12	21

Table 6.4: Design and results of experiments M0–M2, testing goodness-of-fit after applying different Monte Carlo methods. Each column shows a single experiment, including the input data applied (top) and goodness-of-fit after IPF (bottom). The Monte Carlo procedure is non-deterministic, so a series of 30 runs were performed, with the average error shown here for M1 and M2.

As expected, the goodness-of-fit deteriorates after applying Monte Carlo, and deteriorates further using the conditional procedure. The deterioration from M0 to M1 is somewhat larger than the deterioration from M1 to M2. In essence, this shows that the conditional synthesis procedure employed here does not have a major impact on the goodness-of-fit. Even after two stages of conditioning (from dwellings to families to persons), a reasonable goodness-of-fit is maintained.

Additionally, among the tables in validation group 1 (*1D STs, entire PUMA*) in Table 6.4, there is one clear outlier: the fit to CTCODE had an average SRMSE×1000 of 16 for M1 and 30 for M2. This poor fit—and the poorer fits to validation group 2 (*1D STs, by zone*)—might be corrected by stratifying the Monte Carlo synthesis by zone, although this could cause a deterioration of the fits to the non-geographic tables.

Chapter 7

Conclusion

After conducting the experiments in Chapter 6, it appears that using all available data is worthwhile; the multizone method offers small benefits over the zone-by-zone method; and hierarchical margins offer a very small benefit for addressing random rounding issues.

The next stage in the ILUTE synthesis effort will link this population with other data sources to synthesize the vehicles owned by each household, the place-of-work of the household members who are active in the labour market, and the business establishments that provide employment.

In conclusion, several of the problems of existing population synthesis procedures were successfully resolved in this research. The first major contribution is a sparse list-based Iterative Proportional Fitting procedure that combines the advantages of IPF and reweighting: an entropy-maximizing procedure that preserves the association pattern in the PUMS while fitting a set of disparate marginal distributions, making possible a large set of agent attributes with fine categorization. This technique produces results identical to the IPF procedure, but makes more efficient use of memory and time when a large number of attributes are synthesized.

The second major contribution was a technique for synthesizing relationships be-

tween agents using IPF and conditional probabilities. This allows persons to be grouped into aggregations such as families and households while fitting known distributions at the person, family and household level, and enforcing a limited number of constraints between the members of an aggregation. The results show that these relationships can be synthesized with only a minimal impact on the fit at any single level.

Bibliography

[1] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, New York, 2nd edition, 2002.

[2] Theo A. Arentze and Harry J.P. Timmermans. ALBATROSS version 2: A Learning-Based Transportation Oriented Simulation System. Technical report, Eindhoven University of Technology, European Institute of Retailing and Services Studies, Eindhoven, 2005.

[3] Rick A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language*. Wadsworth, Pacific Grove, CA, 1988.

[4] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. Creating synthetic baseline populations. *Transportation Research A*, 30(6):415–435, 1996.

[5] Richard J. Beckman, B.W. Bush, K.M. Henson, and P.E. Stretz. Portland study synthetic population. Technical Report LA-UR-01-4610, Los Alamos National Laboratory, Los Alamos, NM, August 2001.

[6] Grady Booch, James Rumbaugh, and Ivar Jacobson. *The Unified Modelling Language User Guide*. Addison-Wesley, Reading, MA, 1999.

[7] Jean-René Boudreau. Data swapping is not the panacea. In *Proceedings of Statistics Canada's Symposium 2005*, Ottawa, 2005.

[8] Claus Boyens, Oliver Günther, and Hans-Joachim Lenz. Statistical databases. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics*, chapter 9, pages 267–292. Springer, Berlin, 2004.

[9] Graham P. Clarke, editor. *Microsimulation for Urban and Regional Policy Analysis*, volume 6 of *European Research in Regional Science*. Pion, London, 1996.

[10] Clifford C. Clogg and Scott R. Eliason. Some common problems in log-linear analysis. *Sociological Methods Research*, 16(8), 1987.

[11] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1st edition, 1990.

[12] Lawrence H. Cox. On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, 117(2):251–273, 2003.

[13] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–159, February 1975.

[14] Imre Csiszár. Information theoretic methods in probability and statistics (transcript of the 1997 Shannon Lecture). *IEEE Information Theory Society Newsletter*, March 1998.

[15] Juan de Dios Ortúzar and Luis G. Willumsen. *Modelling Transport*. John Wiley & Sons, Chichester, UK, 3rd edition, 2002.

[16] W. Edwards Deming and Frederick F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4):427–444, December 1940.

[17] Richard L. Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Annals of Probability*, 13(3):975–984, 1985.

[18] Federal Committee on Statistical Methodology. Report on statistical disclosure limitation methodology. Working Paper 22, Office of Management and Budget, Executive Office of the President of the United States, Washington, D.C., December 2005.

[19] Stephen E. Fienberg. Log-linear models. In Samuel Kotz, Campbell B. Read, N. Balakrishnan, and Brani Vidakovic, editors, *Encyclopedia of Statistical Sciences*. John Wiley, New York, 2nd edition, 2004.

[20] Stephen E. Fienberg and Michael M. Meyer. Iterative proportional fitting. In Samuel Kotz, Campbell B. Read, N. Balakrishnan, and Brani Vidakovic, editors, *Encyclopedia of Statistical Sciences*. John Wiley, New York, 2nd edition, 2004.

[21] Martin Frick and Kay W. Axhausen. Generating synthetic populations using IPF and Monte Carlo techniques: Some new results. In *Proceedings of the 4th Swiss Transport Research Conference*, Monte Verità, Switzerland, March 2004.

[22] Michael Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, March 1994.

[23] Kenneth P. Furness. Time function iteration. *Traffic Engineering Control*, 7(11):458–460, November 1965.

[24] James E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York, 2nd edition, 2003.

[25] Junfei Jeffrey Guan. Synthesizing family relationships between individuals for the ILUTE micro-simulation model. B.A.Sc. thesis, University of Toronto, Department of Civil Engineering, 2002.

[26] Jessica Y. Guo and Chandra R. Bhat. Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014:92–101, 2007.

[27] Antoine Hobeika. Population synthesizer. In *TRANSIMS Fundamentals*, chapter 3. U.S. Federal Highway Administration, Travel Model Improvement Program, Washington, D.C., 2005.

[28] Zengyi Huang and Paul Williamson. Comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Working Paper 2001/2, University of Liverpool, Department of Geography, Population Microdata Unit, Liverpool, October 2001.

[29] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[30] C. Terrence Ireland and Solomon Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, March 1968.

[31] Daniel C. Knudsen and A. Stewart Fotheringham. Matrix comparison, goodness-of-fit, and spatial interaction modelling. *International Regional Science Review*, 10(2):127–147, 1986.

[32] Michael L. Lahr and Louis de Mesnard. Biproportional techniques in Input-Output analysis: Table updating and structural analysis. *Economic Systems Research*, 16(2):115–134, 2004.

[33] Roderick J.A. Little and Mei-Miau Wu. Models for contingency tables with known marginals when target and sampled populations differ. *Journal of the American Statistical Association*, 86(413):87–95, March 1991.

[34] Eric J. Miller, David S. Kriger, and John Douglas Hunt. Integrated urban models for simulation of transit and land use policies: guidelines for implementation and use. TCRP Report 48, Transit Cooperative Research Program, Transportation Research Board, Washington, D.C., 1998.

[35] Eric J. Miller and Matthew J. Roorda. A prototype model of 24-hour household activity scheduling for the Toronto Area. *Transportation Research Record*, 1831:114–121, 2003.

[36] Eric J. Miller, Matthew J. Roorda, and Juan A. Carrasco. A tour-based model of travel mode choice. *Transportation*, 32(4):399–422, July 2005.

[37] Daniel A. Powers and Yu Xie. *Statistical Methods for Categorical Data Analysis*. Academic Press, Toronto, 2000.

[38] Justin Ryan, Hannah Maoh, and Pavlos Kanarogolou. Population synthesis: Comparing the major techniques using a small, complete population of firms. Working Paper 026, McMaster University, Centre for Spatial Analysis, Hamilton, ON, 2007.

[39] Paul A. Salvini. *Design and development of the ILUTE operational prototype: a comprehensive microsimulation model of urban systems*. PhD thesis, University of Toronto, Department of Civil Engineering, Toronto, 2003.

[40] Paul A. Salvini and Eric J. Miller. ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics*, 5(2):217–234, June 2005.

[41] Statistics Canada. The nation, dwellings and households part 1. Report 93-104, Ottawa, December 1987.

[42] Statistics Canada. 1986 census handbook. Report 99-104E, Ottawa, June 1988.

[43] Statistics Canada. Census of Canada 1986 Public Use Microdata file on households and housing. Documentation and user's guide, Ottawa, April 1989.

[44] Statistics Canada. Census of Canada 1986 Public Use Microdata File on individuals. Documentation and user's guide, Ottawa, November 1989.

[45] Statistics Canada. User's guide to 1986 census data on families. Report 99-113E, Ottawa, 1989.

[46] Statistics Canada. Census of Canada 1986 Public Use Microdata file on families. Documentation and user's guide, Ottawa, May 1990.

[47] Statistics Canada. General review of the 1986 census. Report 99-137E, Ottawa, 1990.

[48] Statistics Canada. User's guide to the quality of 1986 census data: Coverage. Report 99-135E, Ottawa, March 1990.

[49] Statistics Canada. 1996 census handbook. Report 92-352-XPE, Ottawa, June 1997.

[50] Frederick F. Stephan. Iterative methods of adjusting sample frequency tables when expected margins are known. *The Annals of Mathematical Statistics*, 13(2):166–178, June 1942.

[51] Transportation Research Board. Metropolitan travel forecasting: Current practice and future direction. Special Report 288, Washington, D.C., 2007.

[52] David Voas and Paul Williamson. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5):349–366, 2000.

[53] David Voas and Paul Williamson. Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2):177–200, November 2001.

[54] Thomas D. Wickens. *Multiway Contingency Tables Analysis for the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1989.

[55] Leon C.R.J. Willenborg and Ton de Waal. *Elements of Statistical Disclosure Control*. Number 155 in Lecture Notes in Statistics. Springer-Verlag, New York, 2001.

[56] Paul Williamson. The aggregation of small-area synthetic microdata to higher-level geographies: An assessment of fit. Working Paper 2002/1, University of Liverpool, Department of Geography, Population Microdata Unit, Liverpool, 2002.

[57] Paul Williamson, Mark Birkin, and Phil H. Rees. The estimation of population microdata by using data from Small Area Statistics and Samples of Anonymised Records. *Environment and Planning A*, 30(5):785–816, 1998.

Appendix A

Attribute Definitions

The attribute definitions and descriptions below are largely quoted directly from the Census guides to the public use microdata files, with some adaptations for the simpler categories used for population synthesis [43, 44, 46].

A.1 Person Attributes

- AGEP: Age.

  Refers to age at last birthday (as of the census reference date, June 3, 1986). This variable is derived from date of birth.

  1. 15–17.
  2. 18–19.
  3. 20–24.
  4. 25–34.
  5. 35–44.
  6. 45–54.
  7. 55–64.
  8. 65 or older.

- CFSTAT: Census family status and living arrangements.

  Refers to the classification of the population into family and non-family persons. Family persons are household members who belong to a census family (who live in the same dwelling and have a husband-wife or parent-never-married child relationship). Non-family persons are household members who do not belong

to a census family. These categories can be further broken down as indicated by the classes below. (For complete definition of census family status and living arrangements, see 1986 Census Dictionary.)

1. Husband, wife or common-law partner.
2. Child in husband-wife family.
3. Lone parent.
4. Child in a lone-parent family.
5. Non-family person living with others.
6. Non-family person living alone.
7. Not applicable. Includes persons in collectives, persons in households outside Canada and temporary residents

- HLOSP: Highest level of schooling.

  Refers to the highest grade or year of elementary or secondary school attended, or the highest year of university or other non-university completed. University education is considered to be above other non-university. Also, the attainment of a degree, certificate or diploma is considered to be at a higher level than years completed or attended without an educational qualification.

  1. Less than Grade 9. Includes no schooling or kindergarten only.
  2. Grades 9–13.
  3. Secondary (high) school graduation certificate.
  4. Trades certificate or diploma; or other non-university education only, with trades certificate or diploma.
  5. Other non-university education only, without trades or other non-university certificate or diploma.
  6. Other non-university education only, with other non-university certificate or diploma.
  7. University without certificate, diploma or degree.
  8. University with certificate or diploma. Includes trade certificates, other non-university certificate and university certificate below bachelor level.
  9. University with bachelor's degree or higher. Includes university certificate above bachelor level.

- LFACT: Labour force activity.

  Refers to the labour market activity of the population 15 years of age and over, excluding institutional residents, who, in the week prior to enumeration (June 3, 1986) were Employed, Unemployed or Not in the Labour Force. Special note: the census labour force activity concepts have not changed between 1981 and 1986.

However, the processing of the data was modified causing some differences. In the 1986 Census, contrary to previous censuses, a question on school attendance was not asked. This question was used to edit the labour force activity variable, specifically unemployment. Consequently, the processing differences affect the unemployment population and are mostly concentrated among the 15-19-year age group.

1. Employed. The Employed include those persons who, during the week prior to enumeration:
   a. did any work at all excluding housework or other maintenance or repairs around the home and volunteer work; or
   b. were absent from their jobs or businesses because of own temporary illness or disability, vacation, labour dispute at their place of work, or were absent for other reasons.
2. Unemployed. The Unemployed include those persons who, during the week prior to enumeration:
   a. were without work, had actively looked for work in the past four weeks and were available for work; or
   b. had been on lay-off and expected to return to their job; or
   c. had definite arrangements to start a new job in four weeks or less.
3. Not in Labour Force (last worked in 1985–1986). The Not in Labour Force classification refers to those persons who, in the week prior to enumeration, were unwilling or unable to offer or supply their labour services under conditions existing in their labour markets. It includes persons who looked for work during the last four weeks but who were not available to start work in the reference week, as well as persons who did not work, did not have a new job to start in four weeks or less, were not on temporary lay-off or did not look for work in the four weeks prior to enumeration.
4. Not in Labour Force (last worked prior to 1985, or never worked).

- OCC81P: Occupation, 1980 classification basis.

  This refers to the kind of work the person was doing during the reference week, as determined by their reporting of their kind of work and the description of the most important duties. If the person did not have a job during the week prior to enumeration, the data relate to the job of longest duration since January 1, 1985. Persons with two or more jobs were to report the information for the job at which they worked the most hours.

  1. Managerial, administrative and related occupations. Includes major group 11
  2. Occupations in natural sciences, engineering and mathematics. Includes major group 21

3. Occupations in social sciences and related fields. Includes major group 23
4. Teaching and related occupations. Includes major group 27
5. Occupations in medicine and health. Includes major group 31
6. Artistic, literary, recreational and related occupations. Includes major group 33
7. Clerical and related occupations. Includes major group 41
8. Sales occupations. Includes major group 51
9. Service occupations. Includes major group 61
10. Farming, horticultural and animal husbandry occupations, and other primary occupations. Includes major groups 71, 73, 75 and 77
11. Processing occupations. Includes major group 81/82
12. Machining and product fabricating, assembling & repairing occupations. Includes major groups 83 and 85
13. Construction trades occupations. Includes major group 87
14. Transport equipment operating occupations. Includes major group 91
15. Other occupations. Includes major groups 25, 93, 95, 99
16. Not applicable. Includes persons who have not worked since January 1, 1985.

- SEXP: Sex.

  Refers to the gender of the respondent.

  1. Female.
  2. Male.

- TOTINCP: Total income.

  Refers to the total money income received by individuals 15 years of age and over during the calendar year 1985 from the sources listed below.

  1. Wages and Salaries. Refers to gross wages and salaries before deductions for such items as income tax, pensions, unemployment insurance, etc. Included in this source are military pay and allowances, tips, commissions, cash bonuses as well as all types of casual earnings in calendar year 1985. All income "in kind" such as free board and lodging is excluded.
  2. Net Non-farm Self-employment Income. Refers to net income (gross receipts minus expenses of operation such as wages, rents, depreciation, etc.) received during calendar year 1985 from the respondent's non-farm unincorporated business or professional practice. In the case of a partnership, only the respondent's share was to be reported. Also included is net income

from persons baby-sitting in their own homes, operators of direct distributorships such as selling and delivering cosmetics, as well as from free-lance activities of artists, writers, music teachers, hairdressers, dressmakers, etc.

3. Net Farm Self-employment Income. Refers to net income (gross receipts from farm sales minus depreciation and cost of operation) received during calendar year 1985 from the operation of a farm, either on own account or in partnership. In the case of partnerships, only the respondent's share of income was to be reported. Also included are advance, supplementary or assistance payments to farmers by federal or provincial governments. However, the value of income "in kind", such as agricultural products produced and consumed on the farm is excluded.

4. Family Allowances. Refers to total allowances paid in calendar year 1985 by the federal and provincial governments in respect of dependent children under 18 years of age. These allowances, though not collected directly from the respondents, were calculated and included in the income of one of the parents.

5. Federal Child Tax Credits. Refers to federal child tax credits paid in calendar year 1985 by the federal government in respect of dependent children under 18 years of age. No information was collected from the respondents on child tax credits. Instead, these were calculated in the course of processing and assigned, where applicable, to one of the parents in the census family on the basis of information on children in the family and the family income.

6. Old Age Security Pension and Guaranteed Income Supplement. Refers to old age security pensions and guaranteed income supplements paid to persons 65 years of age and over, and spouses' allowances paid to 60 to 64 year-old spouses of old age security recipients by the federal government only during calendar year 1985. Also included are extended spouses' allowances paid to 60 to 64 year-old widows/widowers whose spouse was an old age security pension recipient.

7. Benefits from Canada or Quebec Pension Plan. Refers to benefits received in calendar year 1985 under the Canada or Quebec Pension Plan, e.g., retirement pensions, survivors' benefits, disability pensions. Does not include retirement pensions of civil servants, RCMP and military personnel or lump-sum death benefits.

8. Benefits from Unemployment Insurance. Refers to total unemployment insurance benefits received in calendar year 1985, before income tax deductions. It includes benefits for sickness, maternity, fishing, work sharing, retraining and retirement received under the Federal Unemployment In-

surance program.

9. Other Income from Government Sources. Refers to all transfer payments, excluding those covered as a separate income source (family allowances, federal child tax credits, old age security pensions and guaranteed income supplements, Canada/Quebec Pension Plan benefits and unemployment insurance benefits) received from federal, provincial or municipal programs in calendar year 1985. This source includes transfer payments received by persons in need such as mothers with dependent children, persons temporarily or permanently unable to work, elderly individuals, the blind and the disabled. Included are provincial income supplement payments to seniors to supplement old age security and guaranteed income supplement and provincial payments to seniors to help offset accommodation costs. Also included are other transfer payments such as for training under the National Training Program (NTP), veterans' pensions, war veterans' allowance, pensions to widows and dependants of veterans, workers' compensation, etc. Additionally, provincial tax credits and allowances claimed on the income tax return are included.
10. Dividends and Interest on Bonds, Deposits and Savings Certificates, and Other Investment Income. Refers to interest received in calendar year 1985 from deposits in banks, trust companies, co-operatives, credit unions, caisses populaires, etc., as well as interest on savings certificates, bonds and debentures and all dividends from both Canadian and foreign stocks. Also included is other investment income from either Canadian or foreign sources such as net rents from real estate, mortgage and loan interest received, regular income from an estate or trust fund, and interest from insurance policies.
11. Retirement Pensions, Superannuation and Annuities. Refers to all regular income received during calendar year 1985 as the result of having been a member of a pension plan of one or more employers. It includes payments received from all annuities, including payments from a mature registered retirement savings plan (RRSP) in the form of a life annuity, a fixed term annuity, a registered retirement income fund or an income-averaging annuity contract; pensions paid to widows or other relatives or deceased pensioners; pensions of retired civil servants, Armed Forces personnel and RCMP officers; annuity payments received from the Canadian Government Annuities Fund, an insurance company, etc. Does not include lump-sum death benefits, lump-sum benefits or withdrawals from a pension plan or RRSP or refunds of overcontributions.
12. Other Money Income. Refers to regular cash income received during calendar year 1985 and not reported in any of the other nine sources listed on the questionnaire, e.g., alimony, child support, periodic support from other per-

sons not in the household, net income from roomers and boarders, income from abroad (except dividends and interest), non-refundable scholarships and bursaries, severance pay, royalties, strike pay.

13. Receipts Not Counted as Income. Gambling gains and losses, money inherited during the year in a lump sum, capital gains or losses, receipts from the sale of property or personal belongings, income tax refunds, loan payments received, loans repaid to an individual as the lender, lump sum settlements of insurance policies, rebates of property taxes and other taxes, and refunds of pension contributions were excluded as well as all income in kind such as free meals, living accommodation, or food and fuel produced on own farm.

Individuals immigrating to Canada in 1986 have zero income. Also, because of response problems, all individuals in Hutterite colonies were assigned zero income. Furthermore, data on households, economic families, unattached individuals, census families and non-family persons relate to private households only.

1. Negative income.
2. $0.
3. $1–$999.
4. $1,000–$2,999.
5. $3,000–$4,999.
6. $5,000–$6,999.
7. $7,000–$9,999.
8. $10,000–$14,999.
9. $15,000–$19,999.
10. $20,000–$24,999.
11. $25,000–$29,999.
12. $30,000–$34,999.
13. $35,000 or more.

- CTCODE: Census Tract.

  Census Tract number

  731 different identifying codes.

A.2 Family Attributes

- AGEF: Age of wife or female lone parent.

  Refers to age at last birthday (as of the census reference date, June 3, 1986). This variable is derived from date of birth.

  1. 15–17.
  2. 18–19.
  3. 20–24.
  4. 25–34.
  5. 35–44.
  6. 45–54.
  7. 55–64.
  8. 65 or older.
  9. Not applicable. Includes male lone-parent families.

- AGEM: Age of husband or male lone parent.

  Refers to age at last birthday (as of the census reference date, June 3, 1986). This variable is derived from date of birth.

  1. 15–17.
  2. 18–19.
  3. 20–24.
  4. 25–34.
  5. 35–44.
  6. 45–54.
  7. 55–64.
  8. 65 or older.
  9. Not applicable. Includes female lone-parent families.

- CFSIZE: Number of persons in census family.

  Refers to the classification of census families by the number of persons in the family.

  1. Two persons.
  2. Three persons.
  3. Four persons.
  4. Five persons.
  5. Six persons.

6. Seven persons.
7. Eight or more persons.

- CFSTRUC: Census family structure.

  Refers to the classification of census families into husband-wife families (with or without children present) and lone-parent families by sex of parent.

  The category 'Without children present' for 1986 includes all childless husband-wife families as well as husband-wife families with children no longer at home. In 1981, these two categories were exclusive.

  1. Husband-wife family.
  2. Lone female parent.
  3. Lone male parent.

- CHILDA: Number of children in census family at home under 6 years of age.

  1. None.
  2. One child.
  3. Two or more children.

- CHILDB: Number of children in census family at home 6 to 14 years of age.

  1. None.
  2. One child.
  3. Two children.
  4. Three or more children.

- CHILDC: Number of children in census family at home 15 to 17 years of age.

  1. None.
  2. One child.
  3. Two or more children.

- CHILDDE: Number of children in census family at home 18 to 24 years of age and 25 years of age or over.

  1. No children 18 to 24, no children 25 or over.
  2. One child 18 to 24, no children 25 or over.
  3. Two or more children 18 to 24, no children 25 or over.
  4. No children 18 to 24, one child 25 or over.
  5. One child 18 to 24, one child 25 or over.

6. Two or more children 18 to 24, one child 25 or over.
7. No children 18 to 24, two or more children 25 or over.
8. One child 18 to 24, two or more children 25 or over.
9. Two or more children 18 to 24, two or more children 25 or over.

- HHNUMCF: Number of census families in household.

  1. One census family.
  2. Two or more census families.

- LFACTF: Labour force activity of wife or female lone parent.

  Refers to the labour market activity of the wife or female lone parent, who, in the week prior to enumeration (June 3, 1986) were Employed, Unemployed or Not in the Labour Force. Special note: the census labour force activity concepts have not changed between 1981 and 1986. However, the processing of the data was modified causing some differences. In the 1986 Census, contrary to previous censuses, a question on school attendance was not asked. This question was used to edit the labour force activity variable, specifically unemployment. Consequently, the processing differences affect the unemployment population and are mostly concentrated among the 15-19-year age group.

- Employed. The Employed include those persons who, during the week prior to enumeration:

  1. did any work at all excluding housework or other maintenance or repairs around the home and volunteer work; or
  2. were absent from their jobs or businesses because of own temporary illness or disability, vacation, labour dispute at their place of work, or were absent for other reasons.

  1. Unemployed. The Unemployed include those persons who, during the week prior to enumeration:
     a. were without work, had actively looked for work in the past four weeks and were available for work; or
     b. had been on lay-off and expected to return to their job; or
     c. had definite arrangements to start a new job in four weeks or less.
  2. Not in Labour Force (last worked in 1985–1986). The Not in Labour Force classification refers to those persons who, in the week prior to enumeration, were unwilling or unable to offer or supply their labour services under conditions existing in their labour markets. It includes persons who looked for work during the last four weeks but who were not available to start work in the reference week, as well as persons who did not work, did not have a

new job to start in four weeks or less, were not on temporary lay-off or did not look for work in the four weeks prior to enumeration.

3. Not in Labour Force (last worked prior to 1985, or never worked).
4. Not applicable. Includes male lone parent families.

- LFACTM: Labour force activity of husband or male lone parent.

  Refers to the labour market activity of the husband or male lone parent, who, in the week prior to enumeration (June 3, 1986) were Employed, Unemployed or Not in the Labour Force. Special note: the census labour force activity concepts have not changed between 1981 and 1986. However, the processing of the data was modified causing some differences. In the 1986 Census, contrary to previous censuses, a question on school attendance was not asked. This question was used to edit the labour force activity variable, specifically unemployment. Consequently, the processing differences affect the unemployment population and are mostly concentrated among the 15-19-year age group.

  1. Employed. The Employed include those persons who, during the week prior to enumeration:
     a. did any work at all excluding housework or other maintenance or repairs around the home and volunteer work; or
     b. were absent from their jobs or businesses because of own temporary illness or disability, vacation, labour dispute at their place of work, or were absent for other reasons.
  2. Unemployed. The Unemployed include those persons who, during the week prior to enumeration:
     a. were without work, had actively looked for work in the past four weeks and were available for work; or
     b. had been on lay-off and expected to return to their job; or
     c. had definite arrangements to start a new job in four weeks or less.
  3. Not in Labour Force (last worked in 1985–1986). The Not in Labour Force classification refers to those persons who, in the week prior to enumeration, were unwilling or unable to offer or supply their labour services under conditions existing in their labour markets. It includes persons who looked for work during the last four weeks but who were not available to start work in the reference week, as well as persons who did not work, did not have a new job to start in four weeks or less, were not on temporary lay-off or did not look for work in the four weeks prior to enumeration.
  4. Not in Labour Force (last worked prior to 1985, or never worked).
  5. Not applicable. Includes female lone parent families.

- NUCHILD: Number of children in census family at home.

1. None.
2. One child.
3. Two children.
4. Three children.
5. Four children.
6. Five children.
7. Six children.
8. Seven children.
9. Eight or more children.

- ROOM: Number of rooms.

  Refers to the number of rooms in a dwelling. A room is an enclosed area within a dwelling which is finished and suitable for year-round living.

  1. 1 room.
  2. 2 rooms.
  3. 3 rooms.
  4. 4 rooms.
  5. 5 rooms.
  6. 6 rooms.
  7. 7 rooms.
  8. 8 rooms.
  9. 9 rooms.
  10. 10 or more rooms.

- TENURE: Tenure.

  Refers to whether some member of the household owns or rents the dwelling.

  1. Owned (with or without mortgage).
  2. Rented (for cash, other). Includes families and non-family persons who rent their dwellings and reserve dwellings.

- CTCODE: Census Tract.

  Census Tract number

  731 different identifying codes

A.3 Dwelling/Household Attributes

- BUILTH: Period of construction.

  Refers to the period in time during which the building or dwelling was originally constructed.

  1. 1920 or before.
  2. 1921–1945.
  3. 1946–1960.
  4. 1961–1970.
  5. 1971–1975.
  6. 1976–1980.
  7. 1981–1986. Includes the first five months only of 1986.

- DTYPEH: Structural type of dwelling.

  Refers to the structural characteristics and/or dwelling configuration, that is, whether the dwelling is a detached single house, apartment, etc.

  1. Single-detached house.
  2. Apartment in a building that has five or more storeys.
  3. Apartment in a building that has less than five storeys.
  4. Semi-detached house.
  5. Apartment or flat in a detached duplex; row house or other single attached house.
  6. Mobile and other movable.

- HHNUEF: Number of economic families in household.

  Refers to the presence and number of economic families in the household. An economic family is defined as a group of individuals sharing a common dwelling unit and related by blood, marriage, adoption or common law.

  1. None.
  2. One or more economic families.

- HHNUMCF: Number of census families in household.

  1. None.
  2. One census family.
  3. Two or more census families.

- HHSIZE: Household size.

  Refers to the total number of persons in a private household.

  1. One.
  2. Two.
  3. Three.
  4. Four.
  5. Five.
  6. Six.
  7. Seven.
  8. Eight or more persons.

- PAYH: Monthly gross rent or owner's monthly major payments.

  Refers to the total average monthly payments paid by tenant or owner households to secure shelter. Owner's major payments include payments for electricity, oil, gas, coal, wood or other fuels, water and other municipal services, monthly mortgage payments, and property taxes (municipal and school).

  1. $0–$199.
  2. $200–$399.
  3. $400–$699.
  4. $700–$999.
  5. $1000 or more.

- PPERROOM: Number of persons per room.

  1. 0–0.5.
  2. 0.6–1.0.
  3. 1.1–1.5.
  4. 1.6–2.0.
  5. 2.1 or more.

- ROOM: Number of rooms.

  Refers to the number of rooms in a dwelling. A room is an enclosed area within a dwelling which is finished and suitable for year-round living.

  1. 1 room.
  2. 2 rooms.
  3. 3 rooms.

4. 4 rooms.
5. 5 rooms.
6. 6 rooms.
7. 7 rooms.
8. 8 rooms.
9. 9 rooms.
10. 10 or more rooms.

- TENURH: Tenure.

  Refers to whether some member of the household owns or rents the dwelling.

  1. Owned (with or without mortgage).
  2. Rented (for cash, other). Includes families and non-family persons who rent their dwellings and reserve dwellings.

- CTCODE: Census Tract.

  Census Tract number

  731 different identifying codes.

Appendix B

Detailed Results

Additional details of the results and evaluation procedure are included in this appendix.

1D STs (entire PUMA)	1D STs (by zone)	
AGEP (6)	CTCODE × AGEP (4386)	
CFSTAT (5)	CTCODE × CFSTAT (3655)	
HLOSP (6)	CTCODE × HLOSP (4386)	
LFACT (3)	CTCODE × LFACT (2193)	
OCC81P (16)	CTCODE × OCC81P (11696)	
SEXP (2)	CTCODE × SEXP (1462)	
CTCODE (731)		
2–3D STs (entire PUMA)	**2–3D STs (by zone)**	**2–3D (only in PUMS)**
AGEP × CFSTAT (20)	CTCODE × AGEP × CFSTAT (14620)	AGEP × TOTINCP (104)
AGEP × CFSTAT × SEXP (40)	CTCODE × AGEP × CFSTAT × SEXP (29240)	CFSTAT × HLOSP (63)
AGEP × HLOSP (36)	CTCODE × AGEP × HLOSP (26316)	CFSTAT × OCC81P (112)
AGEP × HLOSP × SEXP (72)	CTCODE × AGEP × HLOSP × SEXP (52632)	CFSTAT × TOTINCP (91)
AGEP × LFACT (18)	CTCODE × AGEP × LFACT (13158)	HLOSP × OCC81P (144)
AGEP × LFACT × SEXP (36)	CTCODE × AGEP × LFACT × SEXP (26316)	HLOSP × TOTINCP (117)
AGEP × SEXP (12)	CTCODE × AGEP × SEXP (11696)	OCC81P × TOTINCP (208)
CFSTAT × SEXP (10)	CTCODE × CFSTAT × SEXP (7310)	AGEP × CFSTAT × HLOSP (504)
HLOSP × LFACT (21)	CTCODE × HLOSP × LFACT (15351)	AGEP × CFSTAT × OCC81P (896)
HLOSP × LFACT × SEXP (42)	CTCODE × HLOSP × LFACT × SEXP (30702)	AGEP × CFSTAT × TOTINCP (728)
HLOSP × SEXP (12)	CTCODE × HLOSP × SEXP (8772)	AGEP × HLOSP × TOTINCP (936)
LFACT × SEXP (6)	CTCODE × LFACT × SEXP (4386)	CFSTAT × HLOSP × OCC81P (1008)
OCC81P × SEXP (32)	CTCODE × OCC81P × SEXP (23392)	CFSTAT × HLOSP × TOTINCP (819)
SEXP × TOTINCP (24)	CTCODE × SEXP × TOTINCP (17544)	CFSTAT × OCC81P × TOTINCP (1456)
		HLOSP × OCC81P × TOTINCP (1872)

Table B.1: Validation tables used to evaluate the goodness-of-fit of synthetic population, with the cell count in parentheses.

	Experiment									
	Almost no geography				Zone-by-zone				Multizone	
	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
1. 1D STs (entire PUMA)										
AGEP	0	1	1	1	5	1	2	1	0	0
CFSTAT	1	0	0	1	3	2	2	1	0	0
CTCODE	0	0	1	1	2	2	1	1	1	1
HLOSP	0	0	1	1	2	1	1	1	0	0
LFACT	3	1	3	1	2	1	2	1	0	0
OCC81P	0	0	0	0	9	3	7	2	0	0
SEXP	0	1	0	0	3	1	1	1	0	0
2. 1D STs (by zone)										
CTCODE × AGEP	311	309	309	309	2	3	1	1	3	1
CTCODE × CFSTAT	408	407	407	407	4	6	3	2	7	2
CTCODE × HLOSP	401	402	402	402	3	3	3	3	3	3
CTCODE × LFACT	157	156	156	156	4	3	3	2	3	2
CTCODE × OCC81P	439	437	437	437	0	0	1	0	1	1
CTCODE × SEXP	0	0	0	0	2	1	1	1	1	1
3. 2–3D STs (entire PUMA)										
AGEP × CFSTAT	286	13	2	1	272	13	3	1	13	0
AGEP × CFSTAT × SEXP	320	14	2	2	303	20	3	1	14	0
AGEP × HLOSP	224	20	2	3	224	39	14	7	20	0
AGEP × HLOSP × SEXP	273	29	2	3	271	44	16	8	30	0
AGEP × LFACT	222	10	3	2	214	8	7	2	10	0
AGEP × LFACT × SEXP	276	12	3	2	265	13	9	4	12	0
AGEP × SEXP	105	3	2	1	99	7	2	1	3	0
CFSTAT × SEXP	68	5	0	1	61	10	2	1	5	0
HLOSP × LFACT	232	38	3	2	229	38	6	2	38	0
HLOSP × LFACT × SEXP	282	44	3	2	280	44	8	3	44	0
HLOSP × SEXP	67	10	1	2	69	7	2	1	10	0
LFACT × SEXP	84	3	3	1	83	2	2	1	2	0
OCC81P × SEXP	244	6	0	0	242	10	14	3	6	0
SEXP × TOTINCP	0	1	0	1	12	7	14	7	0	0
4. 2–3D STs (by zone)										
CTCODE × AGEP × CFSTAT	835	776	776	776	348	153	5	4	154	4
CTCODE × AGEP × CFSTAT × SEXP	871	800	800	800	405	222	7	5	223	5
CTCODE × AGEP × HLOSP	661	618	617	617	334	199	3	2	196	2
CTCODE × AGEP × HLOSP × SEXP	697	637	636	636	410	268	3	2	267	2
CTCODE × AGEP × LFACT	535	475	475	475	259	91	2	2	91	2
CTCODE × AGEP × LFACT × SEXP	580	497	497	497	331	155	3	2	156	2
CTCODE × AGEP × SEXP	388	354	362	354	173	88	0	0	88	0
CTCODE × CFSTAT × SEXP	434	426	426	426	105	83	4	3	82	3
CTCODE × HLOSP × LFACT	643	594	592	592	284	130	2	2	131	2
CTCODE × HLOSP × LFACT × SEXP	689	623	622	622	370	192	3	2	193	3
CTCODE × HLOSP × SEXP	413	408	407	407	114	77	1	1	77	1
CTCODE × LFACT × SEXP	187	166	166	166	95	38	2	2	39	2
CTCODE × OCC81P × SEXP	553	489	489	489	297	125	6	5	126	5
CTCODE × SEXP × TOTINCP	444	444	444	444	3	5	3	3	7	4
5. 2–3D (only in PUMS)										
AGEP × TOTINCP	479	9	434	4	461	27	389	23	9	0
CFSTAT × HLOSP	255	38	176	3	260	52	171	35	38	0
CFSTAT × OCC81P	353	9	253	1	334	27	239	28	9	0
CFSTAT × TOTINCP	358	6	304	3	341	23	278	23	6	0
HLOSP × OCC81P	539	41	298	2	516	54	253	30	40	0
HLOSP × TOTINCP	422	29	373	3	363	47	299	33	29	0
OCC81P × TOTINCP	746	6	751	2	726	48	723	45	6	0
AGEP × CFSTAT × HLOSP	923	73	619	12	919	134	458	74	73	0
AGEP × CFSTAT × OCC81P	1198	50	755	10	1158	75	623	61	49	0
AGEP × CFSTAT × TOTINCP	1117	27	952	9	1072	63	863	56	27	0
AGEP × HLOSP × TOTINCP	1111	54	995	10	1066	100	850	61	54	0
CFSTAT × HLOSP × OCC81P	1189	80	865	6	1152	113	780	81	80	0
CFSTAT × HLOSP × TOTINCP	1071	57	976	7	991	103	849	80	57	0
CFSTAT × OCC81P × TOTINCP	1798	25	1771	6	1742	112	1684	107	26	0
HLOSP × OCC81P × TOTINCP	1684	67	1501	7	1622	127	1427	100	67	0

Table B.2: Detailed results of experiments I1–I10, testing goodness-of-fit of IPF under varying amounts of input data. Each column shows the results of a single experiment, measured using SRMSE $\times 1000$ against a single validation table. The input data and description of these experiments can be found in Table 6.2.