



# Politecnico di Torino

**Politecnico di Torino**

Master degree in Civil Engineering

**Population synthesis and activity chain generation  
for implementing an activity-based model: A study  
for the city of Turin**

**Supervisor:**

Marco Diana

**Candidate:**

Mostaan Hosseinian Daroonkola

Academic Year 2025/2026

# Abstract

Agent-based transport models require detailed disaggregated information on individuals, households, and daily activity patterns. However, such data are not available mainly due to cost of collecting and privacy issue. Census data are either aggregated information, or it lacks detailed residential or in general activities location. Mobility surveys are limited in size and also mostly lack activity's location due to privacy. This thesis addresses this issue in a specific case study by developing a synthetic population for the city of Torino to support the eventual implementation of an agent-based model.

To achieve this result, this study combines three data sources, namely (1) the 2011 census aggregated dataset at census tract level, (2) the 1% census individual and household microdata sample and (3) a very detailed mobility survey that was administered by Politecnico di Torino in the Turin metropolitan area in 2017. Population synthesis was then done by using Iterative Proportional Fitting (IPF) to estimate joint distribution of selected control variables including sex, age group, educational attainment and employment status. After that households with all of their members were assigned to census tracts, synthetic population were enriched by other variables based on available census and microdata information. The generated synthetic population was validated by internal and external validation.

Finally, Bayesian Network model was applied to assign activity-chain sequences to synthetic individuals that were then validated as well. Overall, the thesis provides a detailed synthetic population for Torino, suitable as a basis for future activity-based travel demand modelling applications.

# Riassunto

I modelli di trasporto basati su agenti richiedono informazioni dettagliate e disaggregate su individui, famiglie e schemi di attività quotidiane. Tuttavia, tali dati non sono disponibili principalmente a causa dei costi di raccolta e delle questioni legate alla privacy. I dati censuari forniscono informazioni aggregate oppure mancano di dettagli sulla localizzazione residenziale o, più in generale, sui luoghi delle attività. Le indagini sulla mobilità sono di dimensioni limitate e, inoltre, spesso non includono la localizzazione delle attività per motivi di privacy. Questa tesi affronta tale problematica in uno specifico caso studio, sviluppando una popolazione sintetica per la città di Torino, al fine di supportare l'implementazione di un modello basato su agenti.

Per raggiungere tale risultato, questo studio combina tre differenti fonti di dati: (1) il dataset del censimento del 2011 riportante le statistiche aggregate a livello di sezione censuaria, (2) i microdati individuali e familiari del campione censuario all'1% e (3) un'indagine sulla mobilità molto dettagliata che è stata condotta dal Politecnico di Torino nell'area metropolitana torinese nel 2017. La sintesi della popolazione è stata effettuata utilizzando l'Iterative Proportional Fitting (IPF) per stimare la distribuzione congiunta di alcune variabili di controllo selezionate, tra cui sesso, classe di età, livello di istruzione e condizione occupazionale. Successivamente, le famiglie, insieme a tutti i loro componenti, sono state assegnate alle sezioni censuarie; la popolazione sintetica è stata quindi arricchita con altre variabili sulla base delle informazioni disponibili nei dati censuari e nei microdati. La popolazione sintetica generata è stata validata tramite validazione interna ed esterna.

Infine, è stato applicato un modello di rete bayesiana per assegnare sequenze di catene di attività agli individui sintetici, che sono state successivamente validate. Nel complesso, la tesi fornisce una popolazione sintetica dettagliata per Torino, idonea a costituire una base per future applicazioni di modellazione della domanda di trasporto basata sulle attività.

# GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

Generative Artificial Intelligence tools, specifically ChatGPT (OpenAI), were used in this thesis to support the writing process. The tool was mainly employed for editorial assistance, including improving grammar, clarity, and readability of the text.

ChatGPT was occasionally used to support coding tasks and data processing workflows, particularly in generating and refining R scripts and Python script. However, all generated code was carefully reviewed and tested. No AI tool was used to generate results, perform data analysis autonomously, or produce final outputs.

AI-assisted contributions are mainly present in Chapter 3 (Methodology) and Chapter 4 (Case Study), where support was used to improve clarity and structure of the explanations.

All AI-generated content was critically reviewed, fact-checked, and edited to align with academic writing standards. No citations were taken from AI, and the candidate takes full responsibility for the accuracy, originality, and integrity of the final work.

# Table of Contents

|                                                                        |           |
|------------------------------------------------------------------------|-----------|
| <b>List of Tables</b>                                                  | <b>7</b>  |
| <b>List of Figures</b>                                                 | <b>9</b>  |
| <b>1 Introduction</b>                                                  | <b>10</b> |
| <b>2 Literature review</b>                                             | <b>12</b> |
| 2.1 Population Synthesis Methods . . . . .                             | 12        |
| 2.1.1 Synthetic Reconstruction (SR) . . . . .                          | 13        |
| 2.1.1.1 Introduction to IPF . . . . .                                  | 14        |
| 2.1.1.2 Seed or contingency table . . . . .                            | 14        |
| 2.1.1.3 Control variables . . . . .                                    | 15        |
| 2.1.1.4 Zero-cells problems . . . . .                                  | 15        |
| 2.1.1.5 Integerisation . . . . .                                       | 15        |
| 2.1.2 Combinatorial Optimization . . . . .                             | 16        |
| 2.1.3 Statistical or Machine Learning Models . . . . .                 | 17        |
| 2.1.4 Households Synthetic Approaches . . . . .                        | 17        |
| 2.1.4.1 Households sample with individual information . . . . .        | 18        |
| 2.1.4.2 Individual and Households Independent Data . . . . .           | 18        |
| 2.1.4.3 Individual level data with additional household data . . . . . | 18        |
| 2.2 Population Synthesis Validation . . . . .                          | 19        |
| 2.2.1 Internal Validation . . . . .                                    | 19        |
| 2.2.2 External Validation . . . . .                                    | 19        |
| 2.3 Activity Chain Synthesis . . . . .                                 | 20        |
| <b>3 Methodology</b>                                                   | <b>22</b> |
| 3.1 Control Variables . . . . .                                        | 22        |
| 3.2 Theoretical Formulation of IPF . . . . .                           | 23        |
| 3.3 Integerisation (TRS) . . . . .                                     | 25        |
| 3.3.1 Numerical Example of TRS . . . . .                               | 27        |
| 3.4 Generation of Synthetic Individuals . . . . .                      | 29        |
| 3.5 Enrichment Concept . . . . .                                       | 31        |

|          |                                                                                                                 |           |
|----------|-----------------------------------------------------------------------------------------------------------------|-----------|
| 3.6      | Household Synthesis . . . . .                                                                                   | 32        |
| 3.7      | Validation Framework and Metrics . . . . .                                                                      | 32        |
| 3.7.1    | Validation Metrics . . . . .                                                                                    | 32        |
| 3.8      | Activity Chain Generation . . . . .                                                                             | 34        |
| <b>4</b> | <b>Case Study</b>                                                                                               | <b>38</b> |
| 4.1      | Study Area and Zoning . . . . .                                                                                 | 38        |
| 4.2      | Data Preparation . . . . .                                                                                      | 40        |
| 4.2.1    | Census Statistics Preprocessing . . . . .                                                                       | 42        |
| 4.2.2    | 1% Sample Census Individual Microdata Preprocessing . . . . .                                                   | 43        |
| 4.3      | Selection of Control Variables . . . . .                                                                        | 46        |
| 4.4      | Iterative Proportional Fitting and Integerisation Implementation . . . . .                                      | 46        |
| 4.5      | Synthetic Individuals and Households Generation . . . . .                                                       | 48        |
| 4.5.1    | Synthetic Households Generation . . . . .                                                                       | 49        |
| 4.5.2    | Synthetic Individuals Generation . . . . .                                                                      | 50        |
| 4.5.3    | Collective Buildings . . . . .                                                                                  | 50        |
| 4.5.4    | Final Output . . . . .                                                                                          | 51        |
| 4.6      | Internal Validation Results . . . . .                                                                           | 51        |
| 4.7      | External Validation Results . . . . .                                                                           | 57        |
| 4.8      | Descriptive Analysis of the Synthetic Population . . . . .                                                      | 59        |
| 4.9      | Activity-Chain Synthesis Using Bayesian Network . . . . .                                                       | 61        |
| 4.9.1    | Mobility Survey Preparation . . . . .                                                                           | 61        |
| 4.9.2    | Bayesian Network Implementation for Driving License . . . . .                                                   | 63        |
| 4.9.3    | Bayesian Network Implementation for Activity-Chain Generation . . . . .                                         | 63        |
| 4.9.4    | Activity-Chain Validation . . . . .                                                                             | 66        |
| 4.9.5    | Descriptive Analysis of the Synthesized Activity Chains . . . . .                                               | 68        |
| <b>5</b> | <b>Conclusions</b>                                                                                              | <b>71</b> |
|          | <b>References</b>                                                                                               | <b>73</b> |
|          | <b>Appendices</b>                                                                                               | <b>77</b> |
| A.1      | Available variables from 2011 census . . . . .                                                                  | 77        |
| A.2      | Control variables include original and derived census variables . . . . .                                       | 83        |
| A.3      | Variables and English definitions of the 2011 1% sample census microdata . . . . .                              | 85        |
| A.4      | Categories and codes for the categorical variables in the 2011 census 1%<br>individual sample dataset . . . . . | 87        |
| A.5      | Description of Variables in the Synthetic Population Dataset . . . . .                                          | 97        |
| A.6      | Description of Variables in the Synthetic Household Dataset . . . . .                                           | 99        |
| A.7      | R Script for IPF Implementation and TRS Integerisation . . . . .                                                | 100       |

|                                                                                  |     |
|----------------------------------------------------------------------------------|-----|
| A.8 Python Script for Implementation of Bayesian Network for Driving License     |     |
| Ownership . . . . .                                                              | 108 |
| A.9 Python Script for Implementation of Bayesian Network for Activity Chain      |     |
| Generation . . . . .                                                             | 114 |
| A.10 Distribution of the twenty most frequent activity chains by household group | 125 |

# List of Tables

|    |                                                                                                                              |    |
|----|------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Summary of population synthesis methods (source: own elaboration) . . .                                                      | 13 |
| 2  | Comparison of internal and external validation methods (source: own elaboration) . . . . .                                   | 20 |
| 3  | TRS microdata (source: own elaboration) . . . . .                                                                            | 27 |
| 4  | First step of TRS (source: own elaboration) . . . . .                                                                        | 28 |
| 5  | Summary of datasets used in the study (source: own elaboration). . . . .                                                     | 41 |
| 6  | Census variables used for deriving control variables (source: census aggregate dataset). . . . .                             | 42 |
| 7  | New labels for the <b>SESS0</b> variable in the 1% sample census microdata (source: own definition). . . . .                 | 44 |
| 8  | New classification and labels for educational attainment in the 1% sample census microdata (source: own definition). . . . . | 45 |
| 9  | New classification for employment variables in the 1% sample census microdata (source: own definition). . . . .              | 45 |
| 10 | Marginal constraints used in the IPF procedure. . . . .                                                                      | 47 |
| 11 | Settings used for the <b>mipfp</b> package. . . . .                                                                          | 48 |
| 12 | Example of IPF output after integerisation. . . . .                                                                          | 49 |
| 13 | Census attributes related to household size (source: census aggregate dataset). . . . .                                      | 49 |
| 14 | Example of census statistics for the selected census tract . . . . .                                                         | 59 |
| 15 | Joint distribution of synthetic individuals in a randomly selected census tract . . . . .                                    | 60 |
| 16 | Aligning socio-demographic variables between the mobility survey and the synthetic population dataset. . . . .               | 62 |
| 17 | Aggregation of original activity types into activity categories used for the Bayesian Network. . . . .                       | 64 |
| 18 | Removed records from the mobility survey based on non-positive activity duration. . . . .                                    | 64 |
| 19 | Activity-purpose distribution by household group. . . . .                                                                    | 69 |
| 20 | Observed activity chains in the mobility survey for a selected household group. . . . .                                      | 69 |

|    |                                                                                                                 |     |
|----|-----------------------------------------------------------------------------------------------------------------|-----|
| 21 | Attributes in census aggregate dataset (source: census 2011) . . . . .                                          | 83  |
| 22 | Derived variables (source: own reclassification on census 2011 attributes) .                                    | 85  |
| 23 | Attributes in 1% census individual dataset (source: census 2011) . . . . .                                      | 87  |
| 24 | Categories and codes for the categorical variables in the 2011 census 1%<br>individual sample dataset . . . . . | 97  |
| 25 | Synthetic population attributes and their definition . . . . .                                                  | 99  |
| 26 | Synthetic household attributes and their definition . . . . .                                                   | 99  |
| 27 | Twenty most frequent activity chains by household group . . . . .                                               | 135 |

# List of Figures

|    |                                                                                                                                                 |    |
|----|-------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Numerical example of IPF (source: Zhuo (2017)) . . . . .                                                                                        | 24 |
| 2  | Spatial distribution of population across Torino’s neighborhoods (source: own elaboration on census 2011) . . . . .                             | 39 |
| 3  | Spatial distribution of population across Torino’s census tracts (source: own elaboration on census 2011) . . . . .                             | 40 |
| 4  | Internal validation results for male-age categories (source: own elaboration)                                                                   | 52 |
| 5  | Internal validation results for female-age categories (source: own elaboration)                                                                 | 52 |
| 6  | Internal validation results for male-education categories (source: own elaboration) . . . . .                                                   | 53 |
| 7  | Internal validation results for female-education categories (source: own elaboration) . . . . .                                                 | 53 |
| 8  | Internal validation results for both male and female employment categories (source: own elaboration) . . . . .                                  | 54 |
| 9  | Internal validation results for female aged 15 to 19 before integerisation (source: own elaboration) . . . . .                                  | 55 |
| 10 | Internal validation results for female aged 15 to 19 after integerisation (source: own elaboration) . . . . .                                   | 55 |
| 11 | Total absolute error for female age categories (source: own elaboration) . .                                                                    | 56 |
| 12 | Relative absolute error for female age categories (source: own elaboration)                                                                     | 56 |
| 13 | Marital status validation results for female (source: own elaboration). . . .                                                                   | 58 |
| 14 | Marital status validation results for male (source: own elaboration). . . .                                                                     | 58 |
| 15 | Learned Bayesian Network structure (source: own elaboration on mobility dataset). . . . .                                                       | 65 |
| 16 | Activity-chain frequency comparison between observed and Bayesian Network output (source: own elaboration on the Torino mobility survey). . . . | 67 |
| 17 | Comparison between observed and synthesized work-trip productions at the census-tract level (source: own elaboration). . . . .                  | 68 |

# Chapter 1

## Introduction

Nowadays, agent-based microsimulation models are used for transport planning (e.g. (Nagel, Horni, and Axhausen 2016)) and land use microsimulation (Waddell and Ulfarsson 2004). These models simulate the activity patterns and the related travel choices of each agent during a given day, so such simulated behavior can be used to evaluate the impact of different changes in transportation policy or infrastructure within a simulation framework. Such models simulate decisions at the individual level and therefore depend on disaggregated datasets describing demographic characteristics, household composition, residential location, and daily activity patterns. However, obtaining these data is often difficult, if not impossible, due to several issues such as privacy regulations, high survey costs, and limited data availability.

Traditionally, the primary data sources for this information are census data and travel surveys. Census data are usually available either as aggregated counts for small geographic areas or as anonymized microdata samples with limited spatial detail. Travel survey data also suffer from a similar problem, as the spatial resolution is often too coarse and does not identify where people live at a fine spatial level. As a consequence, neither data source alone is sufficient to support the implementation of detailed agent-based models.

Spatial microsimulation or population synthesis has emerged as an effective solution to overcome this limitation. Population synthesis techniques are methods that use commonly available categorical or ordinal socioeconomic data, typically published by statistical agencies in tabular form, to generate a detailed representation of an entire population. The resulting synthetic population consists of individual records with associated attributes, including geographic location, which gives these techniques their spatial dimension and makes them particularly valuable for geographic analysis. These methods generally combine aggregated population data available for geographic zones (such as census tracts, districts, or wards) with a sample of individual-level population records to estimate the characteristics and distribution of all members of the population. Because the resulting individuals are artificial records rather than real persons, synthetic populations also offer an important advantage in terms of privacy protection.

While a synthetic population provides a detailed representation of individuals and households, agent-based transport models also require information on how these individuals organize their daily activities. Travel demand is generally derived from the need to participate in activities such as work, education, shopping, and leisure rather than being considered an independent phenomenon. Therefore, realistic activity patterns are essential for modelling travel behavior and evaluating transport policies. However, detailed activity schedules are typically available only from travel surveys, which are often limited in sample size and spatial coverage. To overcome this limitation, activity-chain generation techniques are used to assign plausible daily activity sequences to synthetic individuals based on observed behavioral data. By assigning activity-chain to population synthesis, it becomes possible to create a complete synthetic representation of both the socio-demographic characteristics of the population and its daily activity behavior, providing a suitable input for agent-based transport simulation models.

This thesis focuses on the municipality of Torino, one of the largest cities in northern Italy. Although several official datasets describing the population of Torino are available, they do not provide detailed individual-level information at the spatial resolution required for agent-based modelling applications. The smallest census geography, the census tract, contains valuable aggregate information but lacks the multivariate detail necessary to represent individuals and households. Therefore, the main objective of this thesis is to develop a synthetic population that reproduces the demographic and socio-economic characteristics of Torino’s population, together with an activity chain generation that reproduces the travel behavior of those agents at a fine spatial level.

In this regard, the study integrates several complementary datasets. Aggregated census statistics at the census-tract level provide the marginal distributions used as constraints during population synthesis. A 1% sample of census microdata supplies the joint distribution of demographic and socio-economic characteristics and serves as the donor population. Additional household-level information is derived from census household microdata, while a mobility survey conducted in the Torino metropolitan area by Politecnico di Torino provides the behavioral information required for activity-chain generation.

The remainder of this thesis is structured as follows. Chapter 2 reviews the literature on population synthesis and activity-chain generation. Chapter 3 describes the methodology adopted for population synthesis and activity-chain assignment. Chapter 4 presents the resulting synthetic population and activity-chain outputs together with the validation analyses. Finally, Chapter 5 summarizes the main findings, discusses the limitations of the study, and outlines potential directions for future research.

# Chapter 2

## Literature review

Population synthesis emerged from the need for comprehensive individual-level mobility data with detailed spatial attributes to feed agent-based models, particularly when microdata for a study area are unavailable due to privacy restrictions or when such data are available but lack sufficient spatial detail or are only provided at very coarse geographic resolutions. Within population synthesis approaches, Beckman, Baggerly, and Mckay (1996) work is recognized as an early and influential study on constructing human agents from microdata and aggregated counts. Population synthesis research is broadly concerned with generating simulated microdata that reproduces the characteristics of real populations at fine geographic scales (Beckman, Baggerly, and Mckay 1996).

### 2.1 Population Synthesis Methods

Population synthesis methods are commonly grouped into three main categories, as summarized in Table 1. These categories are described in detail in the following three subsections, while the last one will present the approaches related to synthesizing households rather than individuals.

Table 1: Summary of population synthesis methods (source: own elaboration)

| Category                                | Method                                 | Input data                         | Strengths                                                 | Limitations                                                  |
|-----------------------------------------|----------------------------------------|------------------------------------|-----------------------------------------------------------|--------------------------------------------------------------|
| Synthetic Reconstruction (SR)           | IPF, IPU, HIPF                         | Microdata + marginal distributions | Simple; exact match to marginals; preserves relationships | Zero-cell problem; computational burden; non-integer weights |
| Combinatorial Optimization (CO)         | Simulated annealing (SA)               | Microdata + marginal distributions | Generates complete households and individual              | Computationally intensive; limited to observed units         |
| Statistical or Machine learning methods | Monte Carlo Simulation                 | Microdata + marginal distributions | Captures variability and uncertainty                      | Requires large samples; may not satisfy constraints exactly  |
|                                         | Bayesian Networks                      | Microdata                          | Captures dependencies between variables                   | Requires large datasets                                      |
|                                         | Markov Chain Monte Carlo (MCMC)        | Microdata + marginal distributions | Captures joint distributions                              | Depends on quality of conditional distributions              |
|                                         | Generative Adversarial Networks (GANs) | Large microdata                    | Strong predictive performance                             | Loss of diversity; may ignore rare groups                    |

### 2.1.1 Synthetic Reconstruction (SR)

Synthetic Reconstruction (SR) is among the earliest approaches for generating synthetic population by combining microdata with aggregate marginal distributions. These methods use the joint distribution observed in microdata and adjust it to match known aggregate constraints from census statistics or other reliable aggregated count. Core techniques include Iterative Proportional Fitting (IPF), based on the work of (Deming and Stephan 1940), as well as extensions like Iterative Proportional Updating (IPU) (Ye et al. 2009) and Hierarchical Iterative Proportional Fitting (HIPF) (Müller and Axhausen 2012). These extensions adapt the method to hierarchical data, enabling simultaneous adjustment of household- and individual-level characteristics.

One major limitation of SR methods is that if initial microdata are too small or lacks diversity, the resulting synthetic population will be biased or unrealistic (Ballas, Clarke, and Wiemers 2005). Thus, a sufficiently large sample is essential for producing reliable result. Additionally, as the number of attributes increases (e.g. sex, age, income, education, number of vehicles, public transport subscription, employment status), the number of combinations grows exponentially and this leads to a massive contingency table with many empty cells, making computation heavy and reducing statistical reliability (Ballas, Clarke, and Wiemers 2005).

Another limitation is related to the output of IPF-based approaches. IPF often produces non-integer weights, therefore the output must be rounded which can introduce

distortions (Lovelace and D. Ballas 2013). To obtain integer weights, the truncate, replicate, sample (TRS) method was applied, as described in Section 3.3.

One key advantage of SR methods lies in its simplicity and computational efficiency. The method is relatively easy to implement, requires limited modelling assumptions, and can be applied using only aggregate constraints and sample microdata, making it a widely adopted approach in population synthesis (Lovelace, Birkin, et al. 2015). Another important benefit is its ability to accurately match known aggregate data, ensuring consistency with official statistics or observed totals (Deming and Stephan 1940).

#### **2.1.1.1 Introduction to IPF**

Iterative Proportional Fitting (IPF) is a statistical procedure used to adjust contingency tables so that they satisfy a set of known marginal constraints (Deming and Stephan 1940). The algorithm has been independently rediscovered across various disciplines, including demography, transportation research, economics, and engineering. Its widespread use reflects the fundamental problem it addresses, which is how to combine incomplete joint information with reliable aggregate totals (Deming and Stephan 1940).

IPF operates as an iterative scaling procedure applied to multi-dimensional contingency tables or seed table (Beckman, Baggerly, and McKay 1996). The algorithm takes an initial table -typically derived from sample microdata - and a set of known marginal totals - obtained from census statistics or any other aggregated dataset- the algorithm repeatedly rescales cell values so that the resulting table matches known marginal totals (Ballas, Clarke, and Wiemers 2005).

Applications of IPF can be grouped into three categories. First, classical applications involve updating tables to match new marginal totals or adjusting sample relationships to reflect known population distributions. Second, IPF plays a central role in fitting log-linear models for contingency table analysis. Third, hybrid applications arise when some relationships between variables are missing and must be inferred implicitly through constrained estimation (Deming and Stephan 1940; Bishop et al. 1977; Beckman, Baggerly, and McKay 1996; Ballas, Clarke, and Wiemers 2005).

In population synthesis, IPF is primarily used to combine joint distributions from microdata with marginal totals from census statistics. This ensures that the synthetic population reproduces aggregate constraints while preserving relationships observed in the sample microdata, making IPF particularly suitable for generating spatially disaggregated populations.

#### **2.1.1.2 Seed or contingency table**

The seed table is the initial contingency table representing the joint distribution of the selected constraint variables, constructed by cross-tabulation of the available microdata.

It serves as the starting point for the Iterative Proportional Fitting (IPF) procedure, which adjusts this joint distribution to ensure consistency with the marginal totals derived from census data (L. Ma and Srinivasan 2016).

In population synthesis, donor records refer to household or individual observations extracted from sample microdata that provide the detailed joint distribution of socio-demographic characteristics (Beckman, Baggerly, and McKay 1996; Ye et al. 2009). These donor records are subsequently selected, weighted, or replicated during the synthesis process to generate synthetic individuals and households that satisfy the aggregate constraints.

### **2.1.1.3 Control variables**

Control variables and constraint variables are terms used interchangeably. Constraint variables are those variables which are present both in microdata and census statistics and used to guide how individuals in microdata should be allocated to specific zones in spatial microsimulation. In other words, control variables are the set of attributes for which the synthetic population is constrained to reproduce known marginal distributions at the target spatial level. Control variables' selection is based on the relevance to the research context and target variable rather than arbitrary assumptions. Constraint variables, ideally, should be selected in order to provide near-complete population coverage for each spatial unit (Lovelace and Dumont 2016).

### **2.1.1.4 Zero-cells problems**

The zero-cell problem is a well-known issue in IPF-based population synthesis that occurs when a joint combination of attributes is absent in the sample but required by aggregate constraints (Beckman, Baggerly, and McKay 1996). To address this limitation several approaches proposed by different researchers. A common approach is to add a small positive value to all zero-cells although this may introduce bias (Beckman, Baggerly, and McKay 1996; Guo and Bhat 2007).

Another solution involves category reduction, where empty categories merge with adjacent group therefore dimensionality decrease and results in elimination of zero-cells (Guo and Bhat 2007). This causes reduction in the level of detail in the synthetic population. Other methods attempt to preserve local structure by borrowing information across spatial scales, for example by using regional-level distributions to fill missing cells at finer geographic levels (Ye et al. 2009).

### **2.1.1.5 Integerisation**

Integerisation methods are used to convert the non-integer weights produced by iterative proportional fitting (IPF) into whole-number values. This step is necessary because these

weights determine the number of synthetic individuals represented in the final population, and in practice it is not possible to have fractional persons. Simple rounding is the easiest approach where weights are rounded up or down based on 0.5 threshold. However, this often leads to inaccurate results (Ballas, Clarke, and Wiemers 2005). An alternative method is counter-weight, that was introduced by Ballas, Clarke, and Wiemers (2005). This method is a deterministic approach that converts fractional weights into integers by cumulatively adding decimal part of the weights and assigning values when total exceed one. It improves population total with using randomness, but can underrepresent individuals with very small weights (Lovelace and D. Ballas 2013).

The limitations of previous integerisation methods highlight the need for an alternative approach. In response, Lovelace and D. Ballas (2013) introduced the Truncate, Replicate, Sample (TRS) method. TRS separates IPF weights into integer and fractional parts: the integer part is used to replicate individuals, while the remaining decimal parts are used as probabilities in a random sampling process to complete the population.

### 2.1.2 Combinatorial Optimization

Combinatorial Optimization (CO) methods generate synthetic populations by selecting and replicating households and individuals from a sample while matching known aggregate constraints (Aemmer and MacKenzie 2022; L. Ma and Srinivasan 2016; Namazi-Rad, Mokhtarian, and Perez 2014). Combinatorial optimization also requires both microdata and marginal counts as its input. Unlike Synthetic Reconstruction (SR), CO approaches are stochastic and directly produce complete households, eliminating the need for a separate allocation step (Beckman, Baggerly, and McKay 1996; Guo and Bhat 2007; Lovelace and Dumont 2016). They are also more flexible and can handle multiple constraints using optimization techniques such as hill climbing or simulated annealing (Lovelace and Dumont 2016; Namazi-Rad, Mokhtarian, and Perez 2014). These methods iteratively improve a “goodness-of-fit” measure, ensuring that the synthetic population closely aligns with marginal distributions.

Despite these advantages, CO methods have important limitations. They are computationally intensive—particularly for large populations—and, as heuristic optimization approaches, they do not guarantee convergence to a global optimum (Lovelace and Dumont 2016; Namazi-Rad, Mokhtarian, and Perez 2014; Guo and Bhat 2007). The iterative optimization process can require significant runtime, limiting scalability for large populations. Additionally, because CO methods rely on replicating sample households, they are constrained to patterns present in the input data and cannot generate entirely new combinations of attributes (Guo and Bhat 2007; Lovelace and Dumont 2016).

### 2.1.3 Statistical or Machine Learning Models

Model-based generation’s first principle is to capture the statistical structure from observed data, then generate new data respecting the observed statistical structure. These models are called machine learning models or statistical learning. A first line of research relies on Markov Chain Monte Carlo (MCMC) techniques, where variables are iteratively sampled from conditional distributions. In this context, Farooq et al. (2013) proposed a Gibbs sampling framework that generates synthetic individuals consistent with known marginal distributions, estimating missing conditional relationships through discrete choice models when necessary. This approach was shown to outperform traditional methods such as Iterative Proportional Fitting (IPF), although its performance strongly depends on the completeness and reliability of the conditional distributions used (Farooq et al. 2013). To improve scalability and computational efficiency, Saadi, Mustafa, Teller, Farooq, & Cools (2016) introduced a Hidden Markov Model (HMM)-based formulation, where attributes are modeled as ordered sequences of hidden states. By discretizing numerical variables and exploiting sequential dependencies, this method facilitates estimation of conditional probabilities and improves the handling of high-dimensional data (Ismail Saadi et al. 2016).

The second line of research focuses on Bayesian networks, which model probabilistic dependencies between variables through a graphical structure. These methods aim to approximate the joint distribution of population attributes while explicitly representing relationships among them. Sun and Erath (2015) demonstrated that Bayesian networks can effectively capture the structural properties of populations and generate realistic synthetic samples by sampling from the learned network. Extensions of this framework include the work of Zhang et al. (2019), who integrated social network information to generate connected synthetic populations, combining structure learning with probabilistic sampling. Despite their strengths in modeling dependencies, Bayesian networks face several limitations, including the curse of dimensionality, the need for large training datasets, restrictive conditional independence assumptions, and difficulties in handling continuous variables (Sun and Erath 2015; Zhang et al. 2019).

### 2.1.4 Households Synthetic Approaches

The household is another core element of human organization, especially in context of travel behavior, around which important decisions and economic activities are centered.

There are several methods for creating households depending on the availability of data. There are three possible situations (Barthelemy and Toint 2013; Lovelace and Dumont 2016):

- Having access to a household sample that contains information on each individual

member.

- Having separate datasets for individuals and household, stored independently without connection to each other.
- Lacking aggregate household data, but having access to certain individual-level variables that describe household context (such as household size or type).

#### **2.1.4.1 Households sample with individual information**

In this case each household is treated as a single unit, and a synthetic population is created by selecting and replicating these households so that their distribution matches known aggregate counts (e.g., household size or tenure) from census statistics (Lovelace and Dumont 2016). This is typically done through weighting or sampling procedures such as IPF. However, multi-level methods compute household expansion weights that simultaneously satisfy constraints at both the household and individual levels. Examples include some previously reviewed methods such as Iterative Proportional Updating (IPU) proposed by Ye et al. (2009), and Hierarchical Iterative Proportional Fitting (HIPF) introduced by Müller and Axhausen (2012).

#### **2.1.4.2 Individual and Households Independent Data**

When individual and household data are not coming from identical dataset, it is difficult to link them directly. To solve this problem a three-step method proposed by Lovelace and Dumont (2016) is used:

1. Estimate distributions separately: first, the distribution of individuals and households is calculated independently by using available data.
2. Select household types: create household one by one by choosing a type of household that keeps the synthetic household distribution as close as possible to the target or marginal distribution.
3. Assign individuals to households: household's members are selected from a pool of individuals based on their household type attributes.

The process ends when households are completed.

#### **2.1.4.3 Individual level data with additional household data**

When there are individual-level attributes related to household, in other words some cross information, such as household type, size, age gap between household's head and other members, those attributes can be used to build more realistic synthetic populations (Lovelace and Dumont 2016).

Individual are matched using probability distributions derived from real data, such as likely age combination in couples or their household's size. This reduces unrealistic combinations and better reflects real population. Some mismatches may still occur, and leftover individuals may be assigned less accurately (Lovelace and Dumont 2016).

## 2.2 Population Synthesis Validation

The objective of this Section is to present methods to evaluate whether the generated synthetic population accurately reproduces the target population. There are two main approaches used for validation of spatial microsimulation: internal validation and external validation.

### 2.2.1 Internal Validation

Internal validation consists of comparing the model output with the same total counts of constraint variables used in the IPF process (Choupani and Mamdoohi 2016). This involves counting the synthetic individuals for all categories of constraint variables and comparing the result with corresponding census or observed marginals. The purpose is to assess the degree of consistency between the simulated and observed data (Lovelace and Dumont 2016). In other words, internal validation simply answers the question: does synthetic population correctly reproduce the constraint variable distribution?

Internal validation is expected to show near-perfect agreement between simulated and observed values since the Iterative Proportional Fitting (IPF) procedure is designed to match marginal distributions (Lovelace and Dumont 2016). The following issues can result in low internal validation:

- Marginal total between constraint variables are not consistent,
- Integerisation process introduced errors,
- Synthetic population was not implemented correctly.

Therefore, internal validation is the minimum for evaluating spatial microsimulation's result since it confirms that the algorithm has been correctly implemented (Lovelace and Dumont 2016).

### 2.2.2 External Validation

Lovelace and Dumont (2016) stated that external validation involves comparing the synthetic population counts with independent data that were not used in the population synthesis process or comparing those variables were not involved in the synthetic population generation (Beckman, Baggerly, and McKay 1996). Unlike internal validation,

external validation does not necessarily result in near-perfect agreement between simulated and observed values, and some level of error is expected. The objective is not to achieve exact matching but rather to evaluate whether the synthetic population procedure can be considered reliable and whether the synthetic population a realistic estimation of real-world distributions (Choupani and Mamdoohi 2016).

In summary, internal and external validation are complementary. Table 2 presents the main characteristics and differences of validation methods.

Table 2: Comparison of internal and external validation methods (source: own elaboration)

| No | Aspect of validation    | Internal validation                         | External validation                      |
|----|-------------------------|---------------------------------------------|------------------------------------------|
| 1  | Purpose                 | Verify consistency with imposed constraints | Assess realism beyond constraints        |
| 2  | Data used               | Same data used in the synthesis process     | Independent or additional datasets       |
| 3  | Type of comparison      | Simulated vs. target marginals              | Simulated vs. observed real-world data   |
| 4  | Variables involved      | Control (constrained) variables             | Non-constrained variables                |
| 5  | Expected accuracy       | Very high (near-perfect fit)                | Moderate (some deviation expected)       |
| 6  | Interpretation of error | Indicates implementation issues             | Indicates model realism and data quality |
| 7  | Data requirement        | Low                                         | High                                     |

## 2.3 Activity Chain Synthesis

Activity-based models are grounded in the theory that travel demand is not generated independently but rather emerges from individuals’ need to participate in activities across different locations. Consequently, modelling travel demand requires detailed information on the sequence of activities performed by individuals, together with their timing, duration, and location. However, such disaggregated information is often unavailable for the entire population and must therefore be synthesized from available survey and census data.

To overcome this limitation, such information must be synthesized. Existing activity-chain synthesis approaches can generally be classified into constraint-based, utility-based, rule-based, learning-based, and hybrid models (Nguyen, Schweizer, and Rupi 2025). Constraint-based models evaluate whether activity schedules are feasible within given space–time constraints, while utility-based approaches rely on nested logit and discrete choice frameworks to model sequential decisions regarding activities, timing, destinations, and transport modes (Yagi and Mohammadian 2009). However, these methods often assume rational

decision-making and perfect information, which may not accurately represent real human behavior (Raux, T.-Y. Ma, and Cornelis 2016). Rule-based models attempt to improve behavioral realism by generating activity schedules through empirical distributions and heuristic decision rules derived from observed travel behavior (Roorda, Miller, and Habib 2007).

Recently, probabilistic and machine-learning-based approaches for activity-chain synthesis have gained increasing attention due to their ability to capture complex dependencies and stochastic behavioral patterns (Balać, Sallard, and B.V. 2023; Joubert and Waal 2020). Ismaïl Saadi et al. (2016) combined population synthesis with Profile Hidden Markov Models (pHMMs) techniques to assign realistic activity sequences to synthetic individuals using travel survey data. Bayesian Networks (BNs) have also emerged as a suitable approach because they can represent causal relationships between socio-demographic variables, activity participation, and travel decisions while integrating multiple data sources within a single probabilistic framework (Joubert and Waal 2020; Balać, Sallard, and B.V. 2023).

An important consideration when applying Bayesian Network is the availability of sufficient training data. Previous studies have synthesized activity-chain by implementing Bayesian Network model using dataset containing thousands of observations. Joubert and Waal (2020) developed Bayesian Network model to generate activity chain using roughly 7,500 observations extracted from a travel survey. While Balać, Sallard, and B.V. (2023) applied the same approach to Swiss mobility survey data containing approximately 57,000 observations.

Bayesian Networks can also be used to represent activity schedules as sequential processes, where the probability of performing a given activity depends on both individual characteristics and previously performed activities. This structure allows the model to capture temporal dependencies within daily activity chains while preserving behavioral heterogeneity across individuals. Furthermore, structure-learning algorithms let the network to learn directly from observed travel-survey data, reducing the need for manually specified behavioral rules (Joubert and Waal 2020; Balać, Sallard, and B.V. 2023).

Alongside model-based approaches, statistical matching techniques have been increasingly adopted to transfer observed behavioral information from travel-survey respondents to synthetic individuals sharing similar socio-demographic characteristics (Sebastian Hörl and Balac 2021). Rather than estimating behavioral decision processes explicitly, statistical matching assigns activity chains observed in survey data to synthetic individuals using a set of matching variables such as age, sex, household size, education level, employment status, and car ownership.

In this study a Bayesian Network approach is adopted to generate activity chains for the synthetic population of Torino, which is explained in more detail in Section 3.8.

# Chapter 3

## Methodology

This chapter discusses the method used for generating synthetic population and activity chain assignment. In this thesis, Synthetic Reconstruction (SR) method is adopted to implement the population synthesis procedure. The core of this method is based on the IPF algorithm, as described in Subsection 2.1.1.1. According to the explanations given there, the needed input data are:

1. Aggregate counts for census tracts from census statistics.
2. Sample microdata from census.

Both datasets will be duly introduced in Section 4.2.

### 3.1 Control Variables

The selection of Control variables represents a crucial step in synthetic population generation process (see Subsection 2.1.1.3 for detailed definition of control variables). This determines which variable or attribute of the population are reproduced exactly at the census tract level (Beckman, Baggerly, and McKay 1996; Lovelace and Dumont 2016). Control variables were selected by two main criteria:

1. Availability of reliable census marginals at the desired spatial resolution.
2. Relevance of the variables for subsequent modelling and analytical applications (Tanton and Edwards 2013; Lovelace and Dumont 2016).

For each spatial unit, marginal distributions for the selected constraint variables were extracted from the census statistics, as described in detail in Section 4.3. These distributions provide the totals counts for each variable category and serve as target value that the synthetic population must reproduce exactly (Beckman, Baggerly, and McKay 1996; Ballas, Clarke, and Wiemers 2005; Lovelace and Dumont 2016). Since aggregate

data does not provide full joint distributions across all selected control variables, the joint structure was derived from the survey microdata and adjusted through IPF by respecting the marginal constraint.

Special attention was given to the feasibility of the constraint system. In particular, structural zeros-combinations of categories that are logically impossible (for instance, a 10 years old person who has university degree and is employed) were identified prior to the IPF procedure. For example, certain employment categories may not apply to specific age groups. When necessary, sparse or near-zero cells were aggregated to avoid convergence instability of algorithm. This ensures that the seed table derived from the microdata contains strictly positive cells for all admissible combinations. Seed table definition is described in Subsection 2.1.1.2.

However, in practice, due to using a limited sample size, several admissible combinations of control variables were not observed in the microdata, resulting in zero-valued cells in the seed contingency table. Since the Iterative Proportional Fitting (IPF) algorithm requires strictly positive initial values to guarantee convergence and avoid fixed zero constraints, a small positive constant " $\varepsilon = 1 \times 10^{-6}$ " was assigned to these empty but logically admissible cells before running the algorithm. The value of  $\varepsilon$  was chosen sufficiently small so do not change the joint distribution, while ensuring numerical stability and feasibility of the estimation process (Beckman, Baggerly, and McKay 1996).

The synthetic procedure was implemented in R by using MIPFP package, as reported in Appendix A.7, in, independently for each spatial unit. This means that a separate IPF procedure was applied for each census tract. This approach guarantees that the marginal distributions of selected variables match the observed total counts from census statistics at census tract level (Lovelace and Dumont 2016). The manipulation of these control variables to the Torino case study is presented in Chapter 4.

## 3.2 Theoretical Formulation of IPF

The classical IPF algorithm operates by iteratively adjusting the entries of an initial contingency table to match target marginal distributions (Deming and Stephan 1940; Fienberg and Meyer 1972; Beckman, Baggerly, and McKay 1996). Let  $X = (x_{ij})$  denote the initial table and let  $M = (m_{ij})$  represent the adjusted table. At the beginning of the procedure, the adjusted table is initialized as:

$$m_{ij}^{(0)} = x_{ij}$$

The algorithm proceeds by alternately adjusting the table to satisfy row and column marginal constraints.

First, the elements of the table are scaled to match the target row marginals  $u_i$ . This

is done through a weighted normalization:

$$m_{ij}^{(1)} = \frac{m_{ij}^{(0)} u_i}{\sum_j m_{ij}^{(0)}}$$

Next, the table is adjusted to match the target column marginals  $v_j$ :

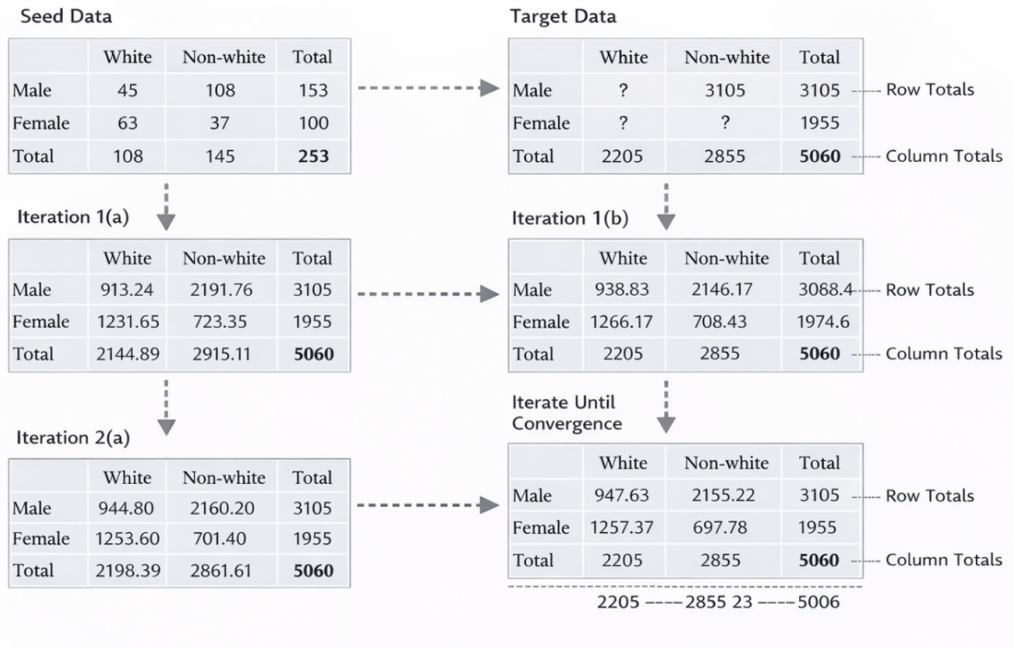
$$m_{ij}^{(2)} = \frac{m_{ij}^{(1)} v_j}{\sum_i m_{ij}^{(1)}}$$

These two steps, row adjustment and column adjustment, constitute one full iteration of the IPF algorithm. The process is repeated iteratively, updating the table at each step, until the marginal totals of the adjusted table converge to the target marginals within a predefined tolerance level.

The IPF algorithm can be extended from two-dimensional tables to multi-dimensional contingency tables. In higher dimensions, the same principle applies: each iteration consists of sequential adjustments across all marginal constraints corresponding to different variables (Müller and Axhausen 2012).

The key property of IPF is that it produces a table that matches the specified marginal totals while minimizing distortion from the original joint distribution. Figure 1 shows a numerical example of the algorithm.

Figure 1: Numerical example of IPF (source: Zhuo (2017))



### 3.3 Integerisation (TRS)

The Iterative Proportional Fitting (IPF) procedure produces fractional values for each cell of the contingency table (Beckman, Baggerly, and McKay 1996; Guo and Bhat 2007; Müller and Axhausen 2010; Lovelace and Dumont 2016). Although these values satisfy the marginal constraints within a numerical tolerance, they are not directly convertible to individuals, since a synthetic population must be an integer number.

Therefore, an integerisation procedure was applied to convert the fractional IPF outputs into integer cell counts while preserving the total population of each census tract and minimizing deviation from the calculated joint distribution (Beckman, Baggerly, and McKay 1996).

The rationale for integerisation based on the TRS method introduced by Lovelace and D. Ballas (2013) is the following.

Let  $\hat{X}_{z,i}$  denote the calculated value by IPF for census tract  $z$  and joint category  $i$ .

In general,

$$\hat{X}_{z,i} \in \mathbb{R}^+$$

It is required to have integerised counts since individuals cannot be fractional:

$$X_{z,i}^* \in \mathbb{N}$$

such that for each census tract  $z$ ,

$$\sum_i X_{z,i}^* = T_z$$

where  $T_z$  is the total fitted population of tract  $z$ , computed as the row sum of the fitted IPF table and rounded to the nearest integer to ensure consistency with discrete totals.

The integerisation procedure must therefore:

- Produce integer-valued counts for each cell in the contingency table;
- Preserve the total population of each census tract exactly;
- Minimize distortion from the fitted IPF joint distribution.

#### **Truncate–Replicate–Sample (TRS) Integerisation**

Integerisation was performed using the Truncate–Replicate–Sample (TRS) method (Lovelace and D. Ballas 2013), adapted here to operate directly on fitted contingency table cell counts rather than on individual survey weights.

The procedure is applied independently to each census tract and consists of three steps.

### Step 1: Truncation

For each calculated cell value  $\hat{X}_{z,i}$ , the integer part is retained:

$$X_{z,i}^{\text{floor}} = \lfloor \hat{X}_{z,i} \rfloor$$

This truncation guarantees that

$$\sum_i X_{z,i}^{\text{floor}} \leq T_z.$$

Truncation separates the calculated value into two components:

$$\hat{X}_{z,i} = X_{z,i}^{\text{floor}} + f_{z,i}$$

where

$$f_{z,i} = \hat{X}_{z,i} - \lfloor \hat{X}_{z,i} \rfloor$$

is the fractional remainder, with

$$0 \leq f_{z,i} < 1.$$

### Step 2: Compute Remaining Units

The number of individuals still to be allocated is

$$R_z = T_z - \sum_i X_{z,i}^{\text{floor}}.$$

Since truncation reduces totals,  $R_z$  is a non-negative integer.

At this stage, the tract population is under-allocated by exactly  $R_z$  individuals.

### Step 3: Probabilistic Allocation

The remaining  $R_z$  individuals are allocated probabilistically across cells according to the magnitude of the fractional remainders.

The selection probability for cell  $i$  is

$$p_{z,i} = \frac{f_{z,i}}{\sum_j f_{z,j}}.$$

Cells are then randomly selected without replacement, and one unit is added to each selected cell until all  $R_z$  units are distributed:

$$X_{z,i}^* = \begin{cases} X_{z,i}^{\text{floor}} + 1, & \text{if selected,} \\ X_{z,i}^{\text{floor}}, & \text{otherwise.} \end{cases}$$

The TRS integerisation applied to the fitted contingency table guarantees that:

- All cell counts are integers;
- The total population per census tract is preserved exactly;
- Cells with larger fractional remainders have higher probability of receiving an additional unit;
- The expected value of the integerised distribution equals the fitted IPF distribution;
- Spatial variation across census tracts is maintained, as each tract is integerised independently.

Unlike simple rounding, TRS makes full use of the fractional information contained in the fitted IPF solution, thereby minimizing the distortion introduced by integerisation (Lovelace and D. Ballas 2013).

The resulting integerised contingency table forms the discrete synthetic population structure for each census tract and constitutes the basis for the generation of individual synthetic agents in the subsequent section.

### 3.3.1 Numerical Example of TRS

Consider four fractional weights for four individuals as follows.

Table 3: TRS microdata (source: own elaboration)

| Individual | Fractional Weight |
|------------|-------------------|
| A          | 2.7               |
| B          | 0.4               |
| C          | 1.9               |
| D          | 3.2               |

So, the total target population is

$$2.7 + 0.4 + 1.9 + 3.2 = 8.2.$$

In TRS the first step is “truncation”, which means separating the integer part from the decimal part.

Table 4: First step of TRS (source: own elaboration)

| Individual | Weight | Truncated part | Decimal remainder |
|------------|--------|----------------|-------------------|
| A          | 2.7    | 2              | 0.7               |
| B          | 0.4    | 0              | 0.4               |
| C          | 1.9    | 1              | 0.9               |
| D          | 3.2    | 3              | 0.2               |

So, after truncation there are already

$$2 + 0 + 1 + 3 = 6$$

individuals in the synthetic population. This is the “replicate” step.

- Individual A appears 2 times.
- Individual B appears 0 times.
- Individual C appears 1 time.
- Individual D appears 3 times.

If we round the target total,

$$8.2 \approx 8,$$

then

$$8 - 6 = 2,$$

meaning that two additional individuals are needed.

The remaining individuals are added by sampling from the records using probabilities proportional to the decimal remainders.

The total remainder is

$$0.7 + 0.4 + 0.9 + 0.2 = 2.2.$$

Therefore,

$$P(A) = \frac{0.7}{2.2} = 0.318$$

$$P(B) = \frac{0.4}{2.2} = 0.182$$

$$P(C) = \frac{0.9}{2.2} = 0.409$$

$$P(D) = \frac{0.2}{2.2} = 0.091.$$

Now sample two extra individuals based on these probabilities.

One possible outcome is

- Extra draw 1: C.
- Extra draw 2: A.

The final integerised population is

$$[A, A, A, C, C, D, D, D].$$

**Note:** Because of the sampling step, a different run could give a slightly different final result.

### 3.4 Generation of Synthetic Individuals

Following integerisation, the calculated contingency tables contain integer counts for each joint combination of control variables. However, these counts represent aggregated category totals rather than individual records. For instance, the IPF result indicates that there are 20 male individuals in census tract A who are employed and have a university degree.

Therefore, it is required to take the final step to generate a fully disaggregated synthetic population composed of individual agents.

The synthetic individuals were generated using a sampling approach. The microdata used to construct the initial seed table also serves as an individual pool for the sampling procedure. Each microdata record represents a real observed individual and contains socio-demographic relationships across variables.

For each individual in the microdata, a joint key was constructed. For instance,

$$\text{Key} = (\text{Sex}, \text{Age group}, \text{Education}, \text{Employment})$$

Microdata records were grouped by this key, forming a set of donor pools corresponding to the possible joint combinations as mentioned in Subsection 2.1.1.2.

For each census tract  $z$  and each joint category  $i$ , the integerised count  $X_{z,i}^*$  obtained from the TRS procedure determines the number of synthetic individuals that must be generated.

The allocation process proceeds as follows:

1. Identify the donor pool corresponding to the joint category  $i$ .

2. Randomly sample  $X_{z,i}^*$  individuals from that pool.
3. Due to the small number of real individuals in the microdata, sampling is performed with replacement for the selection of individuals.
4. For each sampled individual:
  - A new synthetic identifier is assigned.
  - The census tract identifier is attached. Since the microdata used in this study does not contain information on residential location, it is defined at the beginning of the process which individuals belong to which census tract based on the IPF and TRS results. Once each census tract is populated with the required number of individuals, the corresponding census tract identifier is assigned to them as a spatial attribute.
  - The controlled attributes (sex, age group, education, employment) are instead retained from the donor.

Sampling with replacement ensures that population totals can be matched even when the microdata contains limited observations for certain combinations. To guarantee reproducibility of the sampling process, a fixed random seed in Python was used.

Although the synthetic population is generated from the joint distribution observed in the microdata or seed table, in rare cases, some logically possible combinations may have no corresponding records in the microdata. This situation happens because the microdata represents a limited sample of the population, and rare combinations may simply not be observed. During the IPF procedure, marginal constraints derived from census data may require positive population counts in such combinations. Since a small positive constant ( $\varepsilon$ ) was added to all seed cells to ensure that the IPF algorithm performs well, the algorithm is able to allocate mass to these previously unobserved but logically possible combinations. After integerisation, this may result in one or more synthetic individuals belonging to categories for which no donor record exists in the microdata. These cases reflect a sampling problem, which is a known characteristic of spatial microsimulation methods based on reweighting techniques (Lovelace and Dumont 2016). Those individuals were flagged as “no donor for cell”. This ensures that the marginal totals remain consistent, while making these rare and potentially unrealistic cases traceable for post-processing.

The final synthetic population dataset contains one row per synthetic individual. Each row includes:

- Synthetic individual identifier;
- Census tract identifier (Zone);

- Donor reference identifier;
- Sex;
- Age group;
- Educational attainment;
- Employment status.

At this stage, the synthetic population exactly reproduces, at the census tract level, and preserves realistic multivariate relationships derived from the microdata in:

- Age–sex distributions;
- Education–sex distributions;
- Employment–sex distributions.

This synthetic individual population serves as the baseline dataset for subsequent enrichment procedures, including the assignment of additional socio-economic attributes and household structures, as described in the next sections. The generation of synthetic individuals based on this approach is detailed in Section 4.5.

### 3.5 Enrichment Concept

Although the IPF-based synthetic population reproduces the joint distribution of the selected control variables, additional attributes are required for the intended analyses and for building a more realistic representation of individuals and households.

These variables were not constrained directly in the IPF procedure for two main reasons. First, lack of input data. Reliable marginal distributions are not available for several of these variables. Second, including additional dimensions in IPF would increase the number of joint cells, which will lead to empty combinations in the seed table and cause difficulty for the IPF algorithm to converge.

For these reasons, the synthesis step was limited to a set of core demographic control variables, while the remaining variables were assigned afterwards through a post-processing step. Specifically, since each synthetic individual is generated by sampling from the microdata, a donor identifier is retained, allowing additional socio-economic variables (e.g., income, household characteristics, vehicle ownership) to be transferred directly from the corresponding donor record through a merge operation (S. Hörl et al. 2026).

## 3.6 Household Synthesis

Although the microdata considered in this study include a household sample with all its members, a key limitation arises from the lack of detailed aggregate household information at the census tract level. In particular, variables such as household income, number of vehicles, and family type are not available as marginal distributions, and some of them—such as income—are not even present in the household microdata that we used (see Section 3.8). Moreover, the only household-level variable relevant for transportation analysis for which aggregate counts are available at the census tract level is household size.

As a result, a fully household-based synthesis approach could not be implemented. Instead, the adopted methodology in this study combines elements of the different household synthesis strategies described in Subsection 2.1.4. The detailed procedure for household construction is presented in Section 4.5.

## 3.7 Validation Framework and Metrics

The validation of the synthetic population was performed using both internal validation, described in Subsection 2.2.1, and external validation, described in Subsection 2.2.2.

Internal validation was performed by comparing the total counts of the synthetic population with the corresponding census marginal totals for the control variables. For each variable, the synthetic population was aggregated at the census tract level and compared with observed counts from census statistics.

External validation was performed by comparing the aggregated counts, from census statistics, of variables not used as constraints in the synthetic population generation process.

### 3.7.1 Validation Metrics

To evaluate the model fit for both internal validation and external validation, two goodness-of-fit measures were used in this study.

Consider:

- $obs_{ij}$  denotes the observed census count for category  $j$  in zone  $i$ ;
- $sim_{ij}$  denotes the simulated count from the synthetic population;
- $e_{ij} = sim_{ij} - obs_{ij}$  denotes the error.

The following metrics were computed.

### Pearson's Correlation Coefficient ( $r$ )

Pearson's correlation coefficient ( $r$ ) measures the linear relationship between two vectors. In spatial microsimulation, if the model performs correctly, the observed counts and the final simulated counts for each category within every zone should be identical in internal validation. Consequently, when plotting the observed counts against the final counts for each category, the points should lie on a perfect straight line where the values on the x-axis (abscissa) equal those on the y-axis (ordinate). Therefore, a measure of linear correlation is appropriate for evaluating the model.

The Pearson correlation between the vectors  $obs$  and  $sim$  is calculated using the following formula:

$$r = \frac{\sum(obs - \bar{obs})(sim - \bar{sim})}{\sqrt{\sum(obs - \bar{obs})^2} \sqrt{\sum(sim - \bar{sim})^2}}$$

A value of  $r = 1$  indicates perfect agreement.

In spatial microsimulation, values above 0.9 are typically expected when constraints are correctly implemented (Lovelace and Dumont 2016).

### Total Absolute Error (TAE) and Relative Error (RE)

The Total Absolute Error (TAE) measures the total value of deviations between observed and simulated values. It is defined as

$$TAE = \sum_i \sum_j |obs_{ij} - sim_{ij}|.$$

TAE provides a measure of the total discrepancy between observed and simulated data and can be interpreted as the cumulative number of misallocated individuals across all categories and census tracts. However, TAE is sensitive to the scale of the study area, as it increases with population size. To address this limitation, another metric, known as Relative Error (RE), was used.

To enable comparison across different census tracts and different categories for each variable, the Total Absolute Error is normalized by the total population of that census tract or category. The Relative Error (RE) is defined as

$$RE = \frac{TAE}{\sum_i \sum_j obs_{ij}}.$$

RE represents the proportion of the total population that is misallocated by the model and can therefore be interpreted as a percentage error. Unlike TAE, this metric is independent of the absolute size of the population, making it more suitable for comparative analysis.

### 3.8 Activity Chain Generation

As mentioned in Section 2.3, this work employs Bayesian Networks to generate activity chains for the synthesized population. Bayesian Networks are probabilistic graphical models that indicate conditional dependencies between variables through a directed network structure (Jensen and Nielsen 2007). A Bayesian Network consists of two main components: the directed graph or structure, composed of nodes and directed links, and the parameters, represented by numerical probabilities. In activity chain synthesis they are used to model the relationships between socio-demographic characteristics and activity or travel behavior. Nodes represent variables such as age, vehicle ownership, employment status, or work trips, while arrows indicate dependency relationships between variables. For example, employment status may influence the likelihood of making a work trip, and the strength of this relationship is quantified through the model probabilities. In this case the employment status is called parent, and work trip is a child of employment status (Joubert and Waal 2020; Balać, Sallard, and B.V. 2023). The learning process will be explained in more detail in Subsection 4.9.3.

Unlike the population synthesis procedure described in Section 3.4, which relies on the 1% census microdata and census statistics, the activity-chain generation stage uses the dataset generated by the mobility survey as its primary data source. Mobility surveys in fact typically contain information on socio-demographic variables, daily travel behavior, driving-license ownership, and activity participation, which are not available in either the census data or the synthetic population generated through IPF. Therefore, this additional dataset was used to estimate additional mobility-related attributes and assign daily activity chains to synthetic individuals.

In this study, Bayesian Networks were implemented using the Python `pgmpy` (Ankan and Panda 2015) library to estimate driving-license ownership and activity-chain sequences for the synthetic population. The modelling process included data preprocessing, structure definition or learning, parameter estimation, and assignment of activity-chain patterns to synthetic individuals.

Two Bayesian Network models were implemented sequentially. First, a Bayesian Network was used to estimate driving-license ownership for synthetic individuals. The generated driving-license variable was then used as an explanatory variable in the activity-chain generation model, allowing driving-license ownership to influence the generated daily activity patterns.

Before implementing the Bayesian Network model, the mobility survey and synthetic population datasets were preprocessed to ensure consistency among variable categories. Socio-demographic variables were recoded into common categories and driving-license ownership was estimated by using a Bayesian Network model. This process is explained in detail in Subsection 4.9.3.

The mobility survey considered in this study records daily activities that do not always directly correspond to complete home-based activity chains, given the time boundaries of travel diaries. In some cases, the recorded sequence began from an activity occurring later in the day rather than from the first activity performed in the morning. Therefore, a preprocessing procedure was required to reconstruct ordered daily activity sequences. Activity episodes were first ordered chronologically and subsequently reorganized to ensure a consistent home-based representation of daily activity behavior. These reconstructed activity chains formed the input dataset used for training the Bayesian Network model. The implementation of this procedure is described in Subsection 4.9.3.

Bayesian Network structures can be developed either by expert knowledge or learned from data. In this study the network structure was learned from the mobility survey dataset in order to capture the relationship between socio-demographic characteristics and activity chains. Although the network structure was learned from the mobility survey data, several structural constraints were imposed to ensure behavioral plausibility. In particular, socio-demographic variables were specified as root nodes, and temporal constraints were introduced to prevent later activities from influencing earlier activities. The specific constraints adopted in this study are described in Subsection 4.9.3.

For activity-chain generation, daily activity schedules were represented as a sequence of activity positions. Each position in the sequence was represented by a separate Bayesian Network node representing one specific activity, while socio-demographic variables and driving-license ownership were included as explanatory nodes. Following the approach proposed by Joubert and Waal (2020), the network structure was learned from the mobility survey data subject to a set of structural constraints:

- Socio-demographic variables were specified as root nodes and were therefore only allowed to influence activity variables.
- Temporal constraints were imposed so that only earlier activities could affect activities that happened later in the day.

This representation allows the model to capture both the influence of individual characteristics on activity participation and the sequential dependencies between consecutive activities.

Because daily activity schedules contain different numbers of activities, a fixed-length representation was adopted for Bayesian Network estimation, following Joubert and Waal (2020). Each activity chain was represented by six sequential activity positions (labelled `Activity_1` to `Activity_6`). Activity chains containing fewer than six activities were completed using a special `Stop` state indicating the end of the daily sequence. For example, an activity chain represented as “Home–Work–Leisure–Home” was transformed into “Home–Work–Leisure–Home–Stop–Stop”. Activity chains containing more than six

activities were excluded from the modelling dataset. This representation provided a consistent structure for all observations while reducing model complexity and maintaining the activity patterns observed for the vast majority of respondents. The implementation of this procedure is described in Subsection 4.9.3.

To improve the performance of the model and reduce the model complexity, individual activity purposes were aggregated into a smaller set of activity groups before model estimation. The detailed aggregation scheme adopted in this study is presented in Subsection 4.9.3.

After defining the network structure, the model parameters were estimated using the mobility survey dataset. This process involved estimating the conditional probability distributions associated with the connected variables in the network. For instance, the model estimates the probability of an employed individual performing a work activity conditional on employment status and previously performed activities.

These estimated conditional probabilities were then used to generate activity-chain patterns for synthetic individuals probabilistically. This allows different activity chains to be assigned to individuals with similar socio-demographic characteristics in order to reproduce more realistic and heterogeneous activity-chain patterns rather than identical activity patterns for similar individuals (Joubert and Waal 2020; Balać, Sallard, and B.V. 2023).

Once the Bayesian Network parameters were estimated, the model was applied to the synthetic population. For each synthetic individual, the socio-demographic characteristics and driving-license ownership were used as input variables. Based on the probability relationships learned from the mobility survey, the model generated a daily activity chain. This approach allows individuals with similar characteristics to receive different activity chains while preserving the behavioral variability observed in the survey data.

The mobility survey dataset was split into training and testing parts, where 80% of the observations were used to train the model and 20% were used for validation. Model performance was evaluated by comparing the distribution of activity chains generated by the Bayesian Network with the observed activity-chain distribution in the mobility survey. In addition, spatial consistency validation was conducted after assigning activity chains to the synthesized population. This validation compared the number of work trips originating from each census tract in the synthesized population with work-trip production statistics derived from the census commuting matrix at census-tract level. A synthesized individual was considered to generate a work trip if at least one work activity was present in the assigned activity chain. The objective was to assess whether the activity-chain generation procedure reproduced the spatial distribution of work-trip origins across the study area. Detailed validation results are presented in Subsection 4.9.4

The output of the activity-chain generation process consists of a set of activity-sequence variables representing the ordered activities performed by each synthetic in-

dividual throughout the day. Each generated activity chain is represented by up to six activity positions, providing a representation of daily activity behavior suitable for subsequent travel-demand modelling and analysis. However, the methodology generates only the sequence of daily activities (e.g., “Home–Work–Home” or “Home–Shopping–Home”) and does not assign the spatial locations where these activities take place or the trips connecting consecutive activities. Therefore, the output of this stage is an activity schedule rather than a complete synthetic travel demand dataset. The implementation of the model is presented in Chapter 4.

# Chapter 4

## Case Study

### 4.1 Study Area and Zoning

The study focuses on the city of Torino, the second largest metropolitan area in northern Italy and the capital of the Piedmont region. Torino has a population of approximately 850 thousand inhabitants within the municipality and a much larger population in the surrounding metropolitan area according to ISTAT.<sup>1</sup>

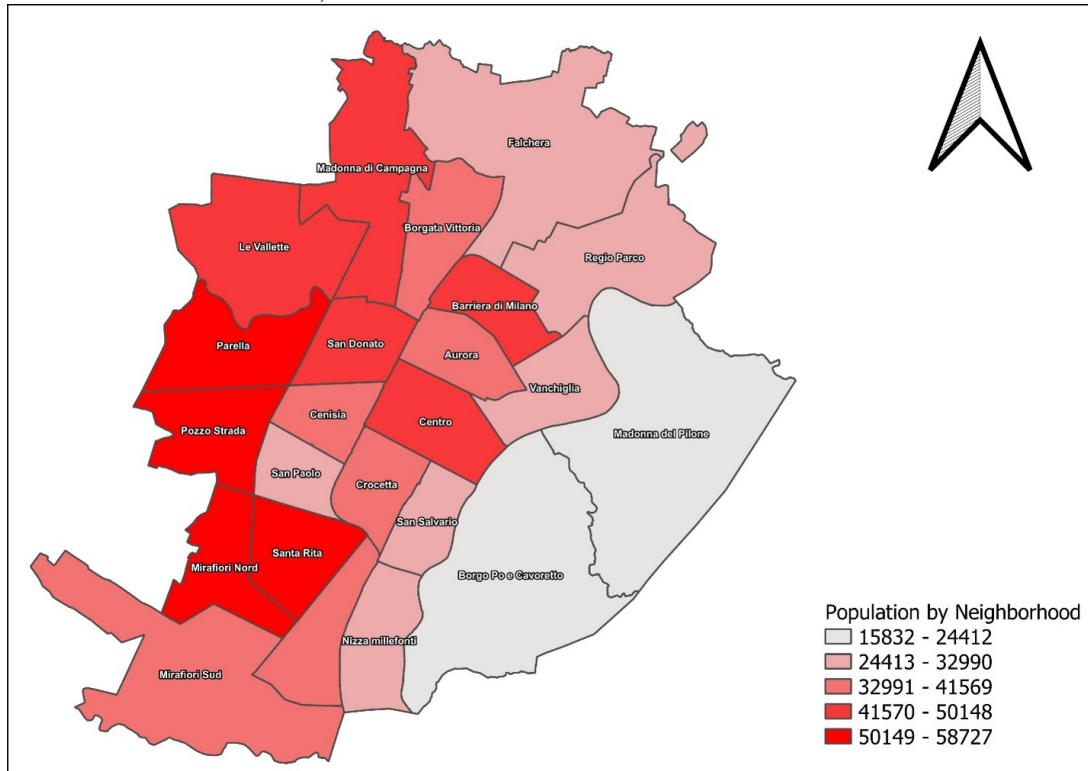
In order to do urban planning, policy analysis and socio-economic research, understanding the spatial distribution of demographic and socio-economic characteristics within the city is needed. There are many official datasets which provide information either at an aggregated spatial level or without detailed geographic identifiers. This limitation makes it difficult to analyze population characteristics at a fine spatial level. Therefore, to address this problem the present study aims to generate a synthetic population that represents individuals and households in the study area while respecting the available aggregate information. This objective is implemented through the population synthesis procedure described in Section 3.4.

The spatial units used in this study are based on two hierarchical geographic levels. The first and larger spatial unit is the neighborhood (*quartiere*). Figure 2 illustrates the spatial distribution of Torino's neighborhoods, showing the division of the municipality into 23 administrative units.

---

<sup>1</sup><https://www.istat.it/wp-content/uploads/2021/11/TORINO-infografica.pdf>

Figure 2: Spatial distribution of population across Torino's neighborhoods (source: own elaboration on census 2011)

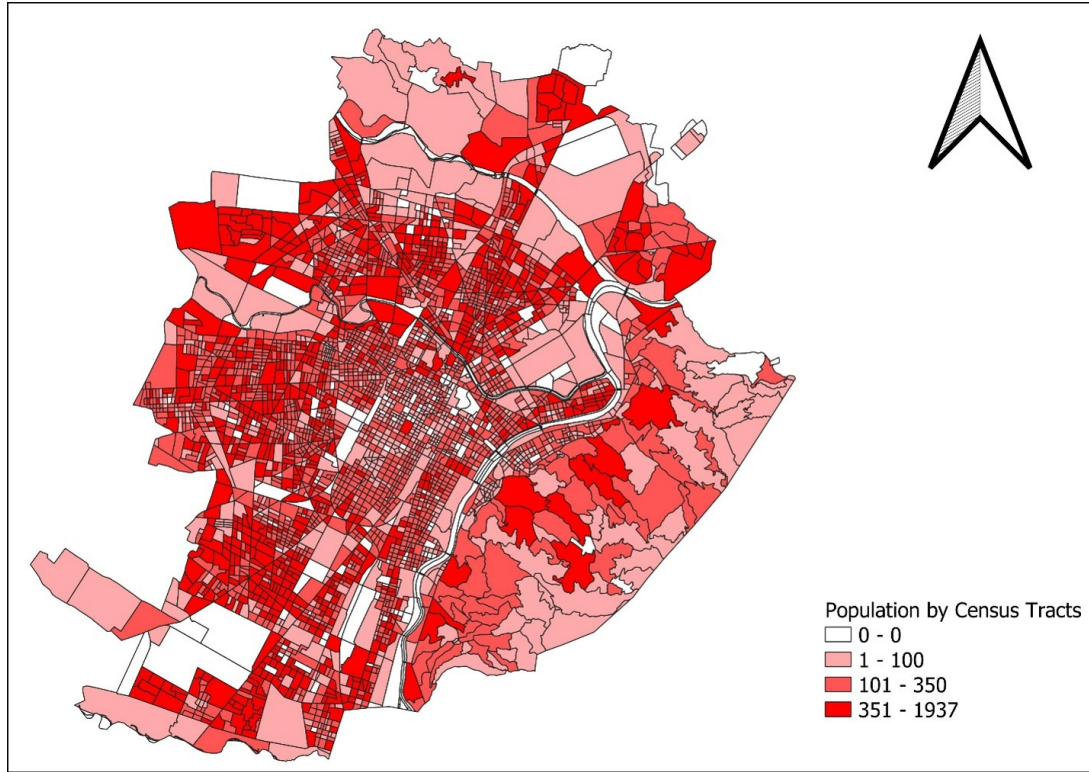


The smallest spatial unit is the census tract (*sezione di censimento*), which represents the basic territorial division used in the national census conducted by the Italian National Institute of Statistics (ISTAT). Census data do not provide detailed information at the individual level at a fine spatial level; instead, they offer aggregated data (e.g., the number of male individuals aged 25–29), along with other variables described in Appendix A.1. This limitation motivates the use of microdata-based reconstruction, as explained in Section 3.4.

The municipality of Torino is divided into 3713 census tracts according to the 2011 Census.<sup>2</sup> From those 3713 census tracts, 169 tracts were removed due to having no population. Figure 3 illustrates the spatial distribution of the population across census tracts in Torino, highlighting the variation in population size at the local level.

<sup>2</sup><https://www.istat.it/notizia/basi-territoriali-e-variabili-censuarie/>

Figure 3: Spatial distribution of population across Torino's census tracts (source: own elaboration on census 2011)



## 4.2 Data Preparation

This study is based on five datasets, each characterized by a different spatial resolution and set of variables. These datasets are used at different stages of this study, as summarized in Table 5. The role of each dataset in the synthesis workflow follows the methodological structure presented in Section 3.1 and Section 3.4.

Table 5: Summary of datasets used in the study (source: own elaboration).

| Dataset                                                                           | Spatial resolution | Unit              | Main variables                                                                                     | Reference (source) | Year |
|-----------------------------------------------------------------------------------|--------------------|-------------------|----------------------------------------------------------------------------------------------------|--------------------|------|
| Census statistics                                                                 | Census tract       | Aggregated counts | Age, sex, education, employment, household size                                                    | ISTAT <sup>3</sup> | 2011 |
| Census 1% individual microdata                                                    | Piedmont region    | Individual        | Age, sex, education, employment status, household variables                                        | ISTAT <sup>4</sup> | 2011 |
| Census 1% household microdata                                                     | Piedmont region    | Household         | Number of vehicles in household                                                                    | ISTAT <sup>4</sup> | 2011 |
| Survey on mobility and modal diversion propensity in the Torino metropolitan area | Geocoded location  | Individual        | Age, sex, education, employment, household attributes, income-related and socio-economic variables | Ceccato (2020)     | 2017 |
| Census commuting matrix (work and study trips)                                    | Census tract       | OD flow           | Number of workers and students traveling between census tracts                                     | ISTAT <sup>5</sup> | 2011 |

The population synthesis combines different datasets that play distinct roles. Census data provide the marginal constraints used in the IPF procedure, as discussed in Subsection 2.1.1.1, and are also used for validation, as described in Section 2.2. The 1% census individual microdata provide the joint distribution, as explained in Subsection 2.1.1.2, and the donor records required for generating and enriching synthetic individuals. The 1% census household microdata provide donor records at the household level, as described in Subsection 2.1.4.1. The mobility survey dataset is finally used for the generation of activity-chain patterns for the synthetic population. For more information about how the survey was conducted, as well as all the variables and their categories, see Riccardo Ceccato (2020) Ph.D. dissertation. In addition, the census commuting matrix provides origin–destination information on work and study trips at the census-tract level and was used for validation of the generated work-trip distributions.

<sup>3</sup>Italian National Institute of Statistics (ISTAT), *Territorial bases and census variables*. <https://www.istat.it/notizia/basi-territoriali-e-variabili-censuarie/>

<sup>4</sup>Italian National Institute of Statistics (ISTAT), *15th Italian General Census of Population and Housing (2011) – 1% Sample*. <https://www.istat.it/en/microdata/15-italian-general-censuses-of-population-and-housing-2011-1-sample/>

<sup>5</sup>Italian National Institute of Statistics (ISTAT), *Contiguity, distance and commuting matrices*. <https://www.istat.it/notizia/matrici-di-contiguita-distanza-e-pendolarismo/>

### 4.2.1 Census Statistics Preprocessing

The primary source of population statistics used in this study is the Italian 2011 official 15th general population and housing census conducted by the Italian National Statistical Office (ISTAT). The 2011 population and housing census is considered here because it was the last traditional Italian census based on a universal, once-every-ten-years survey covering the entire population.<sup>6</sup> Starting from 2018, Italy shifted to a permanent census, replacing the full enumeration with a continuous system that combines administrative data and annual sample surveys involving about 1.4 million households across 2,800 municipalities. Census provides demographic information about the resident population and is widely used for demographic analysis and spatial studies. Since the census covers the entire population, it represents a reliable source for aggregated demographic information at a small spatial scale. For the purpose of this study, census data were used at the census-tract level, which is the smallest possible level for population synthesis. At this level, census data provide aggregated demographic and socio-economic information about the population. The aggregated information is useful because it can be used as the marginal constraints or target population to guide the population synthesis process. The use of census marginals as target constraints is discussed in Section 3.1. A complete list of the available census variables used in this study is provided in Appendix A.1.

Before deriving the control variables, the original census table was preprocessed in several steps. First of all, all the data related to other municipalities were removed from the dataset by filtering the **COMUNE** variable to Torino. Second, all census tracts with total population equal to zero were removed by filtering the **P1** variable equal to zero. Variables not relevant to the selected control variables or not required for the synthesis procedure were removed. Variables related to geographic identifiers, and variables related to sex, age, education and employment status were finally retained, as shown in Table 6.

Table 6: Census variables used for deriving control variables (source: census aggregate dataset).

| Variables               | Definition                  |
|-------------------------|-----------------------------|
| P1, P2                  | Total and male population   |
| P17 to P29              | Age groups (15+)            |
| P33 to P45              | Male age groups (15+)       |
| P47, P48, P49, P50, P52 | Education variables (total) |
| P54, P55, P56, P57, P59 | Education variables (male)  |
| P61, P65                | Employed (total and male)   |

Based on the census variables presented in Table 6, some other variables were derived to serve as control variables for the population synthesis procedure. These derived variables

<sup>6</sup>Italian National Institute of Statistics (ISTAT), *Permanent Census of Population and Housing*. <https://www.istat.it/en/non-categorizzato/permanent-census-of-population-and-housing/>

are required since the original census variables or classifications are not consistent with the requirements of the population synthesis procedure. For instance, the distribution of females by age group was not directly available from the 2011 census dataset, so it was necessary to derive those variables by subtracting the male population from the total population for each age group.

Furthermore, this process ensured internal consistency across all marginal distributions, so that totals remained coherent between variables and could be reliably used as constraints within the Iterative Proportional Fitting (IPF) procedure. The complete list of derived variables along with their definitions and transformation formulas is provided in Appendix A.2.

For each spatial unit, marginal distributions are extracted for a selected set of control variables including basic demographic characteristics such as age and sex, as well as socio-economic attributes such as educational attainment, employment status, and household size.

#### 4.2.2 1% Sample Census Individual Microdata Preprocessing

While census statistics are aggregated data that provide reliable information about the population, they do not provide the multivariate joint distribution of individual characteristics at a fine spatial level such as the census tract. In other words, although census data report how many individuals belong to each category of variables such as age, sex, education, and employment status for each census tract, they do not indicate how these characteristics are jointly distributed across individuals at that spatial level. To overcome this limitation, the joint distribution was approximated using the available 1% sample census microdata, which contain individual-level census records. These records were used to derive the seed contingency table for the IPF procedure.

For instance, census statistics provide the number of individuals aged 20–24, the number of individuals with university degrees, and the number of employed individuals within each census tract. However, they do not directly reveal how many individuals simultaneously belong to all these groups within the same census tract. Such joint distributions may be available only at more aggregated spatial levels or within microdata samples, but not at the level of census tracts.

In order to overcome this problem, a 1% individual-level census microdata sample from 2011 is used<sup>7</sup> to provide detailed information on demographic and socio-economic characteristics. The microdata consist of records describing individual persons, where each record corresponds to a real person and contains several variables such as age, gender, education level, employment status and household characteristics (see Appendix A.2 for

---

<sup>7</sup>Italian National Institute of Statistics (ISTAT), *15th Italian General Census of Population and Housing (2011) – 1% Sample*. <https://www.istat.it/en/microdata/15-italian-general-censuses-of-population-and-housing-2011-1-sample/>

a detailed description of all attributes and Appendix A.4 for all attribute classifications). Using this dataset, it is possible to capture the joint distribution of variables, which is essential for generating a realistic synthetic population.

The limitation of the microdata is that they only contain spatial identifiers at the regional level, which is very coarse. Therefore, even if the microdata provide detailed information about the population, they cannot provide the spatial distribution of individuals across the study area. In order to address this limitation, the population synthesis approach is employed, which combines the joint distributions observed in the microdata with the marginal distributions provided by the census data.

In the population synthesis framework adopted in this study, the 1% sample census microdata serve as the seed population, providing the initial multidimensional contingency table that represents the joint distribution of selected control variables. During the Iterative Proportional Fitting (IPF) procedure, this initial distribution is iteratively adjusted so that its marginal totals match the aggregated census statistics at the census-tract level. In this way, the synthetic population preserves the realistic relationships observed in the 1% sample census microdata while reproducing the spatial demographic structure observed in the census.

A preprocessing phase was required to ensure consistency between the 1% sample census microdata and the census constraints used in the synthesis procedure. Key variables were recoded to align the categorical structure of the 1% sample census microdata with the categories derived from the census aggregate table. Since the original classifications in the census statistics and the 1% sample census microdata do not always coincide exactly, some categories were aggregated or redefined to obtain a common set of classes. First, the 1% sample census microdata were filtered to retain only records referring to Piemonte, using the regional code `REGIONE = 1`. The variables required for the synthesis procedure were then relabelled as follows.

- Sex was recoded into two categories as shown in Table 7.

Table 7: New labels for the `SESSO` variable in the 1% sample census microdata (source: own definition).

| Original code | New label |
|---------------|-----------|
| 1             | Male      |
| 2             | Female    |

- Educational attainment was recoded into five categories consistent with the census-derived education controls as shown in Table 8.

Table 8: New classification and labels for educational attainment in the 1% sample census microdata (source: own definition).

| Original variable | Original category      | Related definition                       | New category      |
|-------------------|------------------------|------------------------------------------|-------------------|
| TIT_STUD          | 1                      | Illiterate or Primary School Certificate | edu_illit_primary |
| TIT_STUD          | 2                      | Lower Secondary School Certificate       | edu_lowersec      |
| TIT_STUD          | 3                      | Vocational Qualification Diploma         | edu_uppersec      |
| TIT_STUD          | 4                      | High School Diploma                      | edu_uppersec      |
| TIT_STUD          | 5                      | Non-university tertiary diploma          | edu_uppersec      |
| TIT_STUD          | 6                      | University diploma                       | edu_tertiary      |
| TIT_STUD          | 7                      | University degree                        | edu_tertiary      |
| TIT_STUD          | Missing or other cases | edu_other_na                             | edu_other_na      |

- Employment status was simplified into a binary classification. Individuals with `COND_PROF = 1` were classified as employed, while all other conditions were classified as not employed, as shown in Table 9.

Table 9: New classification for employment variables in the 1% sample census microdata (source: own definition).

| Original variable | Original code | Original category                                                              | New category |
|-------------------|---------------|--------------------------------------------------------------------------------|--------------|
| COND_PROF         | 1             | Employed                                                                       | Employed     |
| COND_PROF         | 2             | Seeking first employment                                                       | not_employed |
| COND_PROF         | 3             | Unemployed (seeking a new job)                                                 | not_employed |
| COND_PROF         | 4             | Recipient of one or more pensions due to previous employment or capital income | not_employed |
| COND_PROF         | 5             | Student                                                                        | not_employed |
| COND_PROF         | 6             | Homemaker                                                                      | not_employed |
| COND_PROF         | 7             | In another condition                                                           | not_employed |

Records with missing values for the essential variables, sex and age, were excluded from the 1% sample census microdata.

After preprocessing both datasets, namely the 1% individual census microdata and the aggregated census data, they shared the same classification for the selected control variables. Next, a contingency table was constructed from the 1% sample census microdata using the selected control variables. Each unique combination of categorical attributes defines a cell in the multidimensional table, which serves as the input for the IPF algorithm.

The number of individuals for each joint combination represents the initial joint distribution that will be reweighted to match census marginals at the selected spatial unit, namely the census tract, through the IPF implementation.

## 4.3 Selection of Control Variables

From the broad set of variables available in both aggregated and individual-level datasets, a subset was selected to be used as control variables in the population synthesis procedure. The criteria for selecting control variables are described in Section 3.1. These variables represent key demographic and socio-economic characteristics of the population and are available for all census tracts within the study area. The main variables used, which are also called control variables, include:

- Age, grouped into predefined age classes.
- Sex, distinguishing male and female individuals.
- Educational attainment, representing the highest level of education completed.
- Employment status, indicating whether individuals are employed or not employed.
- Household size, describing the number of individuals living in each household.

For each variable (age group, sex, education level, employment status, and household size), the number of individuals in each category was calculated. These aggregated counts represent the marginal totals used as constraints in the IPF procedure.

The most important reason for selecting these variables is because they are common among all the datasets used for the generation and validation of spatial microsimulation. They are also selected because they capture important structural information of the population and are useful for transport microsimulation modelling.

## 4.4 Iterative Proportional Fitting and Integerisation Implementation

The IPF procedure was implemented in R using the `mipfp` package, specifically through the `Ipfp()` function. The corresponding R script is provided in Appendix A.7. The input seed table that was constructed from the preprocessed 1% census microdata used sex, age group, educational attainment, and employment status attributes to extract joint combinations for individuals. The resulting seed table had four dimensions:

$$\text{Sex} \times \text{Age} \times \text{Education} \times \text{Employment}$$

The related categories have been introduced in Subsection 4.2.2, and therefore their cardinality is as follows: 2 sex categories, 14 age groups, 5 education categories, and 2 employment-status categories. Therefore, the seed table contained

$$2 \times 14 \times 5 \times 2 = 280$$

possible combinations.

Because the census constraints were available for each census tract, the seed table was replicated across all census tracts by adding a spatial dimension, namely the census tract. The final IPF input table therefore had five dimensions:

$$\text{Census tract} \times \text{Sex} \times \text{Age} \times \text{Education} \times \text{Employment}$$

A small positive value,  $\varepsilon = 10^{-6}$ , was added to the seed table before running IPF as explained in Subsection 2.1.1.4. This was done to avoid the zero-cell issue that could prevent the algorithm from converging or prevent the algorithm from allocating population to combinations that are logically possible but not observed in the 1% sample census microdata.

Three sets of marginal constraints were used in the IPF procedure, as shown in Table 10.

Table 10: Marginal constraints used in the IPF procedure.

| Constraint     | Dimensions                            |
|----------------|---------------------------------------|
| Age–sex        | Zone $\times$ Sex $\times$ Age        |
| Education–sex  | Zone $\times$ Sex $\times$ Education  |
| Employment–sex | Zone $\times$ Sex $\times$ Employment |

Higher-order joint distributions (e.g., sex by age by education) could in principle provide stronger constraints for the IPF procedure. However, such multidimensional distributions are not available in the census data at the census-tract level. Therefore, the model relies on the highest level of joint information available, namely two-dimensional distributions (sex by age, sex by education, and sex by employment).

Before running the algorithm, consistency checks were performed to verify that the total population was the same across the three constraint sets for each census tract.

The algorithm was configured with a maximum number of 10,000 iterations and a convergence tolerance equal to  $10^{-5}$ . The tolerance defines the stopping criterion: the iterative process terminates when the maximum absolute deviation between the estimated marginal totals and the corresponding target values falls below this threshold. These parameter values were selected to ensure numerical stability and convergence of the algorithm while maintaining computational efficiency. The IPF algorithm was then executed using the settings shown in Table 11.

Table 11: Settings used for the `mipfp` package.

| Parameter             | Value               |
|-----------------------|---------------------|
| R package             | <code>mipfp</code>  |
| Function              | <code>Ipfp()</code> |
| Maximum iterations    | 10,000              |
| Convergence tolerance | $10^{-5}$           |

The output of the IPF procedure was a five-dimensional fitted table containing fractional cell values for each census tract and each joint combination of the control variables. These values matched the target marginal constraints within the specified numerical tolerance.

Since IPF produces fractional values, an integerisation step was required before generating synthetic individuals. This was implemented using the Truncate–Replicate–Sample (TRS) method. The detailed implementation of the TRS procedure is reported in Appendix A.7. For each census tract, the fractional IPF values were first truncated by taking the floor of each cell value. The remaining population units were then allocated probabilistically to cells according to their fractional remainders. Cells with larger fractional parts therefore had a higher probability of receiving one additional individual. A fixed random seed was set before the TRS integerisation step (`set.seed(123)`, as shown in Appendix A.7) to ensure that the probabilistic allocation of the remaining units was reproducible. The IPF procedure itself is deterministic and therefore does not require a random seed.

The TRS procedure produced an integerised contingency table with 280 joint-combination columns for each census tract. This integer table preserves the total population of each census tract while minimizing distortion from the fractional IPF solution. Finally, the integerised table was aggregated back into the same grouped marginal structure as the original constraints, namely age–sex, education–sex, and employment–sex totals.

## 4.5 Synthetic Individuals and Households Generation

After the calculation of joint distributions through Iterative Proportional Fitting (IPF) and the subsequent integerisation step, the result is a contingency table that provides the number of individuals for each combination of control variables at the census-tract level (SEZID). Table 12 shows a sample of a few rows and columns of such a contingency table. According to the previous discussion, the full table presents 3545 rows and 281 columns. However, this table represents aggregate counts rather than individual records. Therefore, an additional step is required to generate a fully disaggregated synthetic population composed of individuals and households.

Table 12: Example of IPF output after integerisation.

| SEZID | Male        | Female     | Male         | Female       |
|-------|-------------|------------|--------------|--------------|
|       | 20–24       | 20–24      | 0–14         | 35–39        |
|       | High School | University | Other        | University   |
|       | Employed    | Employed   | Not Employed | Not Employed |
| 001   | 2           | 1          | 1            | 1            |
| 002   | 1           | 2          | 0            | 1            |
| 003   | 3           | 1          | 2            | 0            |
| 004   | 0           | 2          | 1            | 2            |

The generation process relies mainly on three data sources:

- IPF output after integerisation, which provides the target number of individuals for each census tract and for each combination of control variables.
- Census aggregate dataset, which provides household-size distributions at the census-tract level (attributes PF3–PF9) and the total population target.
- 1% individual census dataset, which serves as the donor of real individuals and households, preserving realistic demographic and socio-economic relationships.

#### 4.5.1 Synthetic Households Generation

Synthetic households are generated by sampling from the 1% census individual dataset. This step extends the individual-generation approach described in Section 3.6. First, the 1% census individual dataset is grouped into households using the household identifier (ID\_ALL). These households are then categorized according to their size.

Household-size distributions can be obtained at the census-tract level from the attributes shown in Table 13, which are available in the census aggregate dataset.

Table 13: Census attributes related to household size (source: census aggregate dataset).

| Variable | Definition                                                          |
|----------|---------------------------------------------------------------------|
| PF3      | Households with 1 member                                            |
| PF4      | Households with 2 members                                           |
| PF5      | Households with 3 members                                           |
| PF6      | Households with 4 members                                           |
| PF7      | Households with 5 members                                           |
| PF8      | Households with 6 or more members                                   |
| PF9      | Total number of persons living in households with 6 or more members |

For each required household size in a given census tract, all households with the same size are first selected. Each candidate household is then evaluated based on the attributes of its members (sex, age, educational attainment, and employment status) with the remaining target counts derived from the IPF procedure. The household that best

matches these target distributions is selected and replicated in the synthetic population. If several households satisfy the same criteria, one of them is selected randomly.

At the beginning of the allocation process, the remaining target counts correspond to the full IPF distributions. After each household is selected, these counts are updated by subtracting the attributes of the assigned individuals. Consequently, subsequent selections progressively match the remaining targets.

## 4.5.2 Synthetic Individuals Generation

Synthetic individuals are generated as members of the selected households. For each synthetic individual:

- A synthetic identifier is assigned.
- The census-tract identifier is assigned to each synthetic individual, as the IPF procedure is performed at the census-tract level.
- Additional attributes such as household roles and marital status are directly transferred from the sampled individual. This corresponds to the enrichment approach introduced in Section 3.5.

## 4.5.3 Collective Buildings

In the census dataset, the population is divided into two main groups: individuals living in private households and those residing in collective living places. Collective living arrangements include hospitals, student dormitories, prisons, retirement homes, and similar institutions. These places cannot be synthesised in the same way as private households since no marginal totals for them are available in the dataset. Including the collective population is essential to ensure consistency with the census data; therefore, collective individuals must be explicitly generated so that the synthetic population accurately reproduces the total population counts at the census-tract level.

The number of individuals living in collective households can be derived by subtracting the population living in private households from the total population of each census tract. In the census aggregate dataset, the variable P2 represents the population living in private households. Therefore, the difference between the total population and P2 provides an estimate of the individuals residing in collective living arrangements.

For each census tract where this difference is greater than zero, these individuals are assigned to a single collective household. This assumption is necessary because the aggregated census data do not provide information on the number of collective households within each census tract. Consequently, all such individuals are grouped into one collective household.

Synthetic individuals living in collective residences are sampled from the subset of records in the 1% census individual microdata corresponding to collective living arrangements. These records are identified using the `TIPCONV` attribute, which indicates the type of collective residence. Therefore, collective synthetic individuals are generated only from those records classified as belonging to collective residences.

#### 4.5.4 Final Output

At the end of the generation process, two complementary datasets are produced:

- A synthetic individual dataset, in which each record represents a synthetic individual.
- A synthetic household dataset, in which each record represents a synthetic household.

The attributes of these datasets are reported in Appendix A.5 and Appendix A.6, respectively.

### 4.6 Internal Validation Results

The internal validation evaluates the ability of synthetic population to reproduce the marginal distribution of the constraint variables at census tract level. This is simply done by comparing the simulated counts derived from population synthesis procedure with the corresponding observed census counts. The evaluation is primarily based on Pearson’s correlation coefficient ( $r$ ), which measures the strength of the linear relationship between observed and simulated values.

Overall, the result demonstrates a very strong level of agreement between the synthetic population and the census data. Pearson’s correlation coefficients are consistently high among all constraint variables categories confirming that the spatial distribution of individuals is successfully reproduced by the model.

For the age–sex constraints, the correlation values remain extremely high for both male and female populations. For males, Pearson’s  $r$  ranges between 0.985 and 0.999, as shown in Figure 4, while for females it ranges between 0.980 and 0.999, as shown in Figure 5. The lowest correlations are observed in the young adult age groups (20–24), with values of 0.985 for males and 0.980 for females, whereas most other age groups exceed 0.99.

Figure 4: Internal validation results for male-age categories (source: own elaboration)

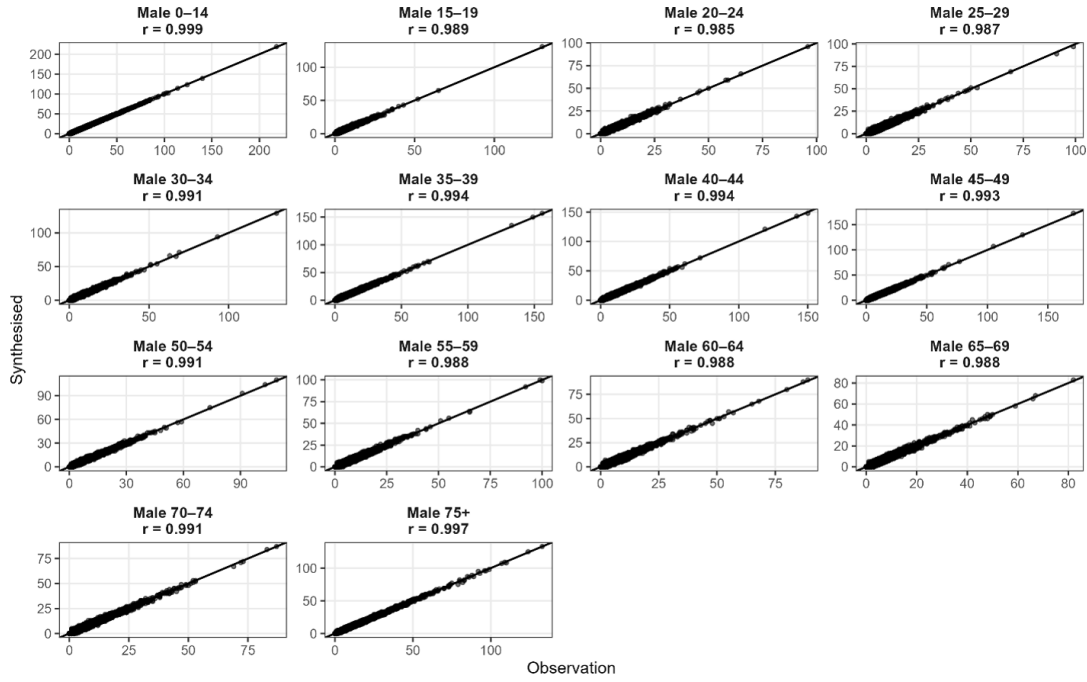
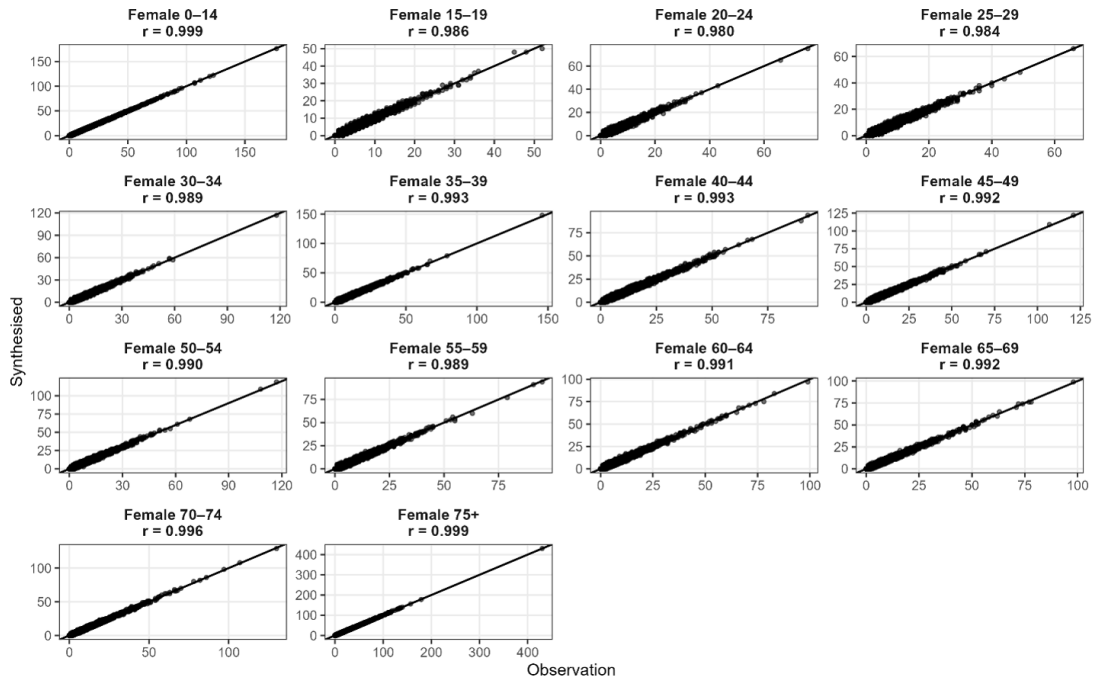


Figure 5: Internal validation results for female-age categories (source: own elaboration)



The sex-education constraints also exhibit a high level of consistency. For the male population, correlation values range between 0.995 and 0.999, as shown in Figure 6, while for females they range between 0.995 and 0.998, as shown in Figure 7. The lowest correlation is observed for the tertiary education category ( $r = 0.995$ ), while most other categories are close to 0.998–0.999.

Figure 6: Internal validation results for male-education categories (source: own elaboration)

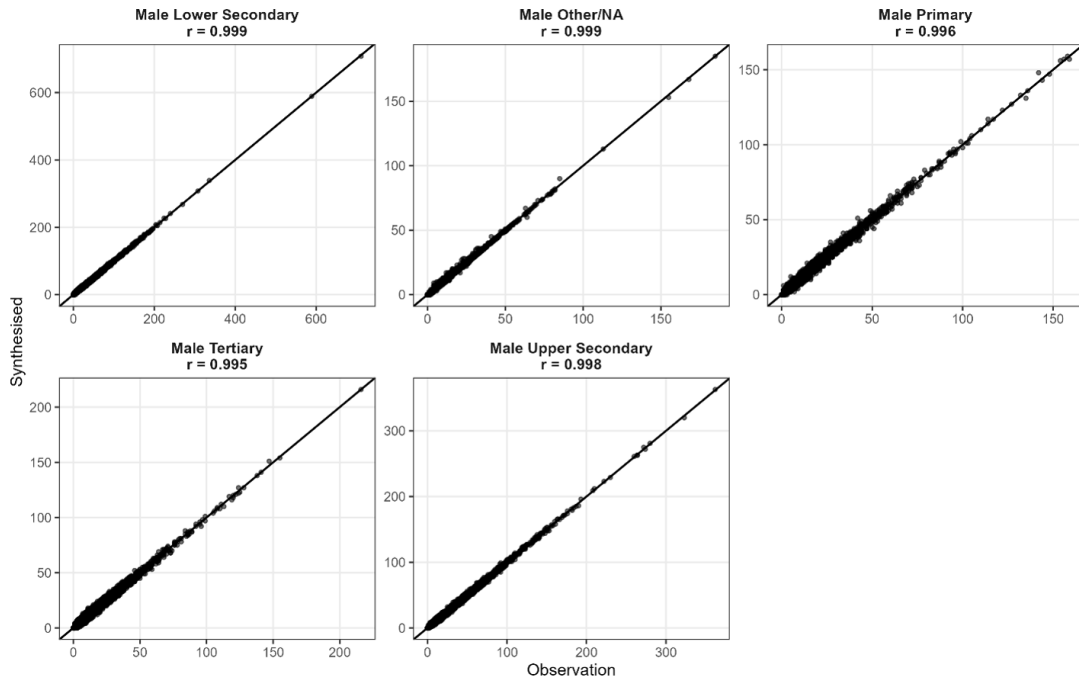
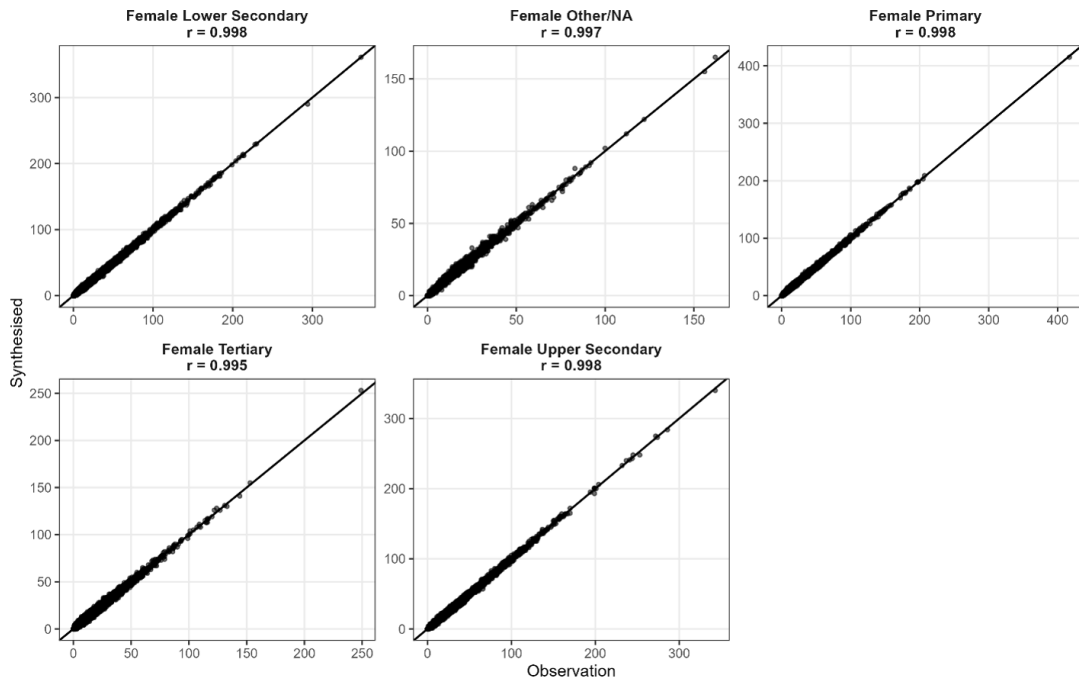
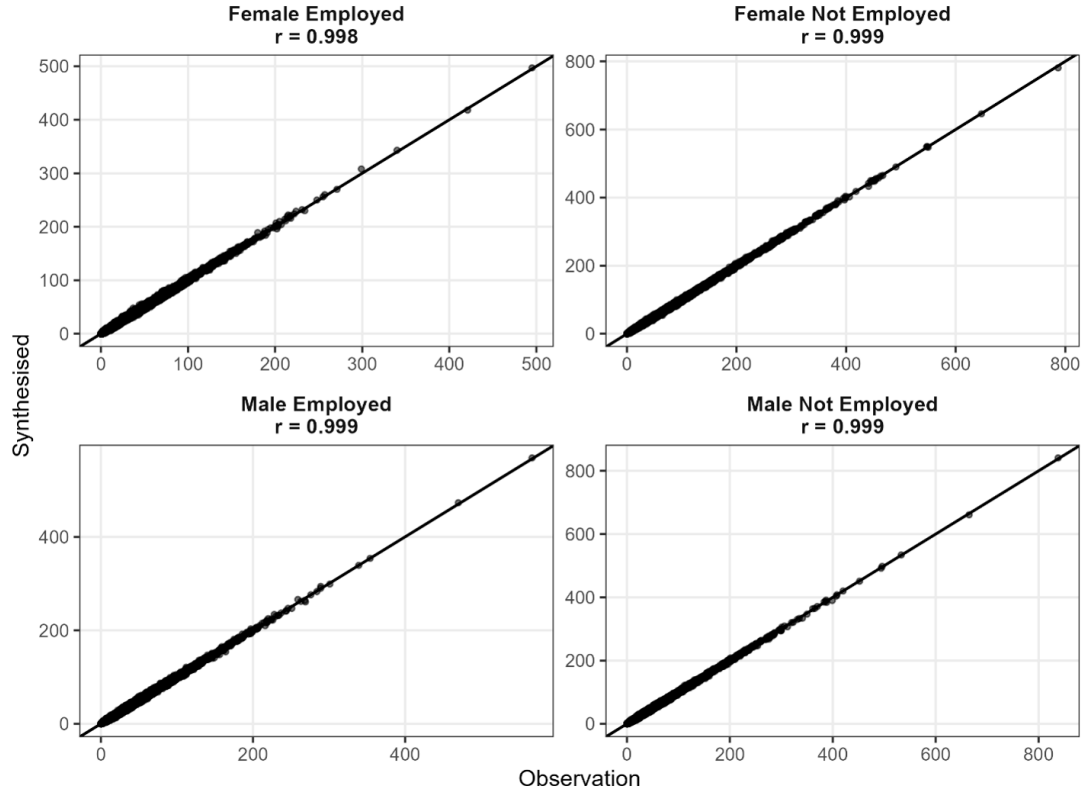


Figure 7: Internal validation results for female-education categories (source: own elaboration)



For the sex-employment status constraints, the agreement is nearly perfect. Pearson's correlation coefficients are equal to or greater than 0.998 for all categories, as shown in Figure 8, with most values reaching 0.999 for both males and females. This indicates an almost exact correspondence between observed and simulated counts.

Figure 8: Internal validation results for both male and female employment categories (source: own elaboration)



The internal validation was also repeated using the fractional IPF output before integerisation. In this case, Pearson’s correlation coefficient reached values equal or extremely close to 1 for almost all categories across the age–sex, education–sex, and employment–sex constraints. This confirms that the IPF procedure itself successfully fitted the census marginals with near-perfect accuracy.

However, after the integerisation step, small deviations appear in some categories. These deviations are expected because the fractional IPF estimates must be converted into integer counts before generating individual synthetic records. The effect is more visible in categories with relatively small counts, such as younger age groups. For example, Figure 9 and Figure 10 shows the validation result for females aged 15–19 before and after integerisation. As can be seen in Figure 10, after integerisation, a small dispersion around the line appears.

Figure 9: Internal validation results for female aged 15 to 19 before integerisation (source: own elaboration)

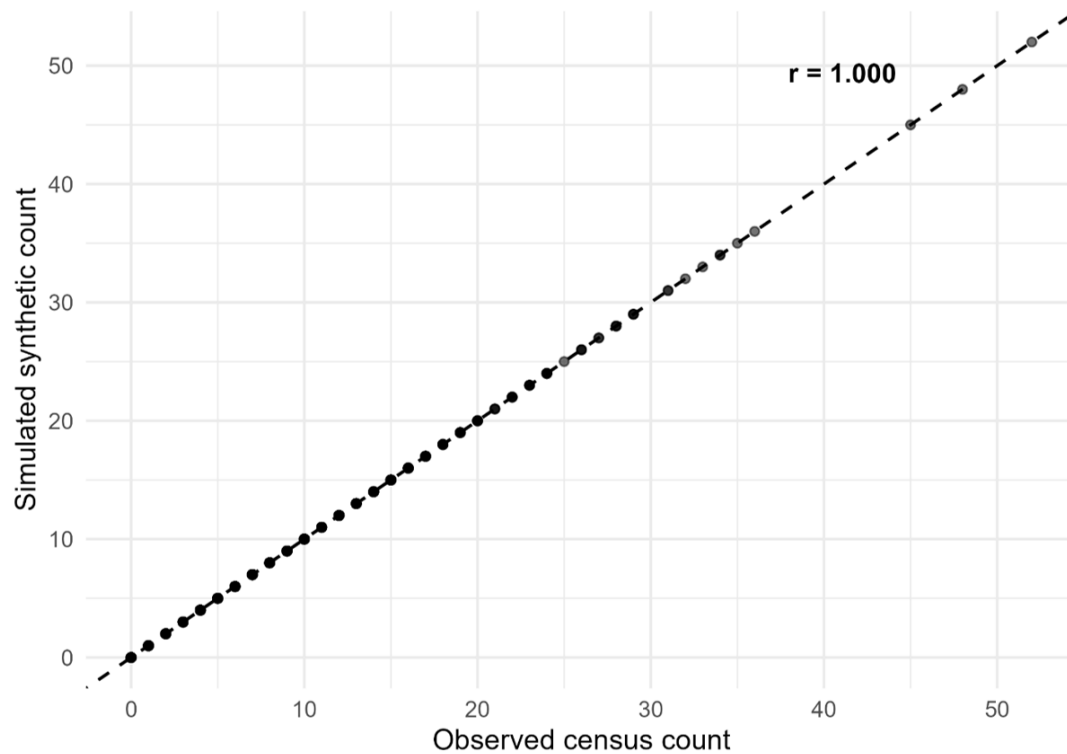
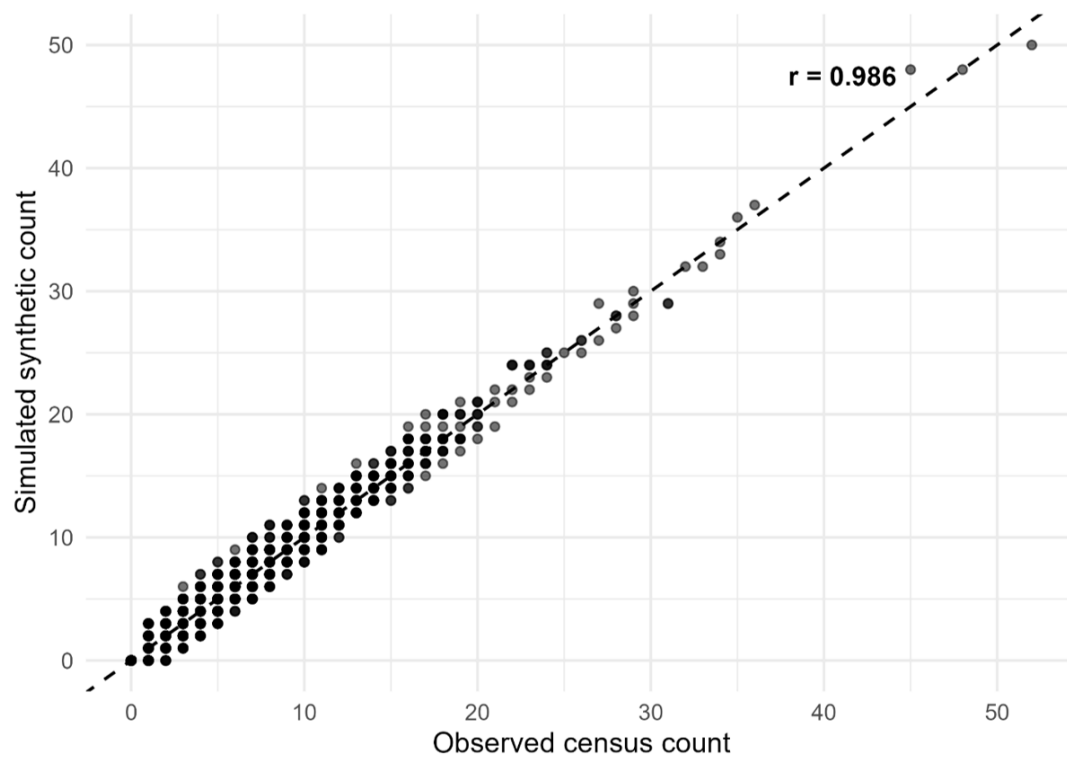


Figure 10: Internal validation results for female aged 15 to 19 after integerisation (source: own elaboration)



An additional analysis was conducted to examine the relationship between census tract population size and model error. The results indicate that more populated census tracts tend to exhibit lower relative error. On the other hand, populated census tracts tend to have lower absolute error, as shown in Figure 11, but higher relative error as this is shown in Figure 12.

Figure 11: Total absolute error for female age categories (source: own elaboration)

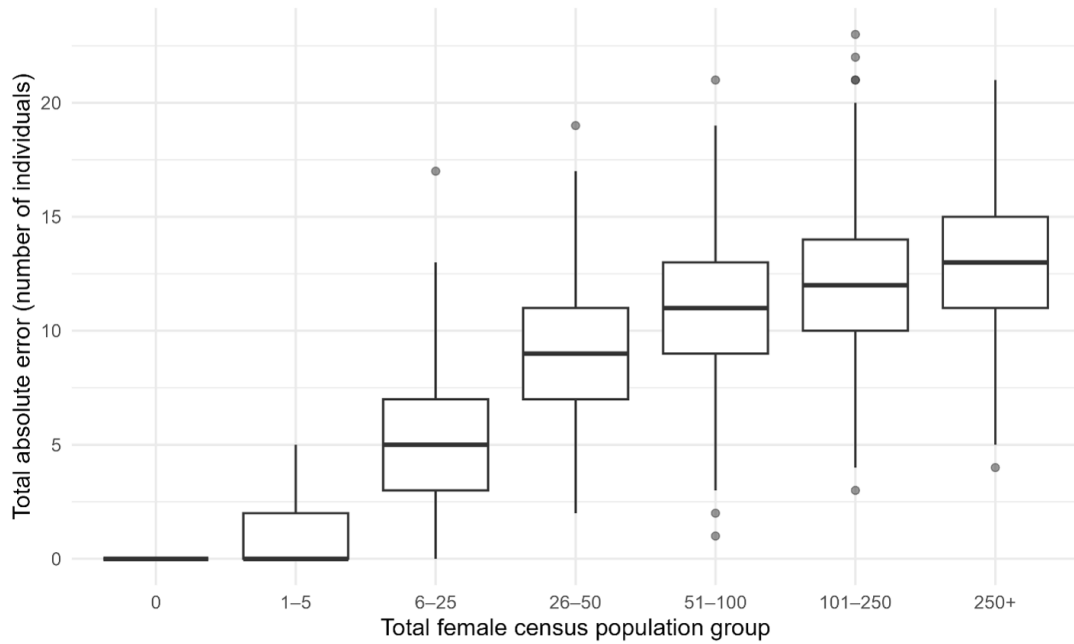
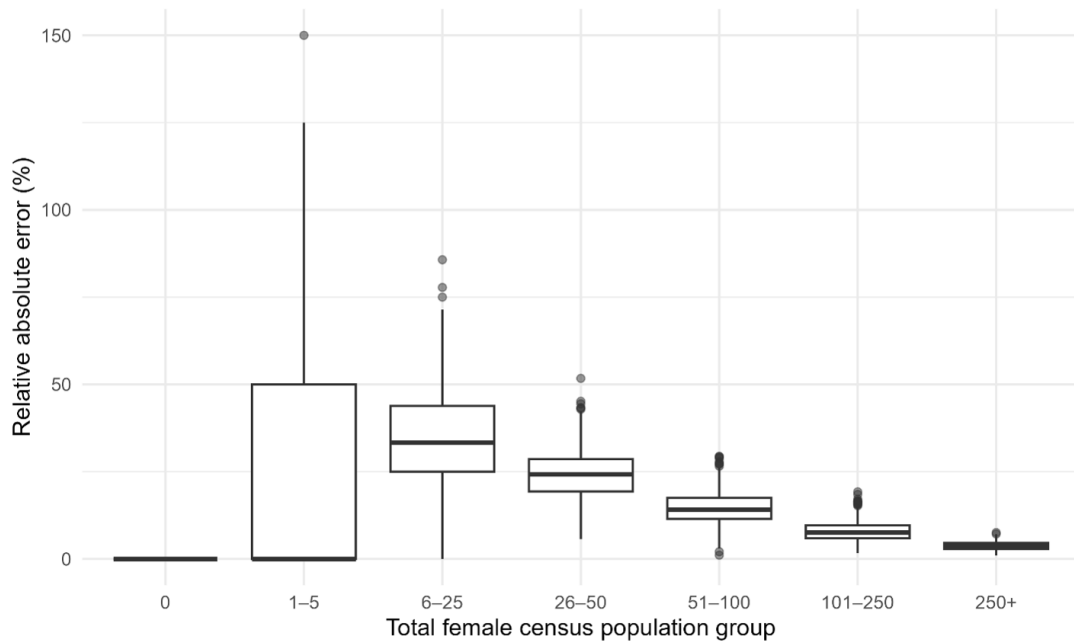


Figure 12: Relative absolute error for female age categories (source: own elaboration)



The points above the whiskers in Figure 12, represent census tracts with unusually high relative error. These cases are mainly associated with very small population counts,

where a small number of misallocated individuals across age categories can produce a large percentage error. For instance, the total female population in census tract 12720002356 is equal to 2: one individual belongs to the 35–39 age group and the other to the 50–54 age group. However, after IPF and integerisation, 3 females were assigned to this census tract in the 50–54 age group. This produces a relative error of approximately 150%.

## 4.7 External Validation Results

While internal validation verifies whether the synthetic population reproduces the census constraints used in the population synthesis procedure, it does not assess whether the resulting population reflects real-world distributions. For this reason, external validation was conducted using an attribute from the census aggregated dataset that was not included among the control variables.

Marital status, disaggregated by gender, was selected as a non-control variable and therefore was not used in the IPF procedure. The comparison between observed census counts and simulated values was performed at the census tract (SEZ2011) level across different marital status categories.

Overall, the results indicate a strong correspondence between the synthetic and observed distributions, particularly for the Married and Single categories. For females, as shown in Figure 13, the Pearson correlation reaches  $r = 0.990$  for Married and  $r = 0.975$  for Single, while for males, as shown in Figure 14, the correlations are  $r = 0.988$  and  $r = 0.987$ , respectively. Lower correlations are observed for Divorced and Legally separated individuals. For females, these are  $r = 0.804$  and  $r = 0.731$ , respectively, while for males they decrease further to  $r = 0.764$  and  $r = 0.700$ .

Figure 13: Marital status validation results for female (source: own elaboration).

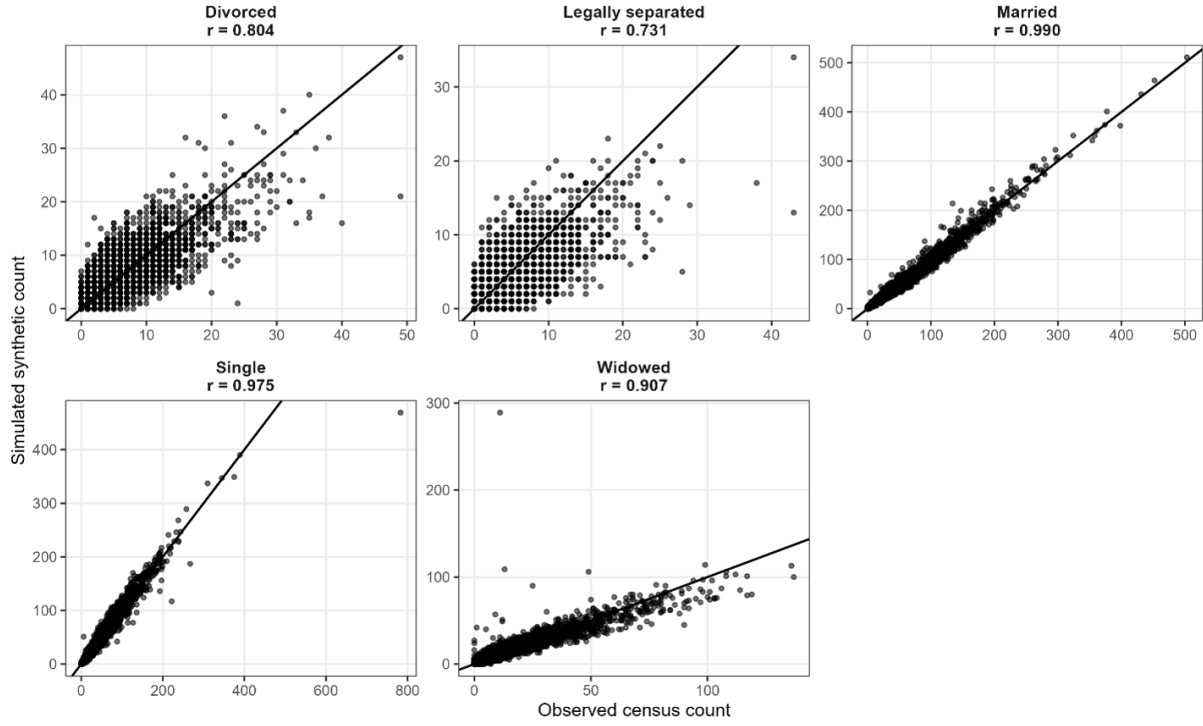
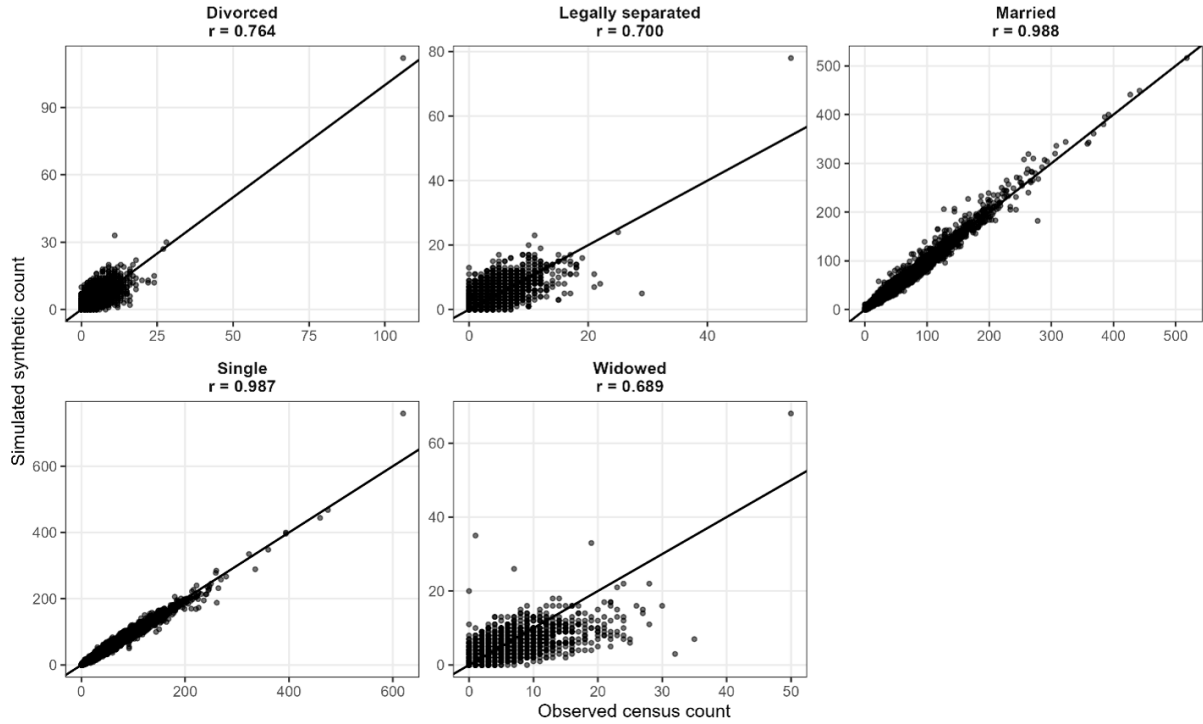


Figure 14: Marital status validation results for male (source: own elaboration).



In summary, the synthetic population reproduces the marital status distribution with higher accuracy for the largest categories (Married and Single), while maintaining ac-

ceptable but lower accuracy for smaller groups such as Divorced and Legally separated. For the female Widowed category, the scatterplot indicates a deviation from the diagonal reference line, suggesting that the model reproduces this group with lower accuracy than the other categories. This result is mainly due to the relatively small number of widowed individuals and their uneven spatial distribution across census tracts.

## 4.8 Descriptive Analysis of the Synthetic Population

To further illustrate the difference between the original census data and the final synthetic population, Table 14 presents an example of the information directly available from the census statistics for a randomly selected census tract (SEZ2011 = 12720001179). The census provides aggregate counts for individual variables but does not provide their joint distribution.

| Variable               | Category           | Count |
|------------------------|--------------------|-------|
| Sex                    | Male               | 36    |
|                        | Female             | 4     |
| Age group              | 0–14               | 0     |
|                        | 15–19              | 0     |
|                        | 20–24              | 1     |
|                        | 25–29              | 7     |
|                        | 30–34              | 7     |
|                        | 35–39              | 1     |
|                        | 40–44              | 2     |
|                        | 45–49              | 4     |
|                        | 50–54              | 0     |
|                        | 55–59              | 0     |
|                        | 60–64              | 1     |
|                        | 65–69              | 2     |
|                        | 70–74              | 6     |
|                        | 75+                | 9     |
| Employment status      | Employed           | 3     |
|                        | Not employed       | 37    |
| Educational attainment | Elementary school  | 7     |
|                        | Middle school      | 17    |
|                        | High school        | 7     |
|                        | Tertiary education | 9     |

Table 14: Example of census statistics for the selected census tract

As shown in Table 14, the census reports only separate marginal distributions. For example, it indicates that the selected census tract contains one resident aged 20–24 and nine residents with tertiary education. However, it does not specify whether the resident

aged 20–24 has tertiary education, whether either individual is employed, or what their marital status is.

Table 15 illustrates how the synthetic population combines these separate marginal distributions into individual-level records. It reports the joint distribution of selected individual attributes for the same census tract shown in Table 14. Only non-zero combinations are reported.

Table 15: Joint distribution of synthetic individuals in a randomly selected census tract

| SEZ2011     | Sex    | Age group | Education         | Employment   | Marital status     | Count |
|-------------|--------|-----------|-------------------|--------------|--------------------|-------|
| 12720001179 | male   | age_20_24 | edu_lowersec      | employed     | single             | 1     |
|             | male   | age_25_29 | edu_tertiary      | not_employed | single             | 2     |
|             | male   | age_25_29 | edu_uppersec      | not_employed | single             | 3     |
|             | male   | age_30_34 | edu_lowersec      | not_employed | single             | 4     |
|             | male   | age_30_34 | edu_uppersec      | not_employed | single             | 1     |
|             | male   | age_35_39 | edu_lowersec      | not_employed | single             | 1     |
|             | male   | age_40_44 | edu_lowersec      | not_employed | single             | 1     |
|             | male   | age_45_49 | edu_lowersec      | not_employed | single             | 3     |
|             | male   | age_45_49 | edu_uppersec      | not_employed | single             | 2     |
|             | male   | age_60_64 | edu_illit_primary | not_employed | single             | 1     |
|             | male   | age_60_64 | edu_lowersec      | not_employed | single             | 1     |
|             | male   | age_60_64 | edu_uppersec      | employed     | single             | 1     |
|             | male   | age_70_74 | edu_illit_primary | not_employed | single             | 1     |
|             | male   | age_70_74 | edu_illit_primary | not_employed | widowed            | 1     |
|             | male   | age_70_74 | edu_illit_primary | not_employed | missing_or_unknown | 1     |
|             | male   | age_70_74 | edu_tertiary      | not_employed | single             | 2     |
|             | male   | age_70_74 | edu_uppersec      | not_employed | single             | 1     |
|             | male   | age_75p   | edu_illit_primary | not_employed | single             | 1     |
|             | male   | age_75p   | edu_illit_primary | not_employed | married            | 1     |
|             | male   | age_75p   | edu_lowersec      | not_employed | single             | 1     |
|             | male   | age_75p   | edu_lowersec      | not_employed | widowed            | 2     |
|             | male   | age_75p   | edu_tertiary      | not_employed | single             | 3     |
|             | female | age_40_44 | edu_lowersec      | employed     | single             | 2     |
|             | female | age_45_49 | edu_tertiary      | employed     | single             | 1     |
|             | female | age_65_69 | edu_lowersec      | not_employed | single             | 2     |

The purpose of this table is to show the type of information that can be derived from the final synthetic population. Each row represents a specific combination of sex, age group, educational attainment, employment status, and marital status within a single census tract. This level of detail is not directly available from the original input datasets, either in census statistics or in census sample microdata.

Although the selected census tract exhibits some unbalanced characteristics, such as a predominance of male individuals and all female individuals being classified as single, these patterns should not be interpreted as a weakness of the synthesis procedure. The selected census tract contains only about 40 synthetic individuals, and therefore census tract demographic imbalances are expected due to the small population size. In addition, marital status was not used as a control variable during the IPF procedure but was assigned later during the enrichment stage. Therefore, its distribution may appear unusual in an individual census tract, even though the overall distribution remains realistic, as shown by the external validation results.

The same type of table can be produced for any other census tract in the study area. Therefore, the example in Table 15 should be interpreted as an illustration of the

analytical potential of the final synthetic dataset rather than as a result limited to the selected tract.

Furthermore, the table can be extended by adding other variables available in the final synthetic population, such as household size, household type, number of household vehicles, or other enriched attributes. The choice of the variables to include is flexible, and this flexibility is important because it shows that the synthetic population is not limited to reproducing census marginals but can also be used to explore more complex relationships between individual, household, and socio-economic characteristics at a fine spatial scale.

## 4.9 Activity-Chain Synthesis Using Bayesian Network

This section presents the implementation of the Bayesian Network framework described in Section 3.8. The framework was applied to estimate driving-license ownership and generate activity-chain sequences for the synthetic population of Torino using the dataset from a mobility survey.

### 4.9.1 Mobility Survey Preparation

Before training the Bayesian Network model, the available mobility survey dataset was cleaned and adjusted to ensure consistency with the synthetic population variable categories. The mobility survey considered in the present work was conducted through three different waves between September 2016 and June 2017 in the whole Turin metropolitan area, as part of a research project carried out by Politecnico di Torino. In total, 4,466 interviews were collected across the three survey waves. Details about this survey can be found in (Ceccato 2020).

In this study, all three waves were combined and considered as one survey in order to maximize the number of observations for training the model. As discussed in Section 2.3, previous studies implementing Bayesian Networks for activity-chain synthesis have relied on datasets containing several thousand observations. Using only a single survey wave would have substantially reduced the sample size and limited the diversity of activity-chain patterns in the study area.

The first step for cleaning the mobility survey is to remove duplicated records across multiple waves, since the same individual could have been interviewed across multiple waves. Duplicated records are linked by identifiers (`ID_INTERV` and `ID_INT2`), which were used to remove duplicated records and retain only the earliest valid interview associated with each individual.

After removing duplicated records, observations with missing or invalid information in essential socio-demographic and household variables were removed as well. Missing

values included blank cells and survey-specific missing codes such as  $-1$  (legitimate skip),  $-6$  (no answer),  $-7$  (do not know),  $-8$  (interview interrupted), and other missing-value conditions. The variables considered mandatory for retention included age (ETA), gender (SESSO), employment status (OCCUPAZ), education level (STUDIO), driving-license ownership (PATENTE), household size (FAM\_N), and household car ownership (FAM\_AUTO).

After cleaning, 4,183 valid observations remained from the original 4,466. A total of 283 observations were removed during preprocessing, including 261 duplicated records across survey waves and 22 observations containing missing information in key socio-demographic or household variables. Among the observations removed due to missing information, four records lacked all major socio-demographic variables, seventeen records lacked education and household-related variables, and one record lacked employment information in addition to household and education variables.

After cleaning, the survey variables were recoded and relabeled to ensure consistency with the synthetic population as described in Section 3.8. Table 16 summarizes all the recoding and relabeling variables between the mobility survey and the synthetic population.

Table 16: Aligning socio-demographic variables between the mobility survey and the synthetic population dataset.

| Original Variable | Original Categories                                                                                                                                                                                                                                                | New Label         | New Categories                                                                                                                                                                   |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SESSO             | M, F                                                                                                                                                                                                                                                               | Sex               | Male, Female                                                                                                                                                                     |
| ETA               | Metric variable in years                                                                                                                                                                                                                                           | age_group         | age_0_14, age_15_19, age_20_24, age_25_29, age_30_34, age_35_39, age_40_44, age_45_49, age_50_54, age_55_59, age_60_64, age_65_69, age_70_74, age_75p                            |
| OCCUPAZ           | 1: Entrepreneur/freelancer<br>2: Officer/manager<br>3: Employee/trade employee<br>4: Worker<br>5: Teacher<br>6: Salesperson<br>7: Artisan/retailer<br>8: Student<br>9: Housewife<br>10: Retired<br>11: Waiting for first employment<br>12: Unemployed<br>13: Other | employment_binary | Employed<br>Employed<br>Employed<br>Employed<br>Employed<br>Employed<br>Employed<br>not_employed<br>not_employed<br>not_employed<br>not_employed<br>not_employed<br>not_employed |
| CASSA.INT         | 1: Cassa integrazione/mobilità<br>2: I work                                                                                                                                                                                                                        | employment_binary | not_employed<br>employed                                                                                                                                                         |
| STUDIO            | 1: No studies<br>2: Primary school<br>3: Middle school<br>4: High school<br>5: University degree                                                                                                                                                                   | education_group   | edu_illit_primary<br>edu_illit_primary<br>edu_lowersec<br>edu_uppersec<br>edu_tertiary                                                                                           |

### 4.9.2 Bayesian Network Implementation for Driving License

The first Bayesian Network model was implemented to generate driving-license ownership for synthetic individuals. The model was trained based on the socio-demographic variables from the mobility survey. The predictor variables include sex, age group, employment status, and household car availability, while driving-license ownership was used as the target variable.

To evaluate the model performance, 80% of records (3346 observations) from the mobility survey were randomly selected to train the model and 20% of the records (837 observations) were used for testing the model performance.

The Bayesian Network achieved an overall accuracy of 88.29% on the testing dataset for predicting driving-license ownership based on socio-demographic and household characteristics. The results demonstrate that the model was able to successfully capture the probabilistic relationship between demographic variables and the ownership of a driving license. The implementation and estimation of the Bayesian Network for estimating driving-license ownership were carried out using the Python script developed for this study and provided in Appendix A.8.

### 4.9.3 Bayesian Network Implementation for Activity-Chain Generation

After the estimation of driving-license ownership, a Bayesian Network model was used again to generate daily activity chains for synthetic individuals. As explained in Section 3.8, a Bayesian Network model is implemented to assign realistic sequences of daily activities while preserving the relationships observed between socio-demographic characteristics and travel behavior. The activity-chain generation process was based on the mobility survey dataset and implemented using the Python script shown in Appendix A.9.

As discussed in Section 3.8, the activity episodes contained in the mobility survey did not always begin at the start of the day. In some cases, the sequence started from an activity occurring later in the daily schedule. Therefore, a preprocessing procedure was implemented to reconstruct complete home-based activity chains. Activity episodes were first ordered chronologically and then reorganized so that the sequence began from the first observed home activity whenever possible. For example, a sequence recorded as Leisure (18:00–19:45) – Home (20:00–08:30) – Work (09:00–17:30) was transformed into Home (00:00–09:00) – Work (09:00–17:30) – Leisure (18:00–19:45) – Home (20:00–24:00). This procedure preserved the observed activities and their chronological order while providing a consistent home-based representation of the daily schedule. The reconstructed activity chains provide a representation of daily travel behavior and are suitable for training the Bayesian Network model.

After reconstructing the activity sequences, the original survey activity types were grouped into a smaller number of activity categories, as shown in Table 17, to reduce model complexity and ensure sufficient observations within each category. Activities with similar purposes were aggregated into common activity groups while preserving the main characteristics of daily travel behavior.

Table 17: Aggregation of original activity types into activity categories used for the Bayesian Network.

| Original Survey Activity           | Aggregated Activity Category |
|------------------------------------|------------------------------|
| Staying at home                    | Home                         |
| Working (usual workplace)          | Work                         |
| Business travel                    | Work                         |
| Working to carry loads/passengers) | Work                         |
| Studying at school/university      | Education                    |
| Medical consultation/treatment     | Maintenance                  |
| Eating and drinking                | Maintenance                  |
| Grocery shopping / shopping center | Shopping                     |
| Taking away or picking up people   | Escort                       |
| Visiting friends or relatives      | Leisure                      |
| Other recreational activities      | Leisure                      |
| Taking a stroll                    | Leisure                      |
| Travelling                         | Leisure                      |
| Other activities                   | Leisure                      |

After activity-chain reconstruction and consistency checks, 133 observations were removed from the dataset, leaving 4050 valid observations. All removed records contained at least one activity episode with a non-positive duration, meaning that the reconstructed end time was equal to or earlier than the corresponding start time. Most records were removed because of intermediate activity episodes, particularly ATTIV\_4 (31 records), ATTIV\_3 (29 records), ATTIV\_2 (23 records), and ATTIV\_5 (19 records). Table 18 summarizes the number of removed observations by activity episode.

Table 18: Removed records from the mobility survey based on non-positive activity duration.

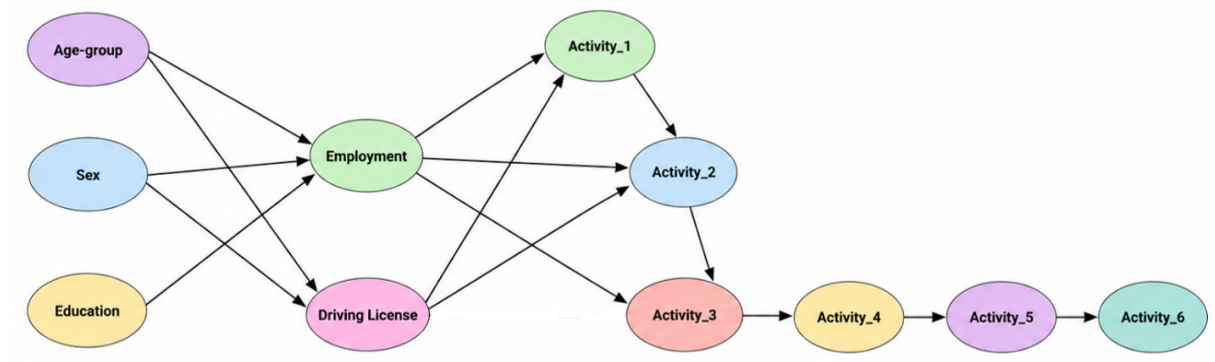
| Removal reason                             | Count |
|--------------------------------------------|-------|
| Non-positive activity duration at ATTIV_1  | 1     |
| Non-positive activity duration at ATTIV_2  | 23    |
| Non-positive activity duration at ATTIV_3  | 29    |
| Non-positive activity duration at ATTIV_4  | 31    |
| Non-positive activity duration at ATTIV_5  | 19    |
| Non-positive activity duration at ATTIV_6  | 9     |
| Non-positive activity duration at ATTIV_7  | 8     |
| Non-positive activity duration at ATTIV_8  | 1     |
| Non-positive activity duration at ATTIV_9  | 2     |
| Non-positive activity duration at ATTIV_11 | 3     |
| Non-positive activity duration at ATTIV_13 | 1     |
| Non-positive activity duration at ATTIV_14 | 1     |
| Non-positive activity duration at ATTIV_15 | 5     |

As described in Section 3.8, to facilitate Bayesian Network estimation, activity chains were represented using a fixed-length structure consisting of six activity positions. After activity-chain reconstruction and cleaning, 4,050 valid individuals remained in the activity-chain dataset. Among these, 49 individuals (1.21%) had activity chains containing more than six activities and were therefore excluded from the Bayesian Network training dataset. Consequently, the final dataset used for Bayesian Network estimation contained 4,001 individuals. The small number of removed observations due to having activity-chain longer than six indicates that the six-activity representation was sufficient to capture the daily activity behavior of the vast majority of survey respondents.

The final Bayesian Network training dataset consisted of socio-demographic variables and six activity-chain variables representing the sequential activities performed in a day. The explanatory variables included age group, sex, education level, employment status, household size, number of household cars, and driving-license ownership. The activity-chain variables were represented by Activity\_1 to Activity\_6, corresponding to the ordered sequence of activities performed by each individual.

The learning algorithm automatically explored different network configurations and identified the dependency relationships that were consistent with the observed activity patterns. The resulting structure, shown in Figure 15, indicates how employment status and driving-license ownership influence the initial activities in the daily schedule, while subsequent activities are primarily influenced by preceding activities in the sequence.

Figure 15: Learned Bayesian Network structure (source: own elaboration on mobility dataset).



To have meaningful relationships, several structural constraints were introduced during the learning process. Socio-demographic variables were specified as root variables and therefore were not allowed to have parent nodes. Since the order of activity should follow the time, later activities prevent from influencing earlier activities. For example, Activity\_5 could not be connected as a parent of Activity\_3. These constraints ensured that the learned network remained consistent with the chronological nature of daily activity behavior.

After learning the structure, the Bayesian Network estimated was trained using mobility survey data. During this process, the model learns how activity can be affected by different socio-demographic variables and prior activity. For instance, the model learned that employed individuals are more likely to have work-related activity than non-employed individuals or the activity performed at one stage of the day influences the activities that are likely to occur afterwards. Based on these learned relationships, the Bayesian Network was able to generate activity-chain for the synthetic population that has been generated in Subsection 4.5.2, according to the observed patterns in the mobility survey.

For each synthetic individual, the model used the socio-demographic and mobility related variables such as employment status and driving license ownership to generate daily activity chain up to six activities. The final output of this process was a synthetic population dataset enriched with six activity variables (Activity\_1 to Activity\_6), representing the daily activity sequence assigned to each synthetic individual.

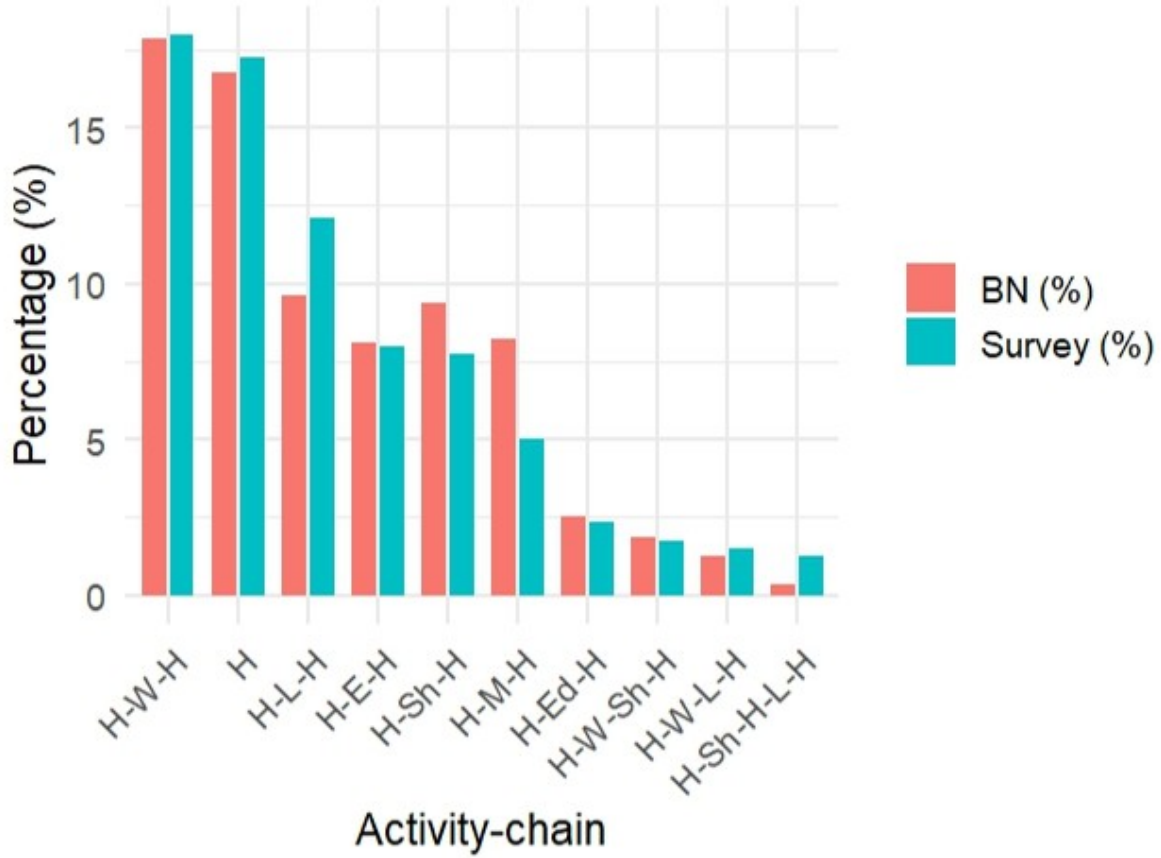
The generated activity chains provide each synthetic individual with a plausible sequence of daily activities, thereby enriching the synthetic population with behavioral information. However, the output remains limited to activity sequences. The spatial locations of activities, trip destinations, travel modes, and travel times were not generated in this study. Consequently, the synthesized activity chains should be regarded as an intermediate output that can be used as input for subsequent activity-location and travel-demand modelling.

#### 4.9.4 Activity-Chain Validation

To evaluate the ability of the Bayesian Network to reproduce observed activity-chain patterns, the dataset was randomly divided into training and testing subsets. Following the same procedure adopted for the driving-license model, 80% of the observations were used for model estimation, while the remaining 20% were reserved for validation.

The activity chains generated by the Bayesian Network were compared with the observed activity chains extracted from the mobility survey dataset. The comparison was performed using the 20% testing dataset that was not used during model estimation. For both datasets, the relative frequency of each activity-chain pattern was calculated. The ten most frequent activity chains were then identified and compared. Figure 16 presents the comparison between the Bayesian Network output and the observed activity chains in the mobility survey.

Figure 16: Activity-chain frequency comparison between observed and Bayesian Network output (source: own elaboration on the Torino mobility survey).



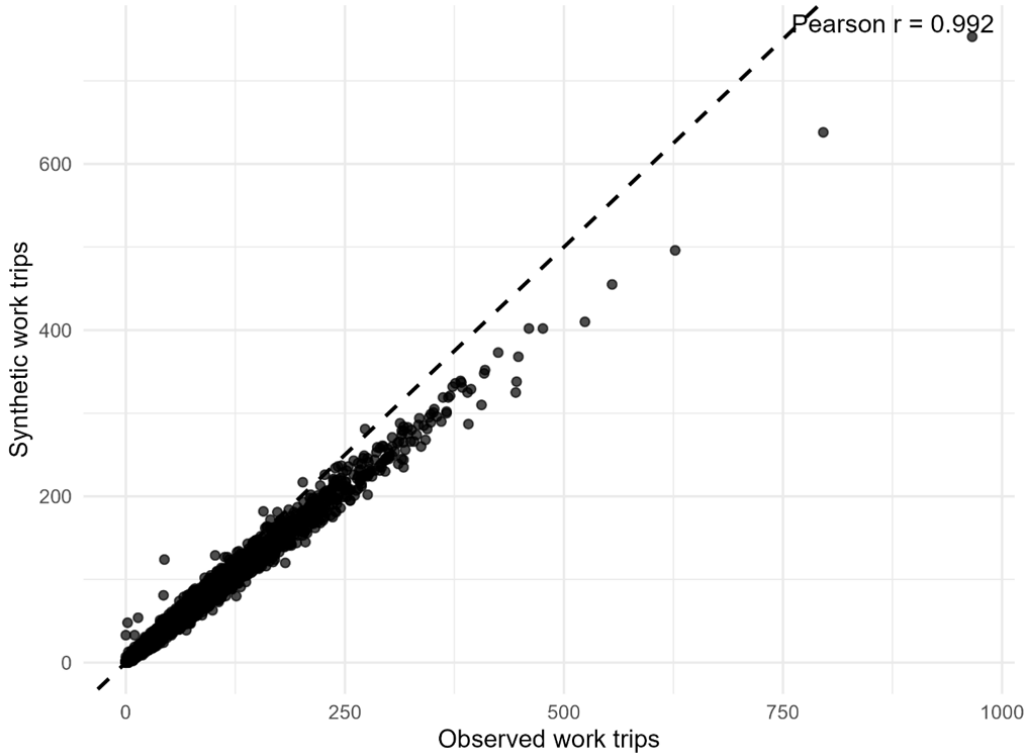
The results show that the Bayesian Network was able to reproduce the dominant activity-chain patterns observed in the mobility survey. In particular, the most common activity chains appeared with similar frequencies in both datasets, indicating that the model was able to capture the overall structure of daily activity behavior. Small differences were observed for less frequent activity chains, which is expected due to the greater behavioral variability associated with these patterns and their limited representation in the survey data.

As an additional validation, the spatial distribution of generated work-related activities was compared with independent census statistics. Since the synthesized population is available at the census-tract level, it is possible to compare the number of work-trip origins generated from each census tract with work-trip production statistics derived from the census commuting matrix.

Figure 17 presents a comparison between the observed number of work-trip productions obtained from the census commuting matrix and the number of synthesized work-trip origins generated from each census tract. The results show a strong agreement between observed and synthesized values, with a Pearson correlation coefficient of 0.992. Most census tracts are located close to the diagonal reference line, indicating that the activity-

chain assignment procedure successfully preserves the spatial distribution of work-related activities across the study area.

Figure 17: Comparison between observed and synthesized work-trip productions at the census-tract level (source: own elaboration).



Overall, the validation results demonstrate that the Bayesian Network successfully reproduces both the behavioral patterns observed in the mobility survey and the spatial distribution of work trip participation observed in the census data. These findings provide confidence that the generated activity chains constitute a realistic behavioral enrichment of the synthetic population and are suitable for subsequent activity-based and agent-based transport modelling applications.

#### 4.9.5 Descriptive Analysis of the Synthesized Activity Chains

To examine the characteristics of the generated activity chains, the synthetic population was stratified according to three household attributes: the presence of children, the presence of at least one employed adult, and household size. Although activity chains were not generated for children, they were considered in the household classification because their presence may influence the activity patterns of other household members, particularly through education- and escort-related trips. For each household group, the distribution of activity purposes were analyzed. The complete ranking of the twenty most frequent activity chains for each household group is reported in Appendix A.10, while Table 19 summarizes the distribution of activity purposes.

Table 19: Activity-purpose distribution by household group.

| Children | Employment         | Household Size | Work (%) | Education (%) | Escort (%) | Shopping (%) | Maintenance (%) | Leisure (%) |
|----------|--------------------|----------------|----------|---------------|------------|--------------|-----------------|-------------|
| With     | ≥1 employed adult  | 2              | 68.09    | 0.49          | 10.61      | 15.48        | 12.64           | 21.03       |
|          | ≥1 employed adult  | 3–4            | 55.93    | 1.93          | 13.36      | 17.12        | 13.42           | 22.60       |
|          | ≥1 employed adult  | 5+             | 42.12    | 2.94          | 15.40      | 19.55        | 13.94           | 23.72       |
|          | No employed adults | 2              | 4.26     | 7.18          | 23.32      | 27.19        | 13.98           | 28.08       |
|          | No employed adults | 3–4            | 4.33     | 6.62          | 23.84      | 26.87        | 15.38           | 28.99       |
|          | No employed adults | 5+             | 3.84     | 6.05          | 23.23      | 27.53        | 15.38           | 28.72       |
| Without  | ≥1 employed adult  | 1              | 69.23    | 0.62          | 11.18      | 14.53        | 12.89           | 20.77       |
|          | ≥1 employed adult  | 2              | 50.92    | 2.11          | 13.72      | 17.38        | 13.54           | 22.68       |
|          | ≥1 employed adult  | 3–4            | 37.73    | 3.69          | 16.67      | 20.20        | 14.09           | 24.55       |
|          | ≥1 employed adult  | 5+             | 38.86    | 2.97          | 15.74      | 18.77        | 13.91           | 24.06       |
|          | No employed adults | 1              | 4.09     | 6.75          | 22.61      | 25.98        | 15.33           | 28.12       |
|          | No employed adults | 2              | 3.95     | 6.11          | 21.46      | 25.98        | 14.99           | 27.59       |
|          | No employed adults | 3–4            | 4.26     | 6.84          | 22.36      | 26.29        | 14.75           | 27.92       |
|          | No employed adults | 5+             | 2.35     | 1.57          | 18.43      | 22.75        | 14.90           | 27.06       |

Table 19 reports, for each household group, the percentage of synthetic individuals with a valid generated activity chain whose chain includes at least one activity of the each purpose. For example, the value reported under Work (%) indicates the percentage of individuals whose activity chain contains at least one work activity. Since a single activity chain may include several activity purposes, the percentages do not sum to 100%.

To further demonstrate the effectiveness of the synthesized activity chains, Table 20 reports the observed activity chains for one household group in the original mobility survey. The selected household group (households with children, no employed adults, and three to four members) contains only six respondents, resulting in only four observed activity-chain patterns. The values reported under Share (%) indicate the percentage of respondents within this household group whose observed daily activity chain corresponds to the listed pattern. In comparison, Appendix A.10 shows that, after assigning activity chains to the complete synthetic population, a much richer distribution of activity-chain patterns becomes available for the same household group. This demonstrates how the synthesis extends the behavioral information contained in the survey while preserving the dominant activity patterns observed in the original data.

Table 20: Observed activity chains in the mobility survey for a selected household group.

| Rank | Activity Chain     | Share (%) |
|------|--------------------|-----------|
| 1    | Home               | 50.0      |
| 2    | Home–Leisure–Home  | 16.7      |
| 3    | Home–Shopping–Home | 16.7      |
| 4    | Home–Work–Home     | 16.6      |

The analysis shows that employment status is the primary factor influencing the generated daily activity patterns. As reported in Appendix A.10, for all household groups containing at least one employed adult, the most frequent activity chain is Home–Work–Home. However, its dominance gradually decreases as household size increases. For example, among one-person households without children, the Home–Work–Home chain

accounts for approximately 40% of all generated activity chains, whereas this proportion decreases to about 21% among households with three or four members. This indicates that larger households exhibit more diverse daily schedules rather than being dominated by simple activity chain pattern.

In contrast, Appendix A.10 shows that households without employed adults display substantially different activity patterns. In these households, the dominant activity chain becomes Home, followed by Home–Shopping–Home, Home–Leisure–Home, and Home–Escort–Home. These results indicate that the proposed methodology successfully generates activity chains that reflect the employment characteristics of the synthetic population.

The presence of dependent children has a more subtle influence on activity-chain structure. Although Home–Work–Home remains the dominant chain in employed households regardless of household composition, Appendix A.10 indicates that households with children exhibit a greater number of escort- and education-related activity chains, such as Home–Escort–Home and Home–Work–Escort–Home. Table 19 complements these results by summarizing the distribution of activity purposes across the different household groups. As household size increases escort, shopping, and maintenance activities become relatively more frequent, suggesting that larger households are characterized by more complex daily activity schedules involving multiple trip purposes.

Table 19 further confirms these differences. Among households containing at least one employed adult, the proportion of work-related activity chains decreases with increasing household size, while the shares of escort and shopping activities gradually increase. Conversely, households without employed adults are characterized by substantially lower proportions of work-related activities and higher proportions of shopping, leisure, and escort activities. In comparison, leisure activities remain relatively stable across the different household groups, indicating that household composition mainly affects mandatory activities rather than discretionary travel.

Overall, these results demonstrate that the generated activity chains are not assigned uniformly across the synthetic population. Instead, the proposed methodology produces heterogeneous daily schedules that vary according to household composition, employment status, and household size, thereby preserving important behavioral differences that are essential for subsequent activity-based transport simulations.

# Chapter 5

## Conclusions

This thesis presented a framework for generating a synthetic population at the census-tract level for Torino city in Italy. The implemented methodology used census statistics with census microdata sample. The proposed method applied Iterative Proportional Fitting (IPF) to produce a spatially disaggregated synthetic population that reproduces the demographic and socio-economic characteristics observed in official census statistics while preserving the multivariate relationships contained in the microdata. The resulting synthetic population provides individual and household records with detailed socio-demographic attributes at fine spatial level, making it suitable for transport modelling and other microsimulation applications where detailed individual-level data are unavailable because of privacy restrictions or data availability limitations.

The quality of the generated population was assessed through both internal and external validation. Internal validation demonstrated that the synthetic population accurately reproduces the census marginal distributions used as constraints at the census-tract level, confirming the correct implementation of the IPF procedures. External validation, based on variables that were not directly constrained during the synthesis process, further showed that the synthetic population is able to reproduce additional socio-demographic characteristics with satisfactory accuracy. The agreement observed in the validation analyses indicates that the methodology preserves realistic relationships between variables while generating a population that is statistically consistent with the available reference data. The descriptive analyses presented in the final chapter also illustrate that the resulting dataset contains detailed multivariate information at the individual and household level that is not directly available from the original input datasets at the census-tract level.

In addition to the generation of the synthetic population, this research assigned daily activity chains to the synthetic individuals using a Bayesian Network trained on mobility survey data. The Bayesian Network captured the probabilistic relationships between socio-demographic characteristics and daily activity patterns, allowing activity chains to be assigned to synthetic individuals. As a result, the final output extends beyond a static synthetic population by adding information on daily activity patterns. However, since the

activity locations were not assigned, this output should be considered an intermediate step toward a complete synthetic travel demand dataset for activity-based transport modelling.

Despite these encouraging results, several limitations should be acknowledged. First, the quality of the synthetic population depends on the representativeness of the available microdata. If some socio-demographic combinations are missing or underrepresented in the survey sample, they cannot be fully represented in the synthetic population. Second, the synthesis was constrained by the limited number of variables available from the census at the census-tract level. As a result, the additional attributes assigned during the enrichment phase were not directly validated using small-area census data. Second, the synthesis was constrained by the limited number of variables available from the census at the census-tract level. Although some additional attributes assigned during the enrichment phase were successfully evaluated through external validation, not all enriched variables could be validated because suitable reference data at the required level of detail were not available.

The activity chain generation also presents opportunities for further development. Although the Bayesian Network successfully assigned plausible daily activity sequences based on the socio-demographic characteristics of the synthetic individuals, the generated chains do not include information on the spatial locations where activities are performed or on the trips connecting consecutive activities. Future research should therefore focus on assigning activity locations and modelling the trips between consecutive activities, including destination, travel mode, and travel duration. This would transform the current framework into a complete synthetic travel demand model suitable for activity-based transport simulation.

Overall, this research demonstrates that a combination of publicly available census data, mobility survey, and established spatial microsimulation techniques can successfully generate a realistic synthetic population at a fine spatial resolution. The proposed framework is fully reproducible, transparent, and adaptable to other study areas where similar data are available. By providing a detailed representation of the resident population while preserving consistency with official statistics, the proposed framework offers a practical foundation for future activity-based transport modelling and other spatial microsimulation applications.

# References

- Aemmer, Z. and D. MacKenzie (2022). “Generative population synthesis for joint household and individual characteristics”. In: *Computers, Environment and Urban Systems*, p. 101852. DOI: 10.1016/j.compenvurbsys.2022.101852.
- Ankan, Ankur and Abinash Panda (2015). “pgmpy: Probabilistic graphical models using python”. In: SciPy Conference. URL: <https://pgmpy.org> (Accessed May 18th, 2025).
- Balać, Milos, Aurore Sallard, and Elsevier B.V. (2023). “Travel demand generation using Bayesian Networks: an application to Switzerland”. In: *Procedia Computer Science* 220. Ed. by Elsevier B.V., pp. 267–274. DOI: 10.1016/j.procs.2023.03.035.
- Ballas, Clarke, and Wiemers (2005). “Building a dynamic spatial microsimulation model for Ireland”. In: *Population, Space and Place*, pp. 157–172. DOI: 10.1002/psp.359.
- Barthelemy, J. and P. L. Toint (2013). “Synthetic Population Generation Without a Sample”. In: *Transportation Science*, pp. 266–279. DOI: 10.1287/trsc.1120.0408.
- Beckman, Baggerly, and Mckay (1996). “Creating synthetic baseline populations”. In: *Transportation Research Part A: Policy and Practice*, pp. 415–429. DOI: 10.1016/0965-8564(96)00004-3.
- Bishop, Y. M. M. et al. (1977). “Book Review: Discrete Multivariate Analysis: Theory and Practice”. In: *Applied Psychological Measurement*, pp. 297–306. DOI: 10.1177/014662167700100218.
- Ceccato, Riccardo (2020). “Switching intentions towards car sharing: Analysis of the relationship with traditional transport modes”. PhD thesis. Torino: Politecnico di Torino. URL: <http://hdl.handle.net/11583/2840371> (Accessed May 25th, 2026).
- Choupani, A.-A. and A. R. Mamdoohi (2016). “Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research”. In: *Transportation Research Procedia*, pp. 223–233. DOI: 10.1016/j.trpro.2016.11.078.

- Deming, W. E. and F. F. Stephan (1940). “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known”. In: *The Annals of Mathematical Statistics*, pp. 427–444. DOI: 10.1214/aoms/1177731829.
- Farooq, Bilal et al. (2013). “Simulation based population synthesis”. In: *Transportation Research Part B* 58, pp. 243–263. DOI: 10.1016/j.trb.2013.09.012.
- Fienberg, S. E. and M. M. Meyer (1972). *Iterative Proportional Fitting*. URL: <https://apps.dtic.mil/sti/html/tr/ADA102618/> (Accessed April 29th, 2026).
- Guo, J. Y. and C. R. Bhat (2007). “Population Synthesis for Microsimulating Travel Behavior”. In: *Transportation Research Record: Journal of the Transportation Research Board*, pp. 92–101. DOI: 10.3141/2014-12.
- Hörl, S. et al. (2026). “Towards a Replicable Synthetic Population and Agent-Based Transport Model for Bavaria”. In: *Advances in Practical Applications of Agents, Multi-Agent Systems, and Computational Social Science: The PAAMS Collection*. Vol. 16031. Springer Nature Switzerland, pp. 179–190. DOI: 10.1007/978-3-032-07638-0\_15.
- Hörl, Sebastian and Milos Balac (2021). “Open synthetic travel demand for Paris and Île-de-France: Inputs and output data”. In: 39.ISSN 2352-3409, p. 107622. DOI: 10.1016/j.dib.2021.107622.
- Jensen, Finn V and Thomas D Nielsen (2007). *Bayesian Networks and Decision Graphs*. second. Springer Science + Business Media, LLC, p. 457. URL: <https://springer.com> (Accessed May 18th, 2025).
- Joubert, Johan W and Alta de Waal (2020). “Activity-based travel demand generation using Bayesian networks”. In: *Transportation Research Part C: Emerging Technologies* 120. DOI: 10.1016/j.trc.2020.102804.
- Lovelace, R. and D. Ballas (2013). “‘Truncate, replicate, sample’: A method for creating integer weights for spatial microsimulation”. In: *Computers, Environment and Urban Systems*, pp. 1–11. DOI: 10.1016/j.compenvurbsys.2013.03.004.
- Lovelace, R., M. Birkin, et al. (2015). “Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique”. In: *Journal of Artificial Societies and Social Simulation*, p. 21. DOI: 10.18564/jasss.2768.
- Lovelace, R. and M. Dumont (2016). *Spatial microsimulation with R*. CRC Press, Taylor & Francis Group. URL: <https://www.ibs.it/spatial-microsimulation-with-r-libro-inglese-robin-lovelace-morgane-dumont/e/9781498711548?srsltid>

=AfmB0oouElZh12NbPYyvwNEnF9kU0CL-vaHM2Ee8umQn96\_CKAmYPcd1 (Accessed April 29th, 2026).

- Ma, L. and S. Srinivasan (2016). “An empirical assessment of factors affecting the accuracy of target-year synthetic populations”. In: *Transportation Research Part A: Policy and Practice*, pp. 247–264. DOI: 10.1016/j.tra.2016.01.016.
- Müller, K. and K. W. Axhausen (2010). *Population synthesis for microsimulation: State of the art*. URL: [https://www.researchgate.net/publication/228973867\\_Population\\_synthesis\\_for\\_microsimulation\\_State\\_of\\_the\\_art](https://www.researchgate.net/publication/228973867_Population_synthesis_for_microsimulation_State_of_the_art) (Accessed April 29th, 2026).
- (2012). *Multi-level fitting algorithms for population synthesis*. DOI: 10.3929/ETHZ-A-010192943.
- Nagel, K., A. Horni, and K. W. Axhausen (2016). “The Multi-Agent Transport Simulation MATSim”. In: *The Multi-Agent Transport Simulation MATSim*. London, pp. 3–8. DOI: 10.5334/baw.1.
- Namazi-Rad, M.-R., P. Mokhtarian, and P. Perez (2014). “Generating a Dynamic Synthetic Population – Using an Age-Structured Two-Sex Model for Household Dynamics”. In: *PLoS ONE*. DOI: 10.1371/journal.pone.0094761.
- Nguyen, N. A., J. Schweizer, and F. Rupi (2025). “Large-scale activity-based demand generation modeling: A literature review and exploration of potential approaches”. In: *Transportation Engineering*. DOI: 10.1016/j.treng.2025.100329.
- Raux, Charles, Tai-Yu Ma, and Eric Cornelis (2016). “Variability in daily activity-travel patterns: the case of a one-week travel diary”. In: *European Transport Research Review* 8.26, p. 26. DOI: 10.1007/s12544-016-0213-9.
- Roorda, Matthew J, Eric J Miller, and Khandker M. Nurul Habib (2007). “Validation of TASHA: A 24-Hour Activity Scheduling Microsimulation Model”. In: *Transportation Research Board Annual Meeting 86th Annual Meeting*, pp. 1–17. DOI: 10.1016/j.tr a.2007.10.004.
- Saadi, Ismail et al. (2016). “Hidden Markov Model-based population synthesis”. In: *Transportation Research Part C*. DOI: 10.1016/j.trb.2016.04.007.
- Saadi, Ismail et al. (2016). “Forecasting travel behavior using Markov Chains-based approaches”. In: *Transportation Research Part C: Emerging Technologies* 69, pp. 402–417. DOI: 10.1016/j.trc.2016.06.020.

- Sun, Lijun and Alexander Erath (2015). “A Bayesian network approach for population synthesis”. In: *Transportation Research Part C: Emerging Technologies* 61, pp. 49–62. DOI: 10.1016/j.trc.2015.10.010.
- Tanton, R. and K. L. Edwards (2013). *Spatial microsimulation: A reference guide for users*. Springer. URL: <https://link.springer.com/book/10.1007/978-94-007-4623-7> (Accessed April 29th, 2026).
- Waddell, Paul and Gudmundur Freyr Ulfarsson (2004). “Introduction to Urban Simulation: Design and Development of Operational Models”. In: *Handbook in Transport*. DOI: 10.1108/9781615832538-013.
- Yagi, Sadayuki and Abolfazl Mohammadian (2009). “An Activity-Based Microsimulation Model of Travel Demand in the Jakarta Metropolitan Area”. In: pp. 32–57. DOI: 10.1016/S1755-5345(13)70028-9.
- Ye, X. et al. (2009). *A methodology to match distributions of both household and person attributes in the generation of synthetic populations*. URL: [https://www.researchgate.net/publication/228963837\\_Methodology\\_to\\_match\\_distributions\\_of\\_both\\_household\\_and\\_person\\_attributes\\_in\\_generation\\_of\\_synthetic\\_populations](https://www.researchgate.net/publication/228963837_Methodology_to_match_distributions_of_both_household_and_person_attributes_in_generation_of_synthetic_populations) (Accessed April 29th, 2026).
- Zhang, Danqing et al. (2019). “Connected Population Synthesis for Transportation Simulation”. In: *Transportation Research Part C: Emerging Technologies*, pp. 1–16. DOI: 10.1016/j.trc.2018.12.014.
- Zhuo, H. (2017). *Synthesizing population for travel activity analysis*. URL: <https://mind.s.wiscnsin.edu/handle/1793/91598> (Accessed April 29th, 2026).

# Appendices

**A.1 Available variables from 2011 census (including variables not used in this study); the variables used in this study are shown in *italics*.**

| Group                                          | Variable       | Definition                                                                                                                               |
|------------------------------------------------|----------------|------------------------------------------------------------------------------------------------------------------------------------------|
| Geographical identifiers                       | CODREG         | Unique numeric identifier of the region within the national territory                                                                    |
| Geographical identifiers                       | REGIONE        | Name of the region                                                                                                                       |
| Geographical identifiers                       | CODPRO         | Unique numeric identifier of the province within the national territory                                                                  |
| Geographical identifiers                       | PROVINCIA      | Name of the province                                                                                                                     |
| Geographical identifiers                       | CODCOM         | Unique numeric identifier of the municipality within the province                                                                        |
| Geographical identifiers                       | COMUNE         | Name of the municipality                                                                                                                 |
| Geographical identifiers                       | PROCUM         | National municipality code obtained by concatenating province and municipality codes                                                     |
| Geographical identifiers                       | <i>SEZ2011</i> | Unique identifier of the 2011 census section at national level                                                                           |
| Geographical identifiers                       | NSEZ           | Census section identifier within the municipality (including special codes for specific cases such as homeless or temporary populations) |
| Geographical identifiers                       | ACE            | Unique identifier of the census enumeration area within the municipality                                                                 |
| Geographical identifiers                       | CODLOC         | Identifier of locality within the municipality (5-digit structured code)                                                                 |
| Geographical identifiers                       | CODASC         | Identifier of sub-municipal statistical area                                                                                             |
| Population structure by sex and marital status | <i>P1</i>      | Total resident population                                                                                                                |
| Population structure by sex and marital status | <i>P2</i>      | Total male resident population                                                                                                           |
| Population structure by sex and marital status | <i>P4</i>      | Single (never married) population                                                                                                        |
| Population structure by sex and marital status | <i>P5</i>      | Married population (including de facto separated)                                                                                        |

| Group                                          | Variable   | Definition                                   |
|------------------------------------------------|------------|----------------------------------------------|
| Population structure by sex and marital status | P6         | Legally separated population                 |
| Population structure by sex and marital status | P7         | Widowed population                           |
| Population structure by sex and marital status | P8         | Divorced population                          |
| Population structure by sex and marital status | P9         | Single males                                 |
| Population structure by sex and marital status | P10        | Married males (including de facto separated) |
| Population structure by sex and marital status | P11        | Legally separated males                      |
| Population structure by sex and marital status | P12        | Widowed males                                |
| Population structure by sex and marital status | P13        | Divorced males                               |
| Population distribution by age group           | P14        | Population aged 0–4                          |
| Population distribution by age group           | P15        | Population aged 5–9                          |
| Population distribution by age group           | P16        | Population aged 10–14                        |
| Population distribution by age group           | <i>P17</i> | Population aged 15–19                        |
| Population distribution by age group           | <i>P18</i> | Population aged 20–24                        |
| Population distribution by age group           | <i>P19</i> | Population aged 25–29                        |
| Population distribution by age group           | <i>P20</i> | Population aged 30–34                        |
| Population distribution by age group           | <i>P21</i> | Population aged 35–39                        |
| Population distribution by age group           | <i>P22</i> | Population aged 40–44                        |
| Population distribution by age group           | <i>P23</i> | Population aged 45–49                        |
| Population distribution by age group           | <i>P24</i> | Population aged 50–54                        |
| Population distribution by age group           | <i>P25</i> | Population aged 55–59                        |
| Population distribution by age group           | <i>P26</i> | Population aged 60–64                        |
| Population distribution by age group           | <i>P27</i> | Population aged 65–69                        |

| <b>Group</b>                                             | <b>Variable</b> | <b>Definition</b>                                             |
|----------------------------------------------------------|-----------------|---------------------------------------------------------------|
| Population distribution by age group                     | <i>P28</i>      | Population aged 70–74                                         |
| Population distribution by age group                     | <i>P29</i>      | Population aged 75 and over                                   |
| Male population distribution by age group                | <i>P30</i>      | Males aged 0–4                                                |
| Male population distribution by age group                | <i>P31</i>      | Males aged 5–9                                                |
| Male population distribution by age group                | <i>P32</i>      | Males aged 10–14                                              |
| Male population distribution by age group                | <i>P33</i>      | Males aged 15–19                                              |
| Male population distribution by age group                | <i>P34</i>      | Males aged 20–24                                              |
| Male population distribution by age group                | <i>P35</i>      | Males aged 25–29                                              |
| Male population distribution by age group                | <i>P36</i>      | Males aged 30–34                                              |
| Male population distribution by age group                | <i>P37</i>      | Males aged 35–39                                              |
| Male population distribution by age group                | <i>P38</i>      | Males aged 40–44                                              |
| Male population distribution by age group                | <i>P39</i>      | Males aged 45–49                                              |
| Male population distribution by age group                | <i>P40</i>      | Males aged 50–54                                              |
| Male population distribution by age group                | <i>P41</i>      | Males aged 55–59                                              |
| Male population distribution by age group                | <i>P42</i>      | Males aged 60–64                                              |
| Male population distribution by age group                | <i>P43</i>      | Males aged 65–69                                              |
| Male population distribution by age group                | <i>P44</i>      | Males aged 70–74                                              |
| Male population distribution by age group                | <i>P45</i>      | Males aged 75 and over                                        |
| Population by educational attainment and literacy status | <i>P46</i>      | Population aged 6 and over                                    |
| Population by educational attainment and literacy status | <i>P47</i>      | Population with tertiary education (university or equivalent) |
| Population by educational attainment and literacy status | <i>P48</i>      | Population with upper secondary education                     |

| <b>Group</b>                                                     | <b>Variable</b> | <b>Definition</b>                         |
|------------------------------------------------------------------|-----------------|-------------------------------------------|
| Population by educational attainment and literacy status         | <i>P49</i>      | Population with lower secondary education |
| Population by educational attainment and literacy status         | <i>P50</i>      | Population with primary education         |
| Population by educational attainment and literacy status         | P51             | Literate population                       |
| Population by educational attainment and literacy status         | <i>P52</i>      | Illiterate population                     |
| Male population by educational attainment and literacy status    | P53             | Males aged 6 and over                     |
| Male population by educational attainment and literacy status    | <i>P54</i>      | Males with tertiary education             |
| Male population by educational attainment and literacy status    | <i>P55</i>      | Males with upper secondary education      |
| Male population by educational attainment and literacy status    | <i>P56</i>      | Males with lower secondary education      |
| Male population by educational attainment and literacy status    | <i>P57</i>      | Males with primary education              |
| Male population by educational attainment and literacy status    | P58             | Literate males                            |
| Male population by educational attainment and literacy status    | <i>P59</i>      | Illiterate males                          |
| Resident population by labor force status (age 15 and over)      | P60             | Population aged 15+ in the labor force    |
| Resident population by labor force status (age 15 and over)      | <i>P61</i>      | Employed population (15+)                 |
| Resident population by labor force status (age 15 and over)      | P62             | Unemployed population seeking work        |
| Male resident population by labor force status (age 15 and over) | P64             | Males aged 15+ in the labor force         |

| Group                                                                           | Variable | Definition                                       |
|---------------------------------------------------------------------------------|----------|--------------------------------------------------|
| Male resident population by labor force status (age 15 and over)                | P65      | Employed males                                   |
| Male resident population by labor force status (age 15 and over)                | P66      | Unemployed males                                 |
| Resident population not in the labor force by activity status (age 15 and over) | P128     | Population aged 15+ not in the labour force      |
| Resident population not in the labor force by activity status (age 15 and over) | P129     | Males aged 15+ not in the labour force           |
| Resident population not in the labor force by activity status (age 15 and over) | P130     | Homemakers                                       |
| Resident population not in the labor force by activity status (age 15 and over) | P131     | Students (total)                                 |
| Resident population not in the labor force by activity status (age 15 and over) | P132     | Male students                                    |
| Resident population not in the labor force by activity status (age 15 and over) | P135     | Population in other inactive conditions          |
| Resident population not in the labor force by activity status (age 15 and over) | P136     | Males in other inactive conditions               |
| Population by commuting behavior and income status                              | P137     | Residents commuting within their municipality    |
| Population by commuting behavior and income status                              | P138     | Residents commuting outside their municipality   |
| Population by commuting behavior and income status                              | P139     | Population with income (labour or capital)       |
| Population by commuting behavior and income status                              | P140     | Males with income                                |
| Foreign and stateless population                                                | ST1      | Total number of foreign and stateless residents  |
| Foreign and stateless population                                                | ST2      | Male foreign and stateless residents             |
| Foreign and stateless population                                                | ST3      | Foreign and stateless residents aged 0–29        |
| Foreign and stateless population                                                | ST4      | Foreign and stateless residents aged 30–54       |
| Foreign and stateless population                                                | ST5      | Foreign and stateless residents aged 55 and over |

| Group                            | Variable   | Definition                                                                                    |
|----------------------------------|------------|-----------------------------------------------------------------------------------------------|
| Foreign and stateless population | ST6        | Male foreign and stateless residents aged 0–29                                                |
| Foreign and stateless population | ST7        | Male foreign and stateless residents aged 30–54                                               |
| Foreign and stateless population | ST8        | Male foreign and stateless residents aged 55 and over                                         |
| Foreign and stateless population | ST9        | Foreign residents from Europe                                                                 |
| Foreign and stateless population | ST10       | Foreign residents from Europe                                                                 |
| Foreign and stateless population | ST11       | Foreign residents from the Americas                                                           |
| Foreign and stateless population | ST12       | Foreign residents from Asia                                                                   |
| Foreign and stateless population | ST13       | Foreign residents from Oceania                                                                |
| Foreign and stateless population | ST14       | Stateless residents                                                                           |
| Foreign and stateless population | ST15       | Total number of foreign residents (excluding stateless where applicable)                      |
| Housing characteristics          | A2         | Dwellings occupied by at least one resident                                                   |
| Housing characteristics          | A3         | Dwellings either vacant or occupied only by non-residents                                     |
| Housing characteristics          | A5         | Other types of occupied accommodation (non-standard housing)                                  |
| Housing characteristics          | A6         | Vacant dwellings                                                                              |
| Housing characteristics          | A7         | Dwellings occupied exclusively by non-residents                                               |
| Housing characteristics          | A44        | Total floor area of dwellings occupied by residents                                           |
| Housing characteristics          | A46        | Households living in rented dwellings                                                         |
| Housing characteristics          | A47        | Households living in owner-occupied dwellings                                                 |
| Housing characteristics          | A48        | Households occupying dwellings under other tenure arrangements                                |
| Household structure              | PF1        | Total number of resident households                                                           |
| Household structure              | PF2        | Total number of individuals living in households                                              |
| Household structure              | <i>PF3</i> | Households with 1 member                                                                      |
| Household structure              | <i>PF4</i> | Households with 2 members                                                                     |
| Household structure              | <i>PF5</i> | Households with 3 members                                                                     |
| Household structure              | <i>PF6</i> | Households with 4 members                                                                     |
| Household structure              | <i>PF7</i> | Households with 5 members                                                                     |
| Household structure              | <i>PF8</i> | Households with 6 or more members                                                             |
| Household structure              | <i>PF9</i> | Total number of persons living in households with 6 or more members                           |
| Building characteristics         | E1         | Total number of buildings and building complexes                                              |
| Building characteristics         | E2         | Buildings currently in use                                                                    |
| Building characteristics         | E3         | Buildings used for residential purposes                                                       |
| Building characteristics         | E4         | Buildings used for non-residential purposes (productive, commercial, services, tourism, etc.) |
| Building characteristics         | E5         | Residential buildings constructed with load-bearing masonry                                   |

| Group                    | Variable | Definition                                                           |
|--------------------------|----------|----------------------------------------------------------------------|
| Building characteristics | E6       | Residential buildings constructed with reinforced concrete           |
| Building characteristics | E7       | Residential buildings constructed with other materials (e.g., steel) |
| Building characteristics | E8       | Residential buildings constructed before 1919                        |
| Building characteristics | E9       | Residential buildings constructed between 1919 and 1945              |
| Building characteristics | E10      | Residential buildings constructed between 1946 and 1960              |
| Building characteristics | E11      | Residential buildings constructed between 1961 and 1970              |
| Building characteristics | E12      | Residential buildings constructed between 1971 and 1980              |
| Building characteristics | E13      | Residential buildings constructed between 1981 and 1990              |
| Building characteristics | E14      | Residential buildings constructed between 1991 and 2000              |
| Building characteristics | E15      | Residential buildings constructed between 2001 and 2005              |
| Building characteristics | E16      | Residential buildings constructed after 2005                         |
| Building characteristics | E17      | Residential buildings with one floor                                 |
| Building characteristics | E18      | Residential buildings with two floors                                |
| Building characteristics | E19      | Residential buildings with three floors                              |
| Building characteristics | E20      | Residential buildings with four or more floors                       |
| Building characteristics | E21      | Residential buildings with one housing unit                          |
| Building characteristics | E22      | Residential buildings with two housing units                         |
| Building characteristics | E23      | Residential buildings with 3–4 housing units                         |
| Building characteristics | E24      | Residential buildings with 5–8 housing units                         |
| Building characteristics | E25      | Residential buildings with 9–15 housing units                        |
| Building characteristics | E26      | Residential buildings with 16 or more housing units                  |
| Building characteristics | E27      | Total number of housing units within residential buildings           |
| Building characteristics | E28      | Residential buildings in excellent condition                         |
| Building characteristics | E29      | Residential buildings in good condition                              |
| Building characteristics | E30      | Residential buildings in fair condition                              |
| Building characteristics | E31      | Residential buildings in poor condition                              |

Table 21: Attributes in census aggregate dataset (source: census 2011)

## A.2 Control variables include original and derived census variables

| Variable (new name) | Definition              | Derivation formula |
|---------------------|-------------------------|--------------------|
| Male                | Total male population   | P2                 |
| Female              | Total female population | P1 – P2            |
| age_0_14            | Population aged 0–14    | P14 + P15 + P16    |
| age_15_19           | Population aged 15–19   | P17                |
| age_20_24           | Population aged 20–24   | P18                |
| age_25_29           | Population aged 25–29   | P19                |
| age_30_34           | Population aged 30–34   | P20                |

| Variable (new name) | Definition                     | Derivation formula                   |
|---------------------|--------------------------------|--------------------------------------|
| age_35_39           | Population aged 35–39          | P21                                  |
| age_40_44           | Population aged 40–44          | P22                                  |
| age_45_49           | Population aged 45–49          | P23                                  |
| age_50_54           | Population aged 50–54          | P24                                  |
| age_55_59           | Population aged 55–59          | P25                                  |
| age_60_64           | Population aged 60–64          | P26                                  |
| age_65_69           | Population aged 65–69          | P27                                  |
| age_70_74           | Population aged 70–74          | P28                                  |
| age_75p             | Population aged 75 and above   | P29                                  |
| mage_0_14           | Male population aged 0–14      | P30 + P31 + P32                      |
| mage_15_19          | Male population aged 15–19     | P33                                  |
| mage_20_24          | Male population aged 20–24     | P34                                  |
| mage_25_29          | Male population aged 25–29     | P35                                  |
| mage_30_34          | Male population aged 30–34     | P36                                  |
| mage_35_39          | Male population aged 35–39     | P37                                  |
| mage_40_44          | Male population aged 40–44     | P38                                  |
| mage_45_49          | Male population aged 45–49     | P39                                  |
| mage_50_54          | Male population aged 50–54     | P40                                  |
| mage_55_59          | Male population aged 55–59     | P41                                  |
| mage_60_64          | Male population aged 60–64     | P42                                  |
| mage_65_69          | Male population aged 65–69     | P43                                  |
| mage_70_74          | Male population aged 70–74     | P44                                  |
| mage_75p            | Male population aged 75+       | P45                                  |
| fage_0_14           | Female population aged 0–14    | age_0_14 – mage_0_14                 |
| fage_15_19          | Female population aged 15–19   | age_15_19 – mage_15_19               |
| fage_20_24          | Female population aged 20–24   | age_20_24 – mage_20_24               |
| fage_25_29          | Female population aged 25–29   | age_25_29 – mage_25_29               |
| fage_30_34          | Female population aged 30–34   | age_30_34 – mage_30_34               |
| fage_35_39          | Female population aged 35–39   | age_35_39 – mage_35_39               |
| fage_40_44          | Female population aged 40–44   | age_40_44 – mage_40_44               |
| fage_45_49          | Female population aged 45–49   | age_45_49 – mage_45_49               |
| fage_50_54          | Female population aged 50–54   | age_50_54 – mage_50_54               |
| fage_55_59          | Female population aged 55–59   | age_55_59 – mage_55_59               |
| fage_60_64          | Female population aged 60–64   | age_60_64 – mage_60_64               |
| fage_65_69          | Female population aged 65–69   | age_65_69 – mage_65_69               |
| fage_70_74          | Female population aged 70–74   | age_70_74 – mage_70_74               |
| fage_75p            | Female population aged 75+     | age_75p – mage_75p                   |
| emp_employed        | Total employed population      | P61                                  |
| emp_not_employed    | Total not employed population  | P1 - P61                             |
| memp_employed       | Employed male population       | P65                                  |
| memp_not_employed   | Not employed male population   | P2 - P65                             |
| femp_employed       | Employed female population     | emp_employed - memp_employed         |
| femp_not_employed   | Not employed female population | emp_not_employed - memp_not_employed |

| Variable (new name) | Definition                      | Derivation formula                                                        |
|---------------------|---------------------------------|---------------------------------------------------------------------------|
| edu_illit_primary   | Illiterate or primary education | P50 + P52                                                                 |
| edu_lowersec        | Lower secondary education       | P49                                                                       |
| edu_uppersec        | Upper secondary education       | P48                                                                       |
| edu_tertiary        | Tertiary education              | P47                                                                       |
| edu_other_na        | Other or missing education      | P1 - (edu_illit_primary + edu_lowersec + edu_uppersec + edu_tertiary)     |
| medu_illit_primary  | Male illiterate/primary         | P57 + P59                                                                 |
| medu_lowersec       | Male lower secondary            | P56                                                                       |
| medu_uppersec       | Male upper secondary            | P55                                                                       |
| medu_tertiary       | Male tertiary                   | P54                                                                       |
| medu_other_na       | Male other/missing              | P2 - (medu_illit_primary + medu_lowersec + medu_uppersec + medu_tertiary) |
| fedu_illit_primary  | Female illiterate/primary       | edu_illit_primary - medu_illit_primary                                    |
| fedu_lowersec       | Female lower secondary          | edu_lowersec - medu_lowersec                                              |
| fedu_uppersec       | Female upper secondary          | edu_uppersec - medu_uppersec                                              |
| fedu_tertiary       | Female tertiary                 | edu_tertiary - medu_tertiary                                              |
| fedu_other_na       | Female other/missing            | edu_other_na - medu_other_na                                              |

Table 22: Derived variables (source: own reclassification on census 2011 attributes)

### A.3 Variables and English definitions of the 2011 1% sample census microdata

| Attribute     | Definition                                                                  |
|---------------|-----------------------------------------------------------------------------|
| ID_FAM        | Household or collective residence identifier                                |
| PROGR_IND     | Individual identifier within the household                                  |
| ID_IND_CONV   | Individual identifier within a collective residence                         |
| TIPCONV       | Type of collective residence                                                |
| MOPERC        | Main reason for staying in a collective residence                           |
| ID_ALL        | Dwelling identifier                                                         |
| REGIONE       | Region of residence                                                         |
| RIP_GEO       | Geographic macro-area of residence                                          |
| NROCOMPO      | Number of household members                                                 |
| TIPOLOGIA_FAM | Household type                                                              |
| REL_PAR       | Relationship to the household reference person or cohabitation relationship |
| SESSO         | Sex                                                                         |
| ETA_CLASSI    | Age group                                                                   |

| Attribute           | Definition                                                                                                                    |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------|
| LUOGO_NAS           | Place of birth                                                                                                                |
| STATO_NAS           | Foreign country of birth code                                                                                                 |
| CITTADINANZA        | Citizenship                                                                                                                   |
| STATO_CITT          | Country of citizenship code                                                                                                   |
| CITT_NAS_ACQ        | Italian citizenship from birth                                                                                                |
| CITT_ITA_ACQ        | Italian citizenship acquired through marriage or other means                                                                  |
| LUONAS_MADRE        | Mother's place of birth                                                                                                       |
| STATO_NAS_MADRE     | Mother's country of birth                                                                                                     |
| LUONAS_PADRE        | Father's place of birth                                                                                                       |
| STATO_NAS_PADRE     | Father's country of birth                                                                                                     |
| RESIDENTE_ESTERO    | Has ever resided abroad                                                                                                       |
| MESE TRASFERIMENTO  | Month of relocation                                                                                                           |
| ANNO TRASFERIMENTO  | Year of relocation                                                                                                            |
| STATO_EST_ULT_RESID | Foreign country of last residence                                                                                             |
| STATO_CIV           | Marital status                                                                                                                |
| STATO_CIV_PREC      | Marital status before the most recent marriage                                                                                |
| LUOGO_PRES_CENS     | Place where the person was present on the census date                                                                         |
| DIM_ANNO_PREC       | Usual residence on 9 October 2010                                                                                             |
| PROV_DIM_ANNO_PREC  | Province code of usual residence on 9 October 2010                                                                            |
| STATO_DIM_ANNO_PREC | Country code of usual residence on 9 October 2010                                                                             |
| FREQ_SCUOLA_FIGLI   | Type of school attendance of the child                                                                                        |
| TIT_STUD            | Highest educational attainment                                                                                                |
| POST_LAUREA         | Postgraduate specialization and/or doctoral qualification                                                                     |
| TIT_STUD_EST        | Highest educational qualification obtained abroad                                                                             |
| ISCR_SCUOLA         | Enrollment in primary school, lower secondary school, upper secondary school, or university                                   |
| FREQ_CORSO_FORM     | Attendance of a vocational training or professional development course                                                        |
| CORSO_STAT6         | Type of course attended: vocational, regional, provincial, or similar course lasting 6 months                                 |
| CORSO_STAT24        | Type of course attended: vocational, regional, provincial, or similar course lasting 24 months                                |
| COND_PROF           | Employment status or non-employment status in the week before the census                                                      |
| FATTE_ORE           | Hours worked                                                                                                                  |
| ASSENZA_LAVORO      | Whether the person had a job but was absent from work during the week 2–8 October                                             |
| CERCA_LAV           | Whether, in the last 4 weeks, the person actively looked for paid work or made arrangements to start a self-employed activity |
| DISPO_LAV           | Availability to start work within 2 weeks                                                                                     |
| SVOLTO_LAV          | Whether the person has ever carried out paid work or worked as an unpaid family worker during their lifetime                  |

| Attribute            | Definition                                                                   |
|----------------------|------------------------------------------------------------------------------|
| CONDIZ_INATTIVI      | Status of economically inactive persons during the week 2–8 October          |
| ORE_LAV_SETT_PREC    | Number of hours worked during the week 14–20 October 2001                    |
| TEMPO_PIENO          | Type of work schedule                                                        |
| POS_PROF             | Employment position                                                          |
| HA_DIP               | Has paid employees                                                           |
| LAV_A_TEMPO          | Permanent or fixed-term employment contract                                  |
| TIPO_LAV             | Type of occupation                                                           |
| SETT_ATECO           | Economic activity sector                                                     |
| STUD_LAV_OGNI_G      | Whether the person travels daily to their usual place of study or work       |
| ALL_STUD_LAV         | From which dwelling the person travels to their usual place of study or work |
| RIENTRA              | Whether the person returns daily from their usual place of study or work     |
| LUOGO_STUD_LAV       | Usual place of study or work                                                 |
| REG_STUD_LAV         | Region code of place of study or work                                        |
| STATO_STUD_LAV       | Country code of place of study or work                                       |
| ORA_VA_STUD_LAV      | Departure time to the usual place of study or work: hour                     |
| MIN_VA_STUD_LAV      | Departure time to the usual place of study or work: minute                   |
| TEMPO_VA_STUD_LAV    | Travel time to the usual place of study or work                              |
| MEZZO_TRAS_STUD_LAV  | Means of transport used to reach the usual place of study or work            |
| DIM_5ANNO_PREC       | Usual residence on 9 October 2006                                            |
| PROV_DIM_5ANNO_PREC  | Province code of usual residence on 9 October 2006                           |
| STATO_DIM_5ANNO_PREC | Foreign country code of usual residence on 9 October 2006                    |
| IND_MODELLO          | Questionnaire form type                                                      |

Table 23: Attributes in 1% census individual dataset (source: census 2011)

## A.4 Categories and codes for the categorical variables in the 2011 census 1% individual sample dataset

| Variable | Category                                                           | Corresponding code |
|----------|--------------------------------------------------------------------|--------------------|
|          | Educational institutions                                           | 1                  |
|          | Residential facilities for minors                                  | 2                  |
|          | Care institutions for people with physical and mental disabilities | 3                  |

| Variable | Category                                                          | Corresponding code |
|----------|-------------------------------------------------------------------|--------------------|
|          | Hospices and nursing homes for disabled adults and elderly people | 4                  |
|          | Reception facilities for immigrants                               | 5                  |
|          | Other care institutions                                           | 6                  |
|          | Public                                                            | 7                  |
|          | Private                                                           | 8                  |
|          | Penal institutions                                                | 9                  |
|          | Ecclesiastical communities                                        | 10                 |
|          | Military communities and barracked corps                          | 11                 |
|          | Hotels, boarding houses, inns, and similar accommodation          | 12                 |
|          | Other collective living arrangements                              | 14                 |
| MOPERC   | Head or manager of the collective residence                       | 1                  |
|          | Healthcare staff                                                  | 2                  |
|          | Social or psychological care staff                                | 3                  |
|          | Administrative or general services staff                          | 4                  |
|          | Maintenance, cleaning, and other services staff                   | 5                  |
|          | Religious member                                                  | 6                  |
|          | Career military personnel, police officer, and similar            | 7                  |
|          | Prisoner with a final sentence                                    | 8                  |
|          | Resident in a reception center for immigrants                     | 9                  |
|          | Long-term inpatient in a care institution                         | 10                 |
|          | Resident in an institution for the elderly                        | 11                 |
|          | Resident in an institution for minors                             | 12                 |
|          | Resident in another reception center                              | 13                 |
|          | Other                                                             | 14                 |
| REGIONE  | Piedmont–Aosta Valley                                             | 1                  |
|          | Lombardy                                                          | 3                  |
|          | Trentino–South Tyrol                                              | 4                  |
|          | Veneto                                                            | 5                  |
|          | Friuli–Venezia Giulia                                             | 6                  |
|          | Liguria                                                           | 7                  |
|          | Emilia–Romagna                                                    | 8                  |
|          | Tuscany                                                           | 9                  |
|          | Umbria                                                            | 10                 |
|          | Marche                                                            | 11                 |
|          | Lazio                                                             | 12                 |
|          | Abruzzo                                                           | 13                 |
|          | Molise                                                            | 14                 |

| Variable      | Category                                                                   | Corresponding code |
|---------------|----------------------------------------------------------------------------|--------------------|
|               | Campania                                                                   | 15                 |
|               | Apulia                                                                     | 16                 |
|               | Basilicata                                                                 | 17                 |
|               | Calabria                                                                   | 18                 |
|               | Sicily                                                                     | 19                 |
|               | Sardinia                                                                   | 20                 |
| RIP_GEO       | North-West                                                                 | 1                  |
|               | North-East                                                                 | 2                  |
|               | Central Italy                                                              | 3                  |
|               | Southern Italy                                                             | 4                  |
|               | Islands                                                                    | 5                  |
| NROCOMPO      | 1 member                                                                   | 1                  |
|               | 2 members                                                                  | 2                  |
|               | 3 members                                                                  | 3                  |
|               | 4 members                                                                  | 4                  |
|               | 5 members                                                                  | 5                  |
|               | 6 members                                                                  | 6                  |
| TIPOLOGIA_FAM | Families without a nucleus                                                 | 1                  |
|               | Not in cohabitation                                                        | 2                  |
|               | In cohabitation                                                            | 3                  |
|               | Families with one nucleus                                                  | 4                  |
|               | Couples without children                                                   | 5                  |
|               | Couples with children                                                      | 6                  |
|               | Mother with children                                                       | 7                  |
|               | Father with children                                                       | 8                  |
|               | Couples without children                                                   | 9                  |
|               | Couples with children                                                      | 10                 |
|               | Mother with children                                                       | 11                 |
|               | Father with children                                                       | 12                 |
|               | Families with two or more nuclei                                           | 13                 |
| REL_PAR       | Head of household                                                          | 1                  |
|               | Spouse of the head of household                                            | 2                  |
|               | Cohabiting partner of the head of household                                | 3                  |
|               | Child of the head of household and of the spouse/cohabiting partner        | 4                  |
|               | Child of the head of household only                                        | 5                  |
|               | Child of the spouse/cohabiting partner only                                | 6                  |
|               | Parent of the head of household, or the parent's spouse/cohabiting partner | 7                  |

| Variable   | Category                                                                                                                           | Corresponding code |
|------------|------------------------------------------------------------------------------------------------------------------------------------|--------------------|
|            | Parent-in-law of the head of household                                                                                             | 8                  |
|            | Brother/sister of the head of household                                                                                            | 9                  |
|            | Brother/sister of the spouse/cohabiting partner                                                                                    | 10                 |
|            | Spouse/cohabiting partner of the brother/sister of the head of household or of the brother/sister of the spouse/cohabiting partner | 11                 |
|            | Son-in-law/daughter-in-law of the head of household and/or of the spouse/cohabiting partner                                        | 12                 |
|            | Grandchild of the head of household and/or of the spouse/cohabiting partner                                                        | 13                 |
|            | Nephew/niece of the head of household and/or of the spouse/cohabiting partner                                                      | 14                 |
|            | Grandparent of the head of household or of the spouse/cohabiting partner                                                           | 15                 |
|            | Other relative of the head of household and/or of the spouse/cohabiting partner                                                    | 16                 |
|            | Other cohabiting person with no couple, kinship, or in-law relationship                                                            | 17                 |
| SESSO      | Male                                                                                                                               | 1                  |
|            | Female                                                                                                                             | 2                  |
| ETA_CLASSI | 0–14                                                                                                                               | 1                  |
|            | 15–19                                                                                                                              | 2                  |
|            | 20–24                                                                                                                              | 3                  |
|            | 25–29                                                                                                                              | 4                  |
|            | 30–34                                                                                                                              | 5                  |
|            | 35–39                                                                                                                              | 6                  |
|            | 40–44                                                                                                                              | 7                  |
|            | 45–49                                                                                                                              | 8                  |
|            | 50–54                                                                                                                              | 9                  |
|            | 55–59                                                                                                                              | 10                 |
|            | 60–64                                                                                                                              | 11                 |
|            | 65–69                                                                                                                              | 12                 |
|            | 70–74                                                                                                                              | 13                 |
|            | 75+                                                                                                                                | 14                 |
| LUOGO_NAS  | In the same municipality of residence                                                                                              | 1                  |
|            | In another municipality in the same province                                                                                       | 2                  |
|            | In another province in the same region                                                                                             | 3                  |
|            | In other regions                                                                                                                   | 4                  |
|            | Abroad                                                                                                                             | 5                  |
| STATO_NAS  | EU Europe                                                                                                                          | 2                  |

| Variable           | Category                 | Corresponding code |
|--------------------|--------------------------|--------------------|
|                    | Non-EU Europe            | 3                  |
|                    | Rest of the world        | 4                  |
| CITTADINANZA       | Italian                  | 1                  |
|                    | Foreign                  | 2                  |
| STATO_CITT         | Italy                    | 1                  |
|                    | EU Europe                | 2                  |
|                    | Non-EU Europe            | 3                  |
|                    | Rest of the world        | 4                  |
| CITT_NAS_ACQ       | Yes                      | 1                  |
|                    | No                       | 2                  |
| CITT_ITA_ACQ       | Marriage                 | 1                  |
|                    | Other                    | 2                  |
| LUONAS_MADRE       | Italy                    | 1                  |
|                    | Abroad                   | 2                  |
| STATO_NAS_MADRE    | EU-27                    | 11                 |
|                    | Central-Eastern Europe   | 12                 |
|                    | Other European countries | 13                 |
|                    | Africa                   | 200                |
|                    | Asia                     | 300                |
|                    | America                  | 400                |
|                    | Oceania                  | 500                |
| LUONAS_PADRE       | Italy                    | 1                  |
|                    | Abroad                   | 2                  |
| STATO_NAS_PADRE    | EU-27                    | 11                 |
|                    | Central-Eastern Europe   | 12                 |
|                    | Other European countries | 13                 |
|                    | Africa                   | 200                |
|                    | Asia                     | 300                |
|                    | America                  | 400                |
|                    | Oceania                  | 500                |
| RESIDENTE_ESTERO   | Yes                      | 1                  |
|                    | No                       | 2                  |
| MESE_TRASFERIMENTO | January                  | 1                  |
|                    | February                 | 2                  |
|                    | March                    | 3                  |
|                    | April                    | 4                  |
|                    | May                      | 5                  |
|                    | June                     | 6                  |
|                    | July                     | 7                  |

| Variable            | Category                                                              | Corresponding code |
|---------------------|-----------------------------------------------------------------------|--------------------|
|                     | August                                                                | 8                  |
|                     | September                                                             | 9                  |
|                     | October                                                               | 10                 |
|                     | November                                                              | 11                 |
|                     | December                                                              | 12                 |
| STATO_EST_ULT_RESID | EU-27                                                                 | 11                 |
|                     | Central-Eastern Europe                                                | 12                 |
|                     | Other European countries                                              | 13                 |
|                     | Africa                                                                | 200                |
|                     | Asia                                                                  | 300                |
|                     | America                                                               | 400                |
|                     | Oceania                                                               | 500                |
| STATO_CIV           | Single                                                                | 1                  |
|                     | Married                                                               | 2                  |
|                     | De facto separated                                                    | 3                  |
|                     | Legally separated                                                     | 4                  |
|                     | Divorced                                                              | 5                  |
|                     | Widowed                                                               | 6                  |
| STATO_CIV_PREC      | Single                                                                | 1                  |
|                     | Divorced                                                              | 2                  |
|                     | Widowed                                                               | 3                  |
| LUOGO_PREC_CENS     | In this dwelling                                                      | 1                  |
|                     | In this municipality, but in another dwelling or collective residence | 2                  |
|                     | In another Italian municipality                                       | 3                  |
|                     | Abroad                                                                | 4                  |
| DIM_ANNO_PREC       | In this dwelling                                                      | 1                  |
|                     | In this municipality, but in another dwelling or collective residence | 2                  |
|                     | In another Italian municipality                                       | 3                  |
|                     | Abroad                                                                | 4                  |
| PROV_DIM_ANNO_PREC  | In the same municipality of residence                                 | 1                  |
|                     | In another municipality in the same province                          | 2                  |
|                     | In another province in the same region                                | 3                  |
|                     | In other regions                                                      | 4                  |
|                     | Abroad                                                                | 5                  |
| STATO_DIM_ANNO_PREC | Italy                                                                 | 1                  |
|                     | EU Europe                                                             | 2                  |
|                     | Non-EU Europe                                                         | 3                  |

| Variable          | Category                                                                       | Corresponding code |
|-------------------|--------------------------------------------------------------------------------|--------------------|
|                   | Rest of the world                                                              | 4                  |
| FREQ_SCUOLA_FIGLI | Nursery, micro-nursery, baby parking, etc. (3–36 months)                       | 1                  |
|                   | Preschool                                                                      | 2                  |
|                   | First year of primary school                                                   | 3                  |
|                   | Neither nursery, nor preschool, nor first year of primary school               | 4                  |
| TIT_STUD          | Illiterate or primary school certificate                                       | 1                  |
|                   | Lower secondary school certificate                                             | 2                  |
|                   | Vocational qualification diploma (2–3 year school course)                      | 3                  |
|                   | Upper secondary school diploma (4–5 year school course)                        | 4                  |
|                   | Non-university tertiary diploma                                                | 5                  |
|                   | University diploma                                                             | 6                  |
|                   | University degree                                                              | 7                  |
| POST_LAUREA       | First-level master's degree                                                    | 1                  |
|                   | Second-level master's degree                                                   | 2                  |
|                   | Specialization school                                                          | 3                  |
|                   | Doctorate (PhD)                                                                | 4                  |
| TIT_STUD_EST      | Yes                                                                            | 1                  |
|                   | No                                                                             | 2                  |
| ISCR_SCUOLA       | Yes                                                                            | 1                  |
|                   | No                                                                             | 2                  |
| FREQ_CORSO_FORM   | Yes                                                                            | 1                  |
|                   | No                                                                             | 2                  |
| CORSO_STAT6       | Yes                                                                            | 1                  |
|                   | No                                                                             | 2                  |
| CORSO_STAT24      | Yes                                                                            | 1                  |
|                   | No                                                                             | 2                  |
| COND_PROF         | Employed                                                                       | 1                  |
|                   | Seeking first employment                                                       | 2                  |
|                   | Unemployed, seeking a new job                                                  | 3                  |
|                   | Recipient of one or more pensions due to previous employment or capital income | 4                  |
|                   | Student                                                                        | 5                  |
|                   | Homemaker                                                                      | 6                  |
|                   | In another condition                                                           | 7                  |
| FATTE_ORE         | Yes                                                                            | 1                  |

| Variable        | Category                                                                                       | Corresponding code |
|-----------------|------------------------------------------------------------------------------------------------|--------------------|
|                 | No                                                                                             | 2                  |
| ASSENZA.LAVORO  | Yes                                                                                            | 1                  |
|                 | No                                                                                             | 2                  |
| CERCA.LAV       | Yes                                                                                            | 1                  |
|                 | No                                                                                             | 2                  |
| DISPO.LAV       | Yes                                                                                            | 1                  |
|                 | No                                                                                             | 2                  |
| SVOLTO.LAV      | Yes                                                                                            | 1                  |
|                 | No                                                                                             | 2                  |
| CONDIZ.INATTIVI | Recipient of one or more pensions due to previous employment or capital income                 | 1                  |
|                 | Student                                                                                        | 2                  |
|                 | Homemaker                                                                                      | 3                  |
|                 | In another condition                                                                           | 4                  |
| TEMPO.PIENO     | Full-time                                                                                      | 1                  |
|                 | Part-time                                                                                      | 2                  |
| POS.PROF        | Entrepreneur                                                                                   | 1                  |
|                 | Freelance professional                                                                         | 2                  |
|                 | Self-employed worker                                                                           | 3                  |
|                 | Cooperative member                                                                             | 4                  |
|                 | Family worker                                                                                  | 5                  |
|                 | Employee                                                                                       | 6                  |
|                 | Semi-dependent worker                                                                          | 7                  |
| HA.DIP          | Yes                                                                                            | 1                  |
|                 | No                                                                                             | 2                  |
| LAV.A.TEMPO     | Permanent contract                                                                             | 1                  |
|                 | Fixed-term contract                                                                            | 2                  |
| TIPO.LAV        | Unskilled manual or service work                                                               | 1                  |
|                 | Plant and machinery operator, assembly line worker, or vehicle driver                          | 2                  |
|                 | Skilled manual work                                                                            | 3                  |
|                 | Crop cultivation and/or animal husbandry                                                       | 4                  |
|                 | Retail sales or personal service activities                                                    | 5                  |
|                 | Clerical work                                                                                  | 6                  |
|                 | Technical, administrative, sports, or artistic activities with medium qualification            | 7                  |
|                 | Highly specialized organizational, technical, intellectual, scientific, or artistic activities | 8                  |

| Variable        | Category                                                                                | Corresponding code |
|-----------------|-----------------------------------------------------------------------------------------|--------------------|
|                 | Business management or leadership of complex public or private organizations            | 9                  |
|                 | Military personnel of any rank in the armed forces                                      | 10                 |
| SETT_ATECO      | Agriculture, forestry, hunting, and fishing                                             | 1                  |
|                 | Mining and quarrying, and support services for extraction                               | 2                  |
|                 | Manufacturing, and the repair, maintenance, and installation of machinery and equipment | 3                  |
|                 | Electricity, gas, steam, and air conditioning supply                                    | 4                  |
|                 | Water supply, sewerage, waste management, and remediation activities                    | 5                  |
|                 | Construction, civil engineering, and installation services in buildings                 | 6                  |
|                 | Wholesale and retail trade, and repair of motor vehicles and motorcycles                | 7                  |
|                 | Transport, warehousing, postal services, and courier activities                         | 8                  |
|                 | Accommodation and food service activities                                               | 9                  |
|                 | Information and communication services                                                  | 10                 |
|                 | Financial and insurance activities                                                      | 11                 |
|                 | Real estate activities                                                                  | 12                 |
|                 | Professional, scientific, and technical activities                                      | 13                 |
|                 | Rental, travel agencies, and business support services                                  | 14                 |
|                 | Central and local public administration, defence, and compulsory social security        | 15                 |
|                 | Public and private education and training                                               | 16                 |
|                 | Health and social care, residential and non-residential                                 | 17                 |
|                 | Arts, sports, entertainment, and recreation activities                                  | 18                 |
|                 | Other service activities and repair of personal and household goods                     | 19                 |
|                 | Activities of households and communal establishments as employers of domestic personnel | 20                 |
|                 | Extraterritorial organizations and bodies                                               | 21                 |
| STUD_LAV_OGNI_G | Yes, travels to the place of study, including vocational training courses               | 1                  |
|                 | Yes, travels to the workplace                                                           | 2                  |
|                 | No, because studies at home                                                             | 3                  |
|                 | No, because works at home                                                               | 4                  |

| Variable            | Category                                                                                  | Corresponding code |
|---------------------|-------------------------------------------------------------------------------------------|--------------------|
|                     | No, because does not have a fixed workplace, such as salespersons or representatives      | 5                  |
|                     | No, because does not study, does not work, or does not attend vocational training courses | 6                  |
| ALL_STUD_LAV        | From this dwelling                                                                        | 1                  |
|                     | From another dwelling                                                                     | 2                  |
| RIENTRA             | Yes                                                                                       | 1                  |
|                     | No                                                                                        | 2                  |
| LUOGO_STUD_LAV      | In the same municipality of residence                                                     | 1                  |
|                     | In another municipality in the same province                                              | 2                  |
|                     | In another province in the same region                                                    | 3                  |
|                     | In other regions                                                                          | 4                  |
|                     | Abroad                                                                                    | 5                  |
| REG_STUD_LAV        | Piedmont–Aosta Valley                                                                     | 1                  |
|                     | Lombardy                                                                                  | 3                  |
|                     | Trentino–South Tyrol                                                                      | 4                  |
|                     | Veneto                                                                                    | 5                  |
|                     | Friuli–Venezia Giulia                                                                     | 6                  |
|                     | Liguria                                                                                   | 7                  |
|                     | Emilia–Romagna                                                                            | 8                  |
|                     | Tuscany                                                                                   | 9                  |
|                     | Umbria                                                                                    | 10                 |
|                     | Marche                                                                                    | 11                 |
|                     | Lazio                                                                                     | 12                 |
|                     | Abruzzo                                                                                   | 13                 |
|                     | Molise                                                                                    | 14                 |
|                     | Campania                                                                                  | 15                 |
|                     | Apulia                                                                                    | 16                 |
|                     | Basilicata                                                                                | 17                 |
|                     | Calabria                                                                                  | 18                 |
|                     | Sicily                                                                                    | 19                 |
|                     | Sardinia                                                                                  | 20                 |
| STATO_STUD_LAV      | Italy                                                                                     | 1                  |
|                     | EU Europe                                                                                 | 2                  |
|                     | Non-EU Europe                                                                             | 3                  |
|                     | Rest of the world                                                                         | 4                  |
| MEZZO_TRAS_STUD_LAV | Train                                                                                     | 1                  |
|                     | Tram                                                                                      | 2                  |
|                     | Metro                                                                                     | 3                  |

| Variable             | Category                                                              | Corresponding code |
|----------------------|-----------------------------------------------------------------------|--------------------|
|                      | Urban bus or trolleybus                                               | 4                  |
|                      | Coach or intercity bus                                                | 5                  |
|                      | Company or school bus                                                 | 6                  |
|                      | Private car, as driver                                                | 7                  |
|                      | Private car, as passenger                                             | 8                  |
|                      | Motorcycle, moped, or scooter                                         | 9                  |
|                      | Bicycle                                                               | 10                 |
|                      | Other means of transport, such as boat or cable car                   | 11                 |
|                      | On foot                                                               | 12                 |
| DIM_5ANNO_PREC       | In this dwelling                                                      | 1                  |
|                      | In this municipality, but in another dwelling or collective residence | 2                  |
|                      | In another municipality                                               | 3                  |
|                      | Abroad                                                                | 4                  |
| PROV_DIM_5ANNO_PREC  | Same municipality of residence                                        | 1                  |
|                      | In another municipality in the same province                          | 2                  |
|                      | In another province in the same region                                | 3                  |
|                      | In another region                                                     | 4                  |
|                      | Abroad                                                                | 5                  |
| STATO_DIM_5ANNO_PREC | EU-27                                                                 | 11                 |
|                      | Central-Eastern Europe                                                | 12                 |
|                      | Other European countries                                              | 13                 |
|                      | Africa                                                                | 200                |
|                      | Asia                                                                  | 300                |
|                      | America                                                               | 400                |
|                      | Oceania                                                               | 500                |
| IND_MODELLO          | Collective residence form                                             | C                  |
|                      | Long form                                                             | L                  |

Table 24: Categories and codes for the categorical variables in the 2011 census 1% individual sample dataset

## A.5 Description of Variables in the Synthetic Population Dataset

| Variable               | Description                                                                     |
|------------------------|---------------------------------------------------------------------------------|
| synthetic_person_id    | Unique identifier assigned to each synthetic individual                         |
| synthetic_household_id | Unique identifier linking individuals to their synthetic household              |
| household_type         | Type of household (private or collective)                                       |
| collective_flag        | Indicator for collective population (1 = collective, 0 = private household)     |
| SEZ2011                | Census tract identifier (spatial unit of analysis)                              |
| CODREG                 | Regional code                                                                   |
| REGIONE                | Name of the region                                                              |
| CODPRO                 | Province code                                                                   |
| PROVINCIA              | Name of the province                                                            |
| CODCOM                 | Municipality code                                                               |
| COMUNE                 | Name of the municipality                                                        |
| PROCOM                 | Combined province–municipality code                                             |
| NSEZ                   | Census section number                                                           |
| ACE                    | Sub-municipal administrative area code                                          |
| CODLOC                 | Locality code                                                                   |
| CODASC                 | Census sub-area code                                                            |
| household_size_synth   | Size of the synthetic household (number of members)                             |
| member_no_in_synth_hh  | Position/order of the individual within the household                           |
| sex                    | Sex of the individual (male/female)                                             |
| age_group              | Age category of the individual                                                  |
| employment_binary      | Employment status (employed / not employed)                                     |
| education_group        | Educational attainment category                                                 |
| REL_PAR                | Relationship to household reference person (from donor data)                    |
| TIPOLOGIA_FAM          | Household typology (from donor data)                                            |
| STATO_CIV              | Marital status (from donor data)                                                |
| POS_PROF               | Professional status/position (from donor data)                                  |
| TIPCONV                | Type of residence (used to identify collective households)                      |
| MOPERC                 | Mode of enumeration (census collection method)                                  |
| donor_ID_FAM           | Original household identifier from 1% sample census microdata (donor reference) |
| donor_PROGR_IND        | Original individual identifier within household (donor reference)               |

| Variable             | Description                                                            |
|----------------------|------------------------------------------------------------------------|
| donor_ID_ALL         | Original household identifier in housing dataset (used for enrichment) |
| donor_household_type | Type of donor household (private or collective)                        |
| NROCOMPO_donor       | Number of members in the original donor household                      |

Table 25: Synthetic population attributes and their definition

## A.6 Description of Variables in the Synthetic Household Dataset

| Attribute              | Description                                                                                                           |
|------------------------|-----------------------------------------------------------------------------------------------------------------------|
| synthetic_household_id | Unique identifier assigned to each synthetic household                                                                |
| SEZ2011                | Census tract identifier where the synthetic household is located                                                      |
| household_size         | Number of individuals belonging to the synthetic household                                                            |
| AUTOMOB                | Number/category of vehicles available to the household, inherited from the donor household 1% sample census microdata |
| donor_ID_ALL           | Identifier of the donor household/dwelling used to assign household-level attributes                                  |

Table 26: Synthetic household attributes and their definition

## A.7 R Script for IPF Implementation and TRS Intergerisation

```
# =====
# FULL PIPELINE (ALL ZONES) - NEW VERSION
# IPF (mipfp) + 280 combos + TRS
#
# MICRODATA INPUT:
#   must contain at least:
#   sex, age_group, employment_binary, education_group
#
# CENSUS INPUT:
#   can contain extra columns, but must contain at least:
#   mage_*, fage_*, memp_*, femp_*, medu_*, fedu_*
#
# OUTPUTS:
# 1) ipf_output_allzones_contingency_280.csv
# 2) ipf_output_allzones_contingency_280_TRS.csv
# 3) ipf_output_allzones_grouped_like_constraints_TRS.csv
# =====

library(mipfp)

# -----
# STEP 1 - Paths + read data
# -----
cons_path <- "D:/PoliTo/Thesis/14042026/Final/01- R01_indicatori_2011_sezioni_ipf_
  controls.csv"
ind_path  <- "D:/PoliTo/Thesis/14042026/Final/02- microdata_seed_ipf.csv"
out_dir   <- "D:/PoliTo/Thesis/14042026/Final"

cons <- read.csv(cons_path, check.names = FALSE, stringsAsFactors = FALSE)
ind  <- read.csv(ind_path,  check.names = FALSE, stringsAsFactors = FALSE)

# -----
# STEP 2 - Zone ID
# Use SEZ2011 if available, otherwise make one
# -----
if ("SEZ2011" %in% names(cons)) {
  cons$ZoneID <- as.character(cons$SEZ2011)
} else {
  cons$ZoneID <- sprintf("Zone_%05d", seq_len(nrow(cons)))
}

rownames(cons) <- cons$ZoneID
```

```

zones <- cons$ZoneID
cat("Zones:", length(zones), "\n")

# -----
# STEP 3 - Define levels
# -----
sex_levels <- c("male", "female")

age_levels <- c(
  "age_0_14", "age_15_19", "age_20_24", "age_25_29",
  "age_30_34", "age_35_39", "age_40_44", "age_45_49",
  "age_50_54", "age_55_59", "age_60_64", "age_65_69",
  "age_70_74", "age_75p"
)

edu_levels <- c(
  "edu_illit_primary",
  "edu_lowersec",
  "edu_uppersec",
  "edu_tertiary",
  "edu_other_na"
)

emp_levels <- c("employed", "not_employed")

# -----
# STEP 4 - Required census columns
# -----
m_age_cols <- c(
  "mage_0_14", "mage_15_19", "mage_20_24", "mage_25_29",
  "mage_30_34", "mage_35_39", "mage_40_44", "mage_45_49",
  "mage_50_54", "mage_55_59", "mage_60_64", "mage_65_69",
  "mage_70_74", "mage_75p"
)

f_age_cols <- c(
  "fage_0_14", "fage_15_19", "fage_20_24", "fage_25_29",
  "fage_30_34", "fage_35_39", "fage_40_44", "fage_45_49",
  "fage_50_54", "fage_55_59", "fage_60_64", "fage_65_69",
  "fage_70_74", "fage_75p"
)

m_edu_cols <- c(
  "medu_illit_primary", "medu_lowersec", "medu_uppersec",
  "medu_tertiary", "medu_other_na"
)

```

```

f_edu_cols <- c(
  "fedu_illit_primary", "fedu_lowersec", "fedu_uppersec",
  "fedu_tertiary", "fedu_other_na"
)

m_emp_cols <- c("memp_employed", "memp_not_employed")
f_emp_cols <- c("femp_employed", "femp_not_employed")

required_cons_cols <- c(m_age_cols, f_age_cols, m_edu_cols, f_edu_cols, m_emp_cols, f_
  emp_cols)
missing_cons <- setdiff(required_cons_cols, names(cons))
if (length(missing_cons) > 0) {
  stop("Missing census columns: ", paste(missing_cons, collapse = ", "))
}

# Convert required census columns to numeric
for (cc in required_cons_cols) {
  cons[[cc]] <- as.numeric(cons[[cc]])
}

# -----
# STEP 5 - Prepare microdata
# -----
required_ind_cols <- c("sex", "age_group", "employment_binary", "education_group")
missing_ind <- setdiff(required_ind_cols, names(ind))
if (length(missing_ind) > 0) {
  stop("Missing microdata columns: ", paste(missing_ind, collapse = ", "))
}

ind$sex <- tolower(trimws(ind$sex))
ind$age_group <- tolower(trimws(ind$age_group))
ind$employment_binary <- tolower(trimws(ind$employment_binary))
ind$education_group <- tolower(trimws(ind$education_group))

ind$sex <- factor(ind$sex, levels = sex_levels)
ind$age_group <- factor(ind$age_group, levels = age_levels)
ind$education_group <- factor(ind$education_group, levels = edu_levels)
ind$employment_binary <- factor(ind$employment_binary, levels = emp_levels)

if (any(is.na(ind$sex)) ||
  any(is.na(ind$age_group)) ||
  any(is.na(ind$education_group)) ||
  any(is.na(ind$employment_binary))) {
  stop("Some microdata values do not match the expected levels.")
}

# -----

```

```

# STEP 6 - Build 1-zone seed (280 cells) + epsilon
# -----
seed_1zone <- table(
  ind$sex,
  ind$age_group,
  ind$education_group,
  ind$employment_binary
)

cat("Seed cells:", length(seed_1zone), "(should be 280)\n")

# avoid structural zeros
seed_1zone <- seed_1zone + 1e-6

# -----
# STEP 7 - Add Zone dimension
# Zone x Sex x Age x Edu x Emp
# -----
init_cells <- rep(seed_1zone, each = nrow(cons))

weight_init <- array(
  init_cells,
  dim = c(nrow(cons), dim(seed_1zone)),
  dimnames = list(
    Zone = zones,
    Sex  = dimnames(seed_1zone)[[1]],
    Age  = dimnames(seed_1zone)[[2]],
    Edu  = dimnames(seed_1zone)[[3]],
    Emp  = dimnames(seed_1zone)[[4]]
  )
)

# -----
# STEP 8 - Convert constraints to arrays
# con_age: Zone x Sex x Age
# con_edu: Zone x Sex x Edu
# con_emp: Zone x Sex x Emp
# -----

# Age
con_age <- array(
  0,
  dim = c(nrow(cons), length(sex_levels), length(age_levels)),
  dimnames = list(Zone = zones, Sex = sex_levels, Age = age_levels)
)

con_age[, "male", ] <- as.matrix(cons[, m_age_cols])

```

```

con_age[, "female", ] <- as.matrix(cons[, f_age_cols])

# Education
con_edu <- array(
  0,
  dim = c(nrow(cons), length(sex_levels), length(edu_levels)),
  dimnames = list(Zone = zones, Sex = sex_levels, Edu = edu_levels)
)

con_edu[, "male", ] <- as.matrix(cons[, m_edu_cols])
con_edu[, "female", ] <- as.matrix(cons[, f_edu_cols])

# Employment
con_emp <- array(
  0,
  dim = c(nrow(cons), length(sex_levels), length(emp_levels)),
  dimnames = list(Zone = zones, Sex = sex_levels, Emp = emp_levels)
)

con_emp[, "male", ] <- as.matrix(cons[, m_emp_cols])
con_emp[, "female", ] <- as.matrix(cons[, f_emp_cols])

# -----
# STEP 9 - Totals consistency check
# -----
tot_age <- apply(con_age, 1, sum)
tot_edu <- apply(con_edu, 1, sum)
tot_emp <- apply(con_emp, 1, sum)

cat("Totals checks:\n")
print(table(tot_edu == tot_age))
print(table(tot_emp == tot_age))
print(table(tot_emp == tot_edu))

# -----
# STEP 10 - target + descript
# weight order: Zone, Sex, Age, Edu, Emp
# -----
target <- list(con_age, con_edu, con_emp)
descript <- list(c(1, 2, 3), c(1, 2, 4), c(1, 2, 5))

# -----
# STEP 11 - Run IPF
# -----
w_mipfp_all <- Ipfp(
  seed = weight_init,
  target.list = descript,

```

```

    target.data = target,
    iter = 10000,
    tol = 1e-5
)

cat("Converged? ", w_mipfp_all$conv, "\n")

w_hat_all <- w_mipfp_all$x.hat

# -----
# STEP 12 - Margin checks
# -----
m_age <- apply(w_hat_all, c(1, 2, 3), sum)
m_edu <- apply(w_hat_all, c(1, 2, 4), sum)
m_emp <- apply(w_hat_all, c(1, 2, 5), sum)

cat("Max abs margin errors:\n")
print(c(
  max_abs_age = max(abs(m_age - con_age)),
  max_abs_edu = max(abs(m_edu - con_edu)),
  max_abs_emp = max(abs(m_emp - con_emp))
))

# -----
# STEP 13 - Export ALL zones contingency (280 columns)
# Column naming:
# male__age_30_34__edu_uppersec__employed
# -----
out_long <- as.data.frame.table(w_hat_all, responseName = "count")

out_long$Joint <- paste(
  out_long$Sex,
  out_long$Age,
  out_long$Edu,
  out_long$Emp,
  sep = "__"
)

out_long2 <- out_long[, c("Zone", "Joint", "count")]

out_wide_all <- reshape(out_long2, idvar = "Zone", timevar = "Joint", direction = "
  wide")
names(out_wide_all) <- sub("^count\\.", "", names(out_wide_all))
out_wide_all[is.na(out_wide_all)] <- 0

# attach geo columns if available
geo_cols <- c("ZoneID", "CODREG", "REGIONE", "CODPRO", "PROVINCIA",

```

```

      "CODCOM", "COMUNE", "PROCOM", "SEZ2011", "NSEZ", "ACE", "CODLOC", "
      CODASC")
geo_cols <- geo_cols[geo_cols %in% names(cons)]

geo_df <- cons[, geo_cols, drop = FALSE]
names(geo_df)[names(geo_df) == "ZoneID"] <- "Zone"

out_wide_all <- merge(geo_df, out_wide_all, by = "Zone", all.y = TRUE, sort = FALSE)

out_csv_280 <- file.path(out_dir, "03- ipf_output_allzones_contingency_280.csv")
write.csv(out_wide_all, out_csv_280, row.names = FALSE)
cat("Saved:\n", out_csv_280, "\n")
cat("Joint columns =", ncol(out_wide_all) - ncol(geo_df), "\n")

# -----
# STEP 14 - TRS integerisation
# -----
set.seed(123)

trs_row <- function(x, total) {
  xf <- floor(x)
  r <- total - sum(xf)

  if (r > 0) {
    frac <- x - xf
    s <- sum(frac)
    if (s <= 0) {
      idx <- sample(seq_along(x), size = r)
    } else {
      p <- frac / s
      idx <- sample(seq_along(x), size = r, prob = p, replace = FALSE)
    }
    xf[idx] <- xf[idx] + 1L
  }
  xf
}

# keep only contingency cells for integerisation
joint_cols <- setdiff(names(out_wide_all), names(geo_df))
X <- as.matrix(out_wide_all[, joint_cols, drop = FALSE])

row_totals <- as.integer(round(rowSums(X)))

X_int <- X
for (i in seq_len(nrow(X))) {
  X_int[i, ] <- trs_row(X[i, ], row_totals[i])
}

```

```

df_280_int <- data.frame(out_wide_all[, names(geo_df), drop = FALSE], X_int, check.
  names = FALSE)

out_csv_280_trs <- file.path(out_dir, "04- ipf_output_allzones_contingency_280_TRS.csv
  ")
write.csv(df_280_int, out_csv_280_trs, row.names = FALSE)
cat("Saved:\n", out_csv_280_trs, "\n")

# -----
# STEP 15 - Group TRS output like constraints
# -----
df <- df_280_int
joint_cols <- setdiff(names(df), names(geo_df))

parts <- strsplit(joint_cols, "__", fixed = TRUE)

Sex <- vapply(parts, '[', character(1), 1)
Age <- vapply(parts, '[', character(1), 2)
Edu <- vapply(parts, '[', character(1), 3)
Emp <- vapply(parts, '[', character(1), 4)

# (A) Age x Sex
age_cols_out <- c(m_age_cols, f_age_cols)
age_block <- matrix(0, nrow(df), length(age_cols_out), dimnames = list(NULL, age_cols_
  out))

for (j in seq_along(joint_cols)) {
  age_out <- if (Sex[j] == "male") paste0("m", Age[j]) else paste0("f", Age[j])
  age_block[, age_out] <- age_block[, age_out] + df[[joint_cols[j]]]
}

# (B) Edu x Sex
edu_cols_out <- c(m_edu_cols, f_edu_cols)
edu_block <- matrix(0, nrow(df), length(edu_cols_out), dimnames = list(NULL, edu_cols_
  out))

for (j in seq_along(joint_cols)) {
  edu_out <- if (Sex[j] == "male") paste0("m", Edu[j]) else paste0("f", Edu[j])
  edu_block[, edu_out] <- edu_block[, edu_out] + df[[joint_cols[j]]]
}

# (C) Emp x Sex
emp_cols_out <- c(m_emp_cols, f_emp_cols)
emp_block <- matrix(0, nrow(df), length(emp_cols_out), dimnames = list(NULL, emp_cols_
  out))

```

```

for (j in seq_along(joint_cols)) {
  emp_out <- if (Sex[j] == "male") paste0("memp-", Emp[j]) else paste0("femp-", Emp[
j])
  emp_block[, emp_out] <- emp_block[, emp_out] + df[[joint_cols[j]]]
}

# combine
out_grouped <- data.frame(
  df[, names(geo_df), drop = FALSE],
  age_block,
  edu_block,
  emp_block,
  check.names = FALSE
)

out_grouped_path <- file.path(out_dir, "05- ipf_output_allzones_grouped_like_
constraints_TRS.csv")
write.csv(out_grouped, out_grouped_path, row.names = FALSE)
cat("Saved:\n", out_grouped_path, "\n")

# -----
# STEP 16 - Internal consistency checks
# -----
totA <- rowSums(out_grouped[, age_cols_out, drop = FALSE])
totE <- rowSums(out_grouped[, edu_cols_out, drop = FALSE])
totP <- rowSums(out_grouped[, emp_cols_out, drop = FALSE])

cat("\nTotals equality checks (should be TRUE):\n")
print(all(totA == totE))
print(all(totA == totP))

```

## A.8 Python Script for Implementation of Bayesian Network for Driving License Ownership

```

import pandas as pd
import numpy as np
from pathlib import Path

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

from pgmpy.models import DiscreteBayesianNetwork
from pgmpy.estimators import BayesianEstimator
from pgmpy.inference import VariableElimination

```

```

# =====
# CONFIGURATION
# =====

INPUT_FILE_PATH = Path(r"PATH_TO_INPUT_FILE.xlsx")
OUTPUT_FILE_PATH = Path(r"PATH_TO_OUTPUT_FILE.xlsx")

SHEET_NAME = "INPUT_SHEET_NAME"

TARGET_COL = "target_variable"

FEATURE_COLS = [
    "feature_1",
    "feature_2",
    "feature_3",
    "feature_4",
]

VARIABLES = FEATURE_COLS + [TARGET_COL]

TEST_SIZE = 0.20
RANDOM_SEED = 42
BDEU_EQUIVALENT_SAMPLE_SIZE = 10

# =====
# READ DATA
# =====

df = pd.read_excel(INPUT_FILE_PATH, sheet_name=SHEET_NAME)

data = df[VARIABLES].copy()

# =====
# CLEAN MISSING VALUES
# =====

MISSING_CODES = [-1, -2, -6, -7, -8, -9, "", " ", "nan", "NaN"]

data = data.replace(MISSING_CODES, np.nan)
data = data.dropna(subset=VARIABLES).copy()

for col in VARIABLES:
    data[col] = data[col].astype(str)

```

```

# =====
# OPTIONAL VARIABLE RECODING
# =====

# Example:
# data["feature_4"] = pd.to_numeric(data["feature_4"], errors="coerce")
# data["feature_4"] = data["feature_4"].apply(
#     lambda x: "category_1" if x == 0
#     else "category_2" if x == 1
#     else "category_3" if pd.notna(x) and x >= 2
#     else np.nan
# )
#
# data = data.dropna(subset=VARIABLES).copy()
# for col in VARIABLES:
#     data[col] = data[col].astype(str)

# =====
# TRAIN / TEST SPLIT
# =====

train_data, test_data = train_test_split(
    data,
    test_size=TEST_SIZE,
    random_state=RANDOM_SEED,
    stratify=data[TARGET_COL]
)

# =====
# DEFINE BAYESIAN NETWORK STRUCTURE
# =====

model = DiscreteBayesianNetwork([
    ("feature_1", "feature_3"),
    ("feature_2", "feature_3"),

    ("feature_1", "feature_4"),
    ("feature_2", "feature_4"),
    ("feature_3", "feature_4"),

    ("feature_1", TARGET_COL),
    ("feature_2", TARGET_COL),
    ("feature_3", TARGET_COL),
    ("feature_4", TARGET_COL),
])

```

```

# =====
# PARAMETER LEARNING
# =====

estimator = BayesianEstimator(model, train_data)

cpds = estimator.get_parameters(
    prior_type="BDeu",
    equivalent_sample_size=BDEU_EQUIVALENT_SAMPLE_SIZE
)

model.add_cpds(*cpds)

print("Model valid:", model.check_model())

# =====
# INFERENCE ON TEST DATA
# =====

infer = VariableElimination(model)

predictions = []
predicted_probabilities = []

for _, row in test_data.iterrows():

    evidence = {
        feature: row[feature]
        for feature in FEATURE_COLS
    }

    posterior = infer.query(
        variables=[TARGET_COL],
        evidence=evidence,
        show_progress=False
    )

    states = posterior.state_names[TARGET_COL]
    probabilities = posterior.values

    best_index = np.argmax(probabilities)

    predictions.append(states[best_index])
    predicted_probabilities.append(probabilities[best_index])

```

```

# =====
# TEST RESULTS
# =====

test_results = test_data.copy()

test_results["observed_target"] = test_data[TARGET_COL].values
test_results["predicted_target"] = predictions
test_results["predicted_probability"] = predicted_probabilities
test_results["correct_prediction"] = (
    test_results["observed_target"].astype(str)
    == test_results["predicted_target"].astype(str)
)

overall_accuracy = accuracy_score(
    test_results["observed_target"],
    test_results["predicted_target"]
)

# =====
# CLASSIFICATION REPORT
# =====

classification_report_dict = classification_report(
    test_results["observed_target"],
    test_results["predicted_target"],
    output_dict=True,
    zero_division=0
)

classification_report_df = pd.DataFrame(
    classification_report_dict
).transpose()

# =====
# CONFUSION MATRIX
# =====

labels = sorted(test_results["observed_target"].astype(str).unique())

confusion_mat = confusion_matrix(
    test_results["observed_target"].astype(str),
    test_results["predicted_target"].astype(str),

```

```

        labels=labels
    )

    confusion_df = pd.DataFrame(
        confusion_mat,
        index=[f"Actual_{label}" for label in labels],
        columns=[f"Predicted_{label}" for label in labels]
    )

# =====
# MODEL SUMMARY
# =====

summary_df = pd.DataFrame({
    "metric": [
        "total_usable_records",
        "training_records",
        "testing_records",
        "training_percentage",
        "testing_percentage",
        "accuracy",
        "random_seed"
    ],
    "value": [
        len(data),
        len(train_data),
        len(test_data),
        f"{int((1 - TEST_SIZE) * 100)}%",
        f"{int(TEST_SIZE * 100)}%",
        overall_accuracy,
        RANDOM_SEED
    ]
})

# =====
# SAVE OUTPUTS
# =====

with pd.ExcelWriter(OUTPUT_FILE_PATH, engine="openpyxl") as writer:
    test_results.to_excel(
        writer,
        sheet_name="test_predictions",
        index=False
    )

```

```

train_data.to_excel(
    writer,
    sheet_name="training_data",
    index=False
)

test_data.to_excel(
    writer,
    sheet_name="testing_data",
    index=False
)

summary_df.to_excel(
    writer,
    sheet_name="model_summary",
    index=False,
    startrow=0
)

classification_report_df.to_excel(
    writer,
    sheet_name="model_summary",
    startrow=10
)

confusion_df.to_excel(
    writer,
    sheet_name="model_summary",
    startrow=25
)

print("Bayesian Network classification completed successfully.")
print("Overall accuracy:", overall_accuracy)
print("Results saved to:", OUTPUT_FILE_PATH)

```

## A.9 Python Script for Implementation of Bayesian Network for Activity Chain Generation

```

import os
import numpy as np
import pandas as pd

from pgmpy.estimators import HillClimbSearch, BDeu, MaximumLikelihoodEstimator

try:

```

```

    from pgmpy.models import DiscreteBayesianNetwork as BayesianNetwork
except ImportError:
    from pgmpy.models import BayesianNetwork

try:
    from pgmpy.estimators import ExpertKnowledge
    HAS_EXPERT_KNOWLEDGE = True
except ImportError:
    HAS_EXPERT_KNOWLEDGE = False

# =====
# CONFIGURATION
# =====

INPUT_FILE_PATH = r"PATH_TO_INPUT_FILE.xlsx"
OUTPUT_DIRECTORY = r"PATH_TO_OUTPUT_DIRECTORY"

ID_COL = "person_id"

DEMO_COLS = [
    "age_class",
    "sex",
    "education",
    "employment",
    "household size",
    "number of cars",
    "license",
    "pt_subscription",
]

ACTIVITY_COLS = [
    "Activity_1",
    "Activity_2",
    "Activity_3",
    "Activity_4",
    "Activity_5",
    "Activity_6",
]

RANDOM_SEED = 123
TEST_SIZE = 0.20

MAX_ITER = 10000
BDEU_EQUIVALENT_SAMPLE_SIZE = 10

```

```

# =====
# OUTPUT PATHS
# =====

os.makedirs(OUTPUT_DIRECTORY, exist_ok=True)

EDGES_OUTPUT_PATH = os.path.join(OUTPUT_DIRECTORY, "learned_bn_edges.csv")
SAMPLED_OUTPUT_PATH = os.path.join(OUTPUT_DIRECTORY, "sampled_activity_chains.csv")
EXCEL_OUTPUT_PATH = os.path.join(OUTPUT_DIRECTORY, "bayesian_network_outputs.xlsx")


# =====
# READ AND PREPARE DATA
# =====

df = pd.read_excel(INPUT_FILE_PATH)

required_cols = [ID_COL] + DEMO_COLS + ACTIVITY_COLS
missing_cols = [c for c in required_cols if c not in df.columns]

if missing_cols:
    raise ValueError(f"Missing columns in input file: {missing_cols}")

df = df[required_cols].copy()

for col in DEMO_COLS + ACTIVITY_COLS:
    df[col] = df[col].astype(str).str.strip()
    df[col] = df[col].replace({"nan": np.nan, "None": np.nan, "": np.nan})

df = df.dropna(subset=DEMO_COLS + ACTIVITY_COLS).copy()


# =====
# TRAIN / TEST SPLIT
# =====

np.random.seed(RANDOM_SEED)

test_idx = df.sample(frac=TEST_SIZE, random_state=RANDOM_SEED).index

test_df = df.loc[test_idx].copy()
train_df = df.drop(index=test_idx).copy()

train_model_df = train_df[DEMO_COLS + ACTIVITY_COLS].copy()


# =====

```

```

# DEFINE STRUCTURE CONSTRAINTS
# =====

all_nodes = DEMO_COLS + ACTIVITY_COLS
forbidden_edges = set()

# Demographic variables are treated as exogenous variables.
for target in DEMO_COLS:
    for source in all_nodes:
        if source != target:
            forbidden_edges.add((source, target))

# Later activities are not allowed to influence earlier activities.
for i, later_activity in enumerate(ACTIVITY_COLS):
    for earlier_activity in ACTIVITY_COLS[:i]:
        forbidden_edges.add((later_activity, earlier_activity))

# Self-loops are not allowed.
for node in all_nodes:
    forbidden_edges.add((node, node))

forbidden_edges = list(forbidden_edges)

# =====
# LEARN BAYESIAN NETWORK STRUCTURE
# =====

hc = HillClimbSearch(train_model_df)

scoring_method = BDeu(
    data=train_model_df,
    equivalent_sample_size=BDEU_EQUIVALENT_SAMPLE_SIZE
)

if HAS_EXPERT_KNOWLEDGE:
    expert_knowledge = ExpertKnowledge(forbidden_edges=forbidden_edges)
    learned_dag = hc.estimate(
        scoring_method=scoring_method,
        expert_knowledge=expert_knowledge,
        max_iter=MAX_ITER
    )
else:
    learned_dag = hc.estimate(
        scoring_method=scoring_method,
        black_list=forbidden_edges,
        max_iter=MAX_ITER
    )

```

```

    )

edges = list(learned_dag.edges())

model = BayesianNetwork(edges)
model.add_nodes_from(all_nodes)

edges_df = pd.DataFrame(edges, columns=["source", "target"])

# =====
# PARAMETER LEARNING
# =====

mle = MaximumLikelihoodEstimator(model, train_model_df)
cpds = mle.get_parameters(n_jobs=1)

model.add_cpds(*cpds)

print("Model valid:", model.check_model())

# =====
# SAMPLE ACTIVITY CHAINS
# =====

rng = np.random.default_rng(RANDOM_SEED)

def sample_from_cpd(model, node, assignment, rng):
    cpd = model.get_cpds(node)
    states = cpd.state_names[node]

    probabilities = []

    for state in states:
        query = {node: state}

        for parent in cpd.variables[1:]:
            query[parent] = assignment[parent]

        try:
            probability = cpd.get_value(**query)
        except Exception:
            probability = 0.0

        probabilities.append(float(probability))

```

```

probabilities = np.array(probabilities, dtype=float)

if probabilities.sum() <= 0:
    probabilities = np.ones(len(states)) / len(states)
else:
    probabilities = probabilities / probabilities.sum()

return rng.choice(states, p=probabilities)

def sample_activity_chain_for_person(model, row, rng):
    assignment = {}

    for col in DEMO_COLS:
        assignment[col] = row[col]

    for activity_col in ACTIVITY_COLS:
        assignment[activity_col] = sample_from_cpd(
            model=model,
            node=activity_col,
            assignment=assignment,
            rng=rng
        )

    return [assignment[c] for c in ACTIVITY_COLS]

sampled_rows = []

for _, row in test_df.iterrows():
    generated_chain = sample_activity_chain_for_person(model, row, rng)

    output_row = {ID_COL: row[ID_COL]}

    for col in DEMO_COLS:
        output_row[col] = row[col]

    for col in ACTIVITY_COLS:
        output_row[f"Observed_{col}"] = row[col]

    for col, value in zip(ACTIVITY_COLS, generated_chain):
        output_row[f"Generated_{col}"] = value

    sampled_rows.append(output_row)

sampled_df = pd.DataFrame(sampled_rows)

```

```

# =====
# VALIDATION
# =====

accuracy_rows = []

for col in ACTIVITY_COLS:
    observed = sampled_df[f"Observed_{col}"]
    generated = sampled_df[f"Generated_{col}"]

    accuracy_rows.append({
        "variable": col,
        "accuracy": (observed == generated).mean(),
        "n_test": len(sampled_df)
    })

accuracy_df = pd.DataFrame(accuracy_rows)

observed_chain = (
    sampled_df[[f"Observed_{c}" for c in ACTIVITY_COLS]]
    .astype(str)
    .agg(" > ".join, axis=1)
)

generated_chain = (
    sampled_df[[f"Generated_{c}" for c in ACTIVITY_COLS]]
    .astype(str)
    .agg(" > ".join, axis=1)
)

chain_accuracy_df = pd.DataFrame([
    {
        "metric": "full_chain_exact_match_accuracy",
        "value": (observed_chain == generated_chain).mean(),
        "n_test": len(sampled_df)
    }
])

# =====
# CHAIN FREQUENCY VALIDATION
# =====

observed_chain_freq = (
    observed_chain.value_counts(normalize=True)
    .mul(100)
    .rename("Observed (%)")
)

```

```

)

generated_chain_freq = (
    generated_chain.value_counts(normalize=True)
    .mul(100)
    .rename("Generated (%)")
)

chain_frequency_df = (
    pd.concat([observed_chain_freq, generated_chain_freq], axis=1)
    .fillna(0)
    .reset_index()
    .rename(columns={"index": "Chain"})
)

chain_frequency_df["Absolute Difference"] = (
    chain_frequency_df["Observed (%)"] -
    chain_frequency_df["Generated (%)"]
).abs()

chain_frequency_df = chain_frequency_df.sort_values(
    "Observed (%)",
    ascending=False
)

# =====
# POSITION-BASED FREQUENCY VALIDATION
# =====

position_tables = []

for col in ACTIVITY_COLS:
    observed_counts = sampled_df[f"Observed_{col}"].value_counts().rename("observed_count")
    generated_counts = sampled_df[f"Generated_{col}"].value_counts().rename("generated_count")

    table = pd.concat([observed_counts, generated_counts], axis=1).fillna(0).reset_index()
    table = table.rename(columns={"index": "activity"})

    table["activity_position"] = col
    table["observed_share"] = table["observed_count"] / table["observed_count"].sum()
    table["generated_share"] = table["generated_count"] / table["generated_count"].sum()
    table["absolute_share_difference"] = (

```

```

        table["observed_share"] - table["generated_share"]
    ).abs()

    position_tables.append(table)

position_frequency_df = pd.concat(position_tables, ignore_index=True)

# =====
# OVERALL ACTIVITY FREQUENCY VALIDATION
# =====

observed_all = pd.concat(
    [sampled_df[f"Observed_{c}"] for c in ACTIVITY_COLS],
    ignore_index=True
)

generated_all = pd.concat(
    [sampled_df[f"Generated_{c}"] for c in ACTIVITY_COLS],
    ignore_index=True
)

overall_observed = observed_all.value_counts().rename("observed_count")
overall_generated = generated_all.value_counts().rename("generated_count")

overall_frequency_df = (
    pd.concat([overall_observed, overall_generated], axis=1)
    .fillna(0)
    .reset_index()
    .rename(columns={"index": "activity"})
)

overall_frequency_df["observed_share"] = (
    overall_frequency_df["observed_count"] /
    overall_frequency_df["observed_count"].sum()
)

overall_frequency_df["generated_share"] = (
    overall_frequency_df["generated_count"] /
    overall_frequency_df["generated_count"].sum()
)

overall_frequency_df["absolute_share_difference"] = (
    overall_frequency_df["observed_share"] -
    overall_frequency_df["generated_share"]
).abs()

```

```

# =====
# TRANSITION VALIDATION
# =====

def build_transitions(data, prefix):
    rows = []

    for _, row in data.iterrows():
        chain = [row[f"{prefix}_{col}"] for col in ACTIVITY_COLS]

        for i in range(len(chain) - 1):
            rows.append({
                "from_activity": chain[i],
                "to_activity": chain[i + 1],
                "transition": f"{chain[i]} -> {chain[i + 1]}"
            })

    return pd.DataFrame(rows)

observed_transitions = build_transitions(sampled_df, "Observed")
generated_transitions = build_transitions(sampled_df, "Generated")

observed_transition_counts = (
    observed_transitions["transition"]
    .value_counts()
    .rename("observed_count")
)

generated_transition_counts = (
    generated_transitions["transition"]
    .value_counts()
    .rename("generated_count")
)

transition_df = (
    pd.concat([observed_transition_counts, generated_transition_counts], axis=1)
    .fillna(0)
    .reset_index()
    .rename(columns={"index": "transition"})
)

transition_df["observed_share"] = (
    transition_df["observed_count"] /
    transition_df["observed_count"].sum()
)

```

```

transition_df["generated_share"] = (
    transition_df["generated_count"] /
    transition_df["generated_count"].sum()
)

transition_df["absolute_share_difference"] = (
    transition_df["observed_share"] -
    transition_df["generated_share"]
).abs()

# =====
# MODEL SUMMARY
# =====

model_summary_df = pd.DataFrame([
    {"metric": "number_of_training_records", "value": len(train_df)},
    {"metric": "number_of_test_records", "value": len(sampled_df)},
    {"metric": "number_of_nodes", "value": len(all_nodes)},
    {"metric": "number_of_edges", "value": len(edges)},
    {"metric": "mean_activity_position_accuracy", "value": accuracy_df["accuracy"].
    mean()},
    {"metric": "full_chain_exact_match_accuracy", "value": chain_accuracy_df.loc[0, "
    value"]},
    {"metric": "mean_absolute_chain_share_difference", "value": chain_frequency_df["
    Absolute Difference"].mean()},
    {"metric": "mean_absolute_share_difference_by_position", "value":
    position_frequency_df["absolute_share_difference"].mean()},
    {"metric": "mean_absolute_share_difference_overall", "value": overall_frequency_df
    ["absolute_share_difference"].mean()},
    {"metric": "mean_absolute_share_difference_transitions", "value": transition_df["
    absolute_share_difference"].mean()},
])

# =====
# SAVE OUTPUTS
# =====

edges_df.to_csv(EDGES_OUTPUT_PATH, index=False)
sampled_df.to_csv(SAMPLED_OUTPUT_PATH, index=False)

with pd.ExcelWriter(EXCEL_OUTPUT_PATH, engine="openpyxl") as writer:
    edges_df.to_excel(writer, sheet_name="learned_edges", index=False)
    sampled_df.to_excel(writer, sheet_name="sampled_activity_chains", index=False)
    accuracy_df.to_excel(writer, sheet_name="activity_accuracy", index=False)

```

```

chain_accuracy_df.to_excel(writer, sheet_name="chain_accuracy", index=False)
chain_frequency_df.to_excel(writer, sheet_name="chain_frequencies", index=False)
position_frequency_df.to_excel(writer, sheet_name="position_frequencies", index=False)
overall_frequency_df.to_excel(writer, sheet_name="overall_frequencies", index=False)
transition_df.to_excel(writer, sheet_name="transition_validation", index=False)
model_summary_df.to_excel(writer, sheet_name="model_summary", index=False)

print("Bayesian Network modelling completed successfully.")
print("Learned edges saved to:", EDGES_OUTPUT_PATH)
print("Sampled activity chains saved to:", SAMPLED_OUTPUT_PATH)
print("Model outputs saved to:", EXCEL_OUTPUT_PATH)

```

## A.10 Distribution of the twenty most frequent activity chains by household group

| Children | Employment         | Household Size | Rank | Activity chain              | Share (%) |
|----------|--------------------|----------------|------|-----------------------------|-----------|
| With     | No employed adults | 2              | 1    | Home                        | 19.12     |
| With     | No employed adults | 2              | 2    | HomeShoppingHome            | 15.37     |
| With     | No employed adults | 2              | 3    | HomeEscortHome              | 15.18     |
| With     | No employed adults | 2              | 4    | HomeLeisureHome             | 14.29     |
| With     | No employed adults | 2              | 5    | HomeMaintenanceHome         | 5.27      |
| With     | No employed adults | 2              | 6    | HomeEducationHome           | 4.96      |
| With     | No employed adults | 2              | 7    | HomeWorkHome                | 1.4       |
| With     | No employed adults | 2              | 8    | HomeEscortShoppingHome      | 1.33      |
| With     | No employed adults | 2              | 9    | HomeLeisureShoppingHome     | 1.21      |
| With     | No employed adults | 2              | 10   | HomeShoppingHomeLeisureHome | 1.21      |
| With     | No employed adults | 2              | 11   | HomeMaintenanceShoppingHome | 1.08      |
| With     | No employed adults | 2              | 12   | HomeLeisureHomeLeisureHome  | 1.02      |

| Children | Employment         | Household Size | Rank | Activity chain              | Share (%) |
|----------|--------------------|----------------|------|-----------------------------|-----------|
| With     | No employed adults | 2              | 13   | HomeLeisureMaintenanceHome  | 1.02      |
| With     | No employed adults | 2              | 14   | HomeShoppingLeisureHome     | 1.02      |
| With     | No employed adults | 2              | 15   | HomeEscortHomeLeisureHome   | 0.89      |
| With     | No employed adults | 2              | 16   | HomeMaintenanceEscortHome   | 0.89      |
| With     | No employed adults | 2              | 17   | LeisureHomeLeisure          | 0.83      |
| With     | No employed adults | 2              | 18   | HomeShoppingEscortHome      | 0.76      |
| With     | No employed adults | 2              | 19   | HomeEscortHomeShoppingHome  | 0.7       |
| With     | No employed adults | 2              | 20   | HomeMaintenanceLeisureHome  | 0.7       |
| With     | No employed adults | 34             | 1    | Home                        | 18.11     |
| With     | No employed adults | 34             | 2    | HomeShoppingHome            | 15.63     |
| With     | No employed adults | 34             | 3    | HomeEscortHome              | 14.75     |
| With     | No employed adults | 34             | 4    | HomeLeisureHome             | 14.18     |
| With     | No employed adults | 34             | 5    | HomeMaintenanceHome         | 6.82      |
| With     | No employed adults | 34             | 6    | HomeEducationHome           | 4.18      |
| With     | No employed adults | 34             | 7    | HomeLeisureHomeLeisureHome  | 1.22      |
| With     | No employed adults | 34             | 8    | HomeEscortShoppingHome      | 1.14      |
| With     | No employed adults | 34             | 9    | HomeShoppingHomeLeisureHome | 1.14      |
| With     | No employed adults | 34             | 10   | LeisureHomeLeisure          | 1.14      |
| With     | No employed adults | 34             | 11   | HomeEscortHomeLeisureHome   | 1.12      |
| With     | No employed adults | 34             | 12   | HomeWorkHome                | 1.09      |
| With     | No employed adults | 34             | 13   | HomeMaintenanceShoppingHome | 0.95      |
| With     | No employed adults | 34             | 14   | HomeShoppingLeisureHome     | 0.87      |

| Children | Employment         | Household Size | Rank | Activity chain               | Share (%) |
|----------|--------------------|----------------|------|------------------------------|-----------|
| With     | No employed adults | 34             | 15   | HomeLeisureShoppingHome      | 0.85      |
| With     | No employed adults | 34             | 16   | HomeShoppingEscortHome       | 0.72      |
| With     | No employed adults | 34             | 17   | HomeEscortHomeShoppingHome   | 0.57      |
| With     | No employed adults | 34             | 18   | HomeMaintenanceLeisureHome   | 0.52      |
| With     | No employed adults | 34             | 19   | HomeShoppingMaintenanceHome  | 0.5       |
| With     | No employed adults | 34             | 20   | HomeEducationHomeLeisureHome | 0.47      |
| With     | No employed adults | 5+             | 1    | Home                         | 18.37     |
| With     | No employed adults | 5+             | 2    | HomeShoppingHome             | 16.11     |
| With     | No employed adults | 5+             | 3    | HomeEscortHome               | 14.3      |
| With     | No employed adults | 5+             | 4    | HomeLeisureHome              | 14.19     |
| With     | No employed adults | 5+             | 5    | HomeMaintenanceHome          | 6.78      |
| With     | No employed adults | 5+             | 6    | HomeEducationHome            | 4.3       |
| With     | No employed adults | 5+             | 7    | HomeEscortShoppingHome       | 1.53      |
| With     | No employed adults | 5+             | 8    | HomeLeisureShoppingHome      | 1.36      |
| With     | No employed adults | 5+             | 9    | LeisureHomeLeisure           | 1.36      |
| With     | No employed adults | 5+             | 10   | HomeWorkHome                 | 1.24      |
| With     | No employed adults | 5+             | 11   | HomeShoppingHomeLeisureHome  | 1.19      |
| With     | No employed adults | 5+             | 12   | HomeEscortHomeLeisureHome    | 0.9       |
| With     | No employed adults | 5+             | 13   | HomeLeisureHomeLeisureHome   | 0.79      |
| With     | No employed adults | 5+             | 14   | HomeMaintenanceShoppingHome  | 0.79      |
| With     | No employed adults | 5+             | 15   | HomeShoppingEscortHome       | 0.68      |
| With     | No employed adults | 5+             | 16   | HomeShoppingLeisureHome      | 0.68      |

| Children | Employment              | Household Size | Rank | Activity chain                 | Share (%) |
|----------|-------------------------|----------------|------|--------------------------------|-----------|
| With     | No employed adults      | 5+             | 17   | HomeLeisureMaintenanceHome     | 0.62      |
| With     | No employed adults      | 5+             | 18   | HomeShoppingHomeShoppingHome   | 0.62      |
| With     | No employed adults      | 5+             | 19   | HomeMaintenanceHomeLeisureHome | 0.57      |
| With     | No employed adults      | 5+             | 20   | HomeEscortLeisureHome          | 0.51      |
| With     | $\geq 1$ employed adult | 2              | 1    | HomeWorkHome                   | 39.81     |
| With     | $\geq 1$ employed adult | 2              | 2    | Home                           | 8.58      |
| With     | $\geq 1$ employed adult | 2              | 3    | HomeLeisureHome                | 5.45      |
| With     | $\geq 1$ employed adult | 2              | 4    | HomeShoppingHome               | 4.26      |
| With     | $\geq 1$ employed adult | 2              | 5    | HomeWorkShoppingHome           | 4.19      |
| With     | $\geq 1$ employed adult | 2              | 6    | HomeWorkHomeLeisureHome        | 3.07      |
| With     | $\geq 1$ employed adult | 2              | 7    | HomeEscortHome                 | 3.01      |
| With     | $\geq 1$ employed adult | 2              | 8    | HomeMaintenanceHome            | 2.2       |
| With     | $\geq 1$ employed adult | 2              | 9    | HomeWorkLeisureHome            | 1.99      |
| With     | $\geq 1$ employed adult | 2              | 10   | Work                           | 1.68      |
| With     | $\geq 1$ employed adult | 2              | 11   | HomeMaintenanceWorkHome        | 1.64      |
| With     | $\geq 1$ employed adult | 2              | 12   | HomeWorkMaintenanceHome        | 1.6       |
| With     | $\geq 1$ employed adult | 2              | 13   | HomeWorkEscortHome             | 1.43      |
| With     | $\geq 1$ employed adult | 2              | 14   | HomeWorkHomeWorkHome           | 1.07      |
| With     | $\geq 1$ employed adult | 2              | 15   | HomeWorkHomeShoppingHome       | 0.96      |
| With     | $\geq 1$ employed adult | 2              | 16   | HomeWorkHomeMaintenanceHome    | 0.68      |
| With     | $\geq 1$ employed adult | 2              | 17   | HomeShoppingLeisureHome        | 0.66      |
| With     | $\geq 1$ employed adult | 2              | 18   | HomeWorkHomeEscortHome         | 0.66      |
| With     | $\geq 1$ employed adult | 2              | 19   | HomeMaintenanceLeisureHome     | 0.64      |
| With     | $\geq 1$ employed adult | 2              | 20   | WorkHomeWork                   | 0.64      |
| With     | $\geq 1$ employed adult | 34             | 1    | HomeWorkHome                   | 32.39     |
| With     | $\geq 1$ employed adult | 34             | 2    | Home                           | 10.09     |
| With     | $\geq 1$ employed adult | 34             | 3    | HomeLeisureHome                | 7.17      |
| With     | $\geq 1$ employed adult | 34             | 4    | HomeShoppingHome               | 6.12      |
| With     | $\geq 1$ employed adult | 34             | 5    | HomeEscortHome                 | 5.26      |
| With     | $\geq 1$ employed adult | 34             | 6    | HomeWorkShoppingHome           | 3.2       |
| With     | $\geq 1$ employed adult | 34             | 7    | HomeMaintenanceHome            | 3.2       |
| With     | $\geq 1$ employed adult | 34             | 8    | HomeWorkHomeLeisureHome        | 2.54      |
| With     | $\geq 1$ employed adult | 34             | 9    | HomeWorkLeisureHome            | 1.73      |
| With     | $\geq 1$ employed adult | 34             | 10   | Work                           | 1.64      |
| With     | $\geq 1$ employed adult | 34             | 11   | HomeWorkMaintenanceHome        | 1.47      |
| With     | $\geq 1$ employed adult | 34             | 12   | HomeWorkEscortHome             | 1.39      |
| With     | $\geq 1$ employed adult | 34             | 13   | HomeMaintenanceWorkHome        | 1.37      |
| With     | $\geq 1$ employed adult | 34             | 14   | HomeEducationHome              | 0.95      |
| With     | $\geq 1$ employed adult | 34             | 15   | HomeShoppingLeisureHome        | 0.75      |
| With     | $\geq 1$ employed adult | 34             | 16   | HomeWorkHomeShoppingHome       | 0.75      |

| Children | Employment              | Household Size | Rank | Activity chain              | Share (%) |
|----------|-------------------------|----------------|------|-----------------------------|-----------|
| With     | $\geq 1$ employed adult | 34             | 17   | HomeWorkHomeWorkHome        | 0.73      |
| With     | $\geq 1$ employed adult | 34             | 18   | HomeEscortWorkHome          | 0.68      |
| With     | $\geq 1$ employed adult | 34             | 19   | HomeLeisureHomeLeisureHome  | 0.61      |
| With     | $\geq 1$ employed adult | 34             | 20   | HomeWorkHomeMaintenanceHome | 0.59      |
| With     | $\geq 1$ employed adult | 5+             | 1    | HomeWorkHome                | 23.56     |
| With     | $\geq 1$ employed adult | 5+             | 2    | Home                        | 14.14     |
| With     | $\geq 1$ employed adult | 5+             | 3    | HomeLeisureHome             | 8.82      |
| With     | $\geq 1$ employed adult | 5+             | 4    | HomeShoppingHome            | 8.29      |
| With     | $\geq 1$ employed adult | 5+             | 5    | HomeEscortHome              | 7.13      |
| With     | $\geq 1$ employed adult | 5+             | 6    | HomeMaintenanceHome         | 4.14      |
| With     | $\geq 1$ employed adult | 5+             | 7    | HomeWorkShoppingHome        | 2.5       |
| With     | $\geq 1$ employed adult | 5+             | 8    | HomeEducationHome           | 1.76      |
| With     | $\geq 1$ employed adult | 5+             | 9    | HomeWorkHomeLeisureHome     | 1.76      |
| With     | $\geq 1$ employed adult | 5+             | 10   | HomeWorkLeisureHome         | 1.37      |
| With     | $\geq 1$ employed adult | 5+             | 11   | Work                        | 1.26      |
| With     | $\geq 1$ employed adult | 5+             | 12   | HomeWorkEscortHome          | 1.14      |
| With     | $\geq 1$ employed adult | 5+             | 13   | HomeMaintenanceWorkHome     | 1.09      |
| With     | $\geq 1$ employed adult | 5+             | 14   | HomeWorkMaintenanceHome     | 1.06      |
| With     | $\geq 1$ employed adult | 5+             | 15   | HomeShoppingLeisureHome     | 0.73      |
| With     | $\geq 1$ employed adult | 5+             | 16   | HomeLeisureHomeLeisureHome  | 0.71      |
| With     | $\geq 1$ employed adult | 5+             | 17   | HomeEscortShoppingHome      | 0.71      |
| With     | $\geq 1$ employed adult | 5+             | 18   | HomeLeisureShoppingHome     | 0.65      |
| With     | $\geq 1$ employed adult | 5+             | 19   | HomeShoppingHomeLeisureHome | 0.62      |
| With     | $\geq 1$ employed adult | 5+             | 20   | HomeWorkHomeShoppingHome    | 0.62      |
| Without  | No employed adults      | 1              | 1    | Home                        | 20.62     |
| Without  | No employed adults      | 1              | 2    | HomeShoppingHome            | 14.97     |
| Without  | No employed adults      | 1              | 3    | HomeLeisureHome             | 14.24     |
| Without  | No employed adults      | 1              | 4    | HomeEscortHome              | 13.8      |
| Without  | No employed adults      | 1              | 5    | HomeMaintenanceHome         | 6.63      |
| Without  | No employed adults      | 1              | 6    | HomeEducationHome           | 4.38      |
| Without  | No employed adults      | 1              | 7    | HomeEscortShoppingHome      | 1.42      |
| Without  | No employed adults      | 1              | 8    | HomeShoppingHomeLeisureHome | 1.12      |
| Without  | No employed adults      | 1              | 9    | HomeLeisureHomeLeisureHome  | 1.08      |
| Without  | No employed adults      | 1              | 10   | HomeEscortHomeLeisureHome   | 1.01      |

| Children | Employment         | Household Size | Rank | Activity chain                 | Share (%) |
|----------|--------------------|----------------|------|--------------------------------|-----------|
| Without  | No employed adults | 1              | 11   | HomeLeisureShoppingHome        | 0.9       |
| Without  | No employed adults | 1              | 12   | HomeWorkHome                   | 0.89      |
| Without  | No employed adults | 1              | 13   | LeisureHomeLeisure             | 0.85      |
| Without  | No employed adults | 1              | 14   | HomeMaintenanceShoppingHome    | 0.85      |
| Without  | No employed adults | 1              | 15   | HomeShoppingLeisureHome        | 0.76      |
| Without  | No employed adults | 1              | 16   | HomeLeisureMaintenanceHome     | 0.71      |
| Without  | No employed adults | 1              | 17   | HomeShoppingEscortHome         | 0.65      |
| Without  | No employed adults | 1              | 18   | HomeMaintenanceEscortHome      | 0.64      |
| Without  | No employed adults | 1              | 19   | HomeMaintenanceLeisureHome     | 0.59      |
| Without  | No employed adults | 1              | 20   | HomeMaintenanceHomeLeisureHome | 0.46      |
| Without  | No employed adults | 2              | 1    | Home                           | 22.91     |
| Without  | No employed adults | 2              | 2    | HomeShoppingHome               | 15.05     |
| Without  | No employed adults | 2              | 3    | HomeLeisureHome                | 13.83     |
| Without  | No employed adults | 2              | 4    | HomeEscortHome                 | 12.96     |
| Without  | No employed adults | 2              | 5    | HomeMaintenanceHome            | 6.46      |
| Without  | No employed adults | 2              | 6    | HomeEducationHome              | 3.95      |
| Without  | No employed adults | 2              | 7    | HomeEscortShoppingHome         | 1.31      |
| Without  | No employed adults | 2              | 8    | HomeShoppingHomeLeisureHome    | 1.14      |
| Without  | No employed adults | 2              | 9    | HomeEscortHomeLeisureHome      | 1.01      |
| Without  | No employed adults | 2              | 10   | HomeLeisureHomeLeisureHome     | 1         |
| Without  | No employed adults | 2              | 11   | HomeLeisureShoppingHome        | 0.99      |
| Without  | No employed adults | 2              | 12   | HomeWorkHome                   | 0.86      |

| Children | Employment         | Household Size | Rank | Activity chain                 | Share (%) |
|----------|--------------------|----------------|------|--------------------------------|-----------|
| Without  | No employed adults | 2              | 13   | HomeMaintenanceShoppingHome    | 0.84      |
| Without  | No employed adults | 2              | 14   | LeisureHomeLeisure             | 0.79      |
| Without  | No employed adults | 2              | 15   | HomeShoppingLeisureHome        | 0.77      |
| Without  | No employed adults | 2              | 16   | HomeLeisureMaintenanceHome     | 0.73      |
| Without  | No employed adults | 2              | 17   | HomeShoppingEscortHome         | 0.71      |
| Without  | No employed adults | 2              | 18   | HomeMaintenanceEscortHome      | 0.65      |
| Without  | No employed adults | 2              | 19   | HomeMaintenanceLeisureHome     | 0.56      |
| Without  | No employed adults | 2              | 20   | HomeMaintenanceHomeLeisureHome | 0.49      |
| Without  | No employed adults | 34             | 1    | Home                           | 20.82     |
| Without  | No employed adults | 34             | 2    | HomeShoppingHome               | 15.29     |
| Without  | No employed adults | 34             | 3    | HomeLeisureHome                | 13.89     |
| Without  | No employed adults | 34             | 4    | HomeEscortHome                 | 13.88     |
| Without  | No employed adults | 34             | 5    | HomeMaintenanceHome            | 6.34      |
| Without  | No employed adults | 34             | 6    | HomeEducationHome              | 4.38      |
| Without  | No employed adults | 34             | 7    | HomeEscortShoppingHome         | 1.44      |
| Without  | No employed adults | 34             | 8    | HomeLeisureHomeLeisureHome     | 1.14      |
| Without  | No employed adults | 34             | 9    | HomeShoppingHomeLeisureHome    | 1.08      |
| Without  | No employed adults | 34             | 10   | HomeEscortHomeLeisureHome      | 1         |
| Without  | No employed adults | 34             | 11   | HomeLeisureShoppingHome        | 0.92      |
| Without  | No employed adults | 34             | 12   | LeisureHomeLeisure             | 0.91      |
| Without  | No employed adults | 34             | 13   | HomeWorkHome                   | 0.88      |
| Without  | No employed adults | 34             | 14   | HomeShoppingLeisureHome        | 0.83      |

| Children | Employment         | Household Size | Rank | Activity chain                 | Share (%) |
|----------|--------------------|----------------|------|--------------------------------|-----------|
| Without  | No employed adults | 34             | 15   | HomeLeisureMaintenanceHome     | 0.8       |
| Without  | No employed adults | 34             | 16   | HomeMaintenanceShoppingHome    | 0.74      |
| Without  | No employed adults | 34             | 17   | HomeMaintenanceLeisureHome     | 0.63      |
| Without  | No employed adults | 34             | 18   | HomeShoppingEscortHome         | 0.63      |
| Without  | No employed adults | 34             | 19   | HomeMaintenanceEscortHome      | 0.62      |
| Without  | No employed adults | 34             | 20   | HomeMaintenanceHomeLeisureHome | 0.46      |
| Without  | No employed adults | 5+             | 1    | Home                           | 34.9      |
| Without  | No employed adults | 5+             | 2    | HomeLeisureHome                | 13.73     |
| Without  | No employed adults | 5+             | 3    | HomeShoppingHome               | 11.37     |
| Without  | No employed adults | 5+             | 4    | HomeEscortHome                 | 10.2      |
| Without  | No employed adults | 5+             | 5    | HomeMaintenanceHome            | 6.67      |
| Without  | No employed adults | 5+             | 6    | HomeShoppingHomeLeisureHome    | 2.75      |
| Without  | No employed adults | 5+             | 7    | HomeLeisureShoppingHome        | 1.57      |
| Without  | No employed adults | 5+             | 8    | HomeEducationHome              | 1.18      |
| Without  | No employed adults | 5+             | 9    | HomeMaintenanceEscortHome      | 1.18      |
| Without  | No employed adults | 5+             | 10   | HomeShoppingEscortHome         | 1.18      |
| Without  | No employed adults | 5+             | 11   | HomeEscortLeisureHome          | 0.78      |
| Without  | No employed adults | 5+             | 12   | HomeLeisureHomeLeisureHome     | 0.78      |
| Without  | No employed adults | 5+             | 13   | HomeLeisureMaintenanceHome     | 0.78      |
| Without  | No employed adults | 5+             | 14   | HomeMaintenanceHomeLeisureHome | 0.78      |
| Without  | No employed adults | 5+             | 15   | HomeMaintenanceLeisureHome     | 0.78      |
| Without  | No employed adults | 5+             | 16   | HomeShoppingLeisureHome        | 0.78      |

| Children | Employment              | Household Size | Rank | Activity chain              | Share (%) |
|----------|-------------------------|----------------|------|-----------------------------|-----------|
| Without  | No employed adults      | 5+             | 17   | HomeShoppingMaintenanceHome | 0.78      |
| Without  | No employed adults      | 5+             | 18   | EscortMaintenanceHomeEscort | 0.39      |
| Without  | No employed adults      | 5+             | 19   | HomeEducationShoppingHome   | 0.39      |
| Without  | No employed adults      | 5+             | 20   | HomeEscortHomeEscortHome    | 0.39      |
| Without  | $\geq 1$ employed adult | 1              | 1    | HomeWorkHome                | 40.29     |
| Without  | $\geq 1$ employed adult | 1              | 2    | Home                        | 8.27      |
| Without  | $\geq 1$ employed adult | 1              | 3    | HomeLeisureHome             | 5.31      |
| Without  | $\geq 1$ employed adult | 1              | 4    | HomeWorkShoppingHome        | 3.87      |
| Without  | $\geq 1$ employed adult | 1              | 5    | HomeShoppingHome            | 3.62      |
| Without  | $\geq 1$ employed adult | 1              | 6    | HomeWorkHomeLeisureHome     | 3.06      |
| Without  | $\geq 1$ employed adult | 1              | 7    | HomeEscortHome              | 2.99      |
| Without  | $\geq 1$ employed adult | 1              | 8    | HomeMaintenanceHome         | 2.33      |
| Without  | $\geq 1$ employed adult | 1              | 9    | HomeWorkLeisureHome         | 2.2       |
| Without  | $\geq 1$ employed adult | 1              | 10   | Work                        | 1.98      |
| Without  | $\geq 1$ employed adult | 1              | 11   | HomeWorkEscortHome          | 1.81      |
| Without  | $\geq 1$ employed adult | 1              | 12   | HomeWorkMaintenanceHome     | 1.78      |
| Without  | $\geq 1$ employed adult | 1              | 13   | HomeMaintenanceWorkHome     | 1.73      |
| Without  | $\geq 1$ employed adult | 1              | 14   | HomeWorkHomeShoppingHome    | 1.01      |
| Without  | $\geq 1$ employed adult | 1              | 15   | HomeWorkHomeWorkHome        | 0.92      |
| Without  | $\geq 1$ employed adult | 1              | 16   | HomeEscortWorkHome          | 0.83      |
| Without  | $\geq 1$ employed adult | 1              | 17   | HomeWorkHomeEscortHome      | 0.73      |
| Without  | $\geq 1$ employed adult | 1              | 18   | HomeWorkHomeMaintenanceHome | 0.69      |
| Without  | $\geq 1$ employed adult | 1              | 19   | HomeShoppingLeisureHome     | 0.67      |
| Without  | $\geq 1$ employed adult | 1              | 20   | HomeMaintenanceLeisureHome  | 0.56      |
| Without  | $\geq 1$ employed adult | 2              | 1    | HomeWorkHome                | 29.4      |
| Without  | $\geq 1$ employed adult | 2              | 2    | Home                        | 12.65     |
| Without  | $\geq 1$ employed adult | 2              | 3    | HomeLeisureHome             | 7.72      |
| Without  | $\geq 1$ employed adult | 2              | 4    | HomeShoppingHome            | 6.68      |
| Without  | $\geq 1$ employed adult | 2              | 5    | HomeEscortHome              | 5.63      |
| Without  | $\geq 1$ employed adult | 2              | 6    | HomeMaintenanceHome         | 3.55      |
| Without  | $\geq 1$ employed adult | 2              | 7    | HomeWorkShoppingHome        | 2.83      |
| Without  | $\geq 1$ employed adult | 2              | 8    | HomeWorkHomeLeisureHome     | 2.16      |
| Without  | $\geq 1$ employed adult | 2              | 9    | Work                        | 1.57      |
| Without  | $\geq 1$ employed adult | 2              | 10   | HomeWorkLeisureHome         | 1.56      |
| Without  | $\geq 1$ employed adult | 2              | 11   | HomeWorkMaintenanceHome     | 1.34      |
| Without  | $\geq 1$ employed adult | 2              | 12   | HomeWorkEscortHome          | 1.32      |
| Without  | $\geq 1$ employed adult | 2              | 13   | HomeMaintenanceWorkHome     | 1.22      |
| Without  | $\geq 1$ employed adult | 2              | 14   | HomeEducationHome           | 1.12      |
| Without  | $\geq 1$ employed adult | 2              | 15   | HomeShoppingLeisureHome     | 0.72      |
| Without  | $\geq 1$ employed adult | 2              | 16   | HomeWorkHomeShoppingHome    | 0.71      |

| Children | Employment              | Household Size | Rank | Activity chain              | Share (%) |
|----------|-------------------------|----------------|------|-----------------------------|-----------|
| Without  | $\geq 1$ employed adult | 2              | 17   | HomeWorkHomeWorkHome        | 0.65      |
| Without  | $\geq 1$ employed adult | 2              | 18   | HomeLeisureHomeLeisureHome  | 0.64      |
| Without  | $\geq 1$ employed adult | 2              | 19   | HomeLeisureShoppingHome     | 0.59      |
| Without  | $\geq 1$ employed adult | 2              | 20   | HomeMaintenanceLeisureHome  | 0.58      |
| Without  | $\geq 1$ employed adult | 34             | 1    | HomeWorkHome                | 21.22     |
| Without  | $\geq 1$ employed adult | 34             | 2    | Home                        | 13.95     |
| Without  | $\geq 1$ employed adult | 34             | 3    | HomeLeisureHome             | 9.54      |
| Without  | $\geq 1$ employed adult | 34             | 4    | HomeShoppingHome            | 9.12      |
| Without  | $\geq 1$ employed adult | 34             | 5    | HomeEscortHome              | 8.17      |
| Without  | $\geq 1$ employed adult | 34             | 6    | HomeMaintenanceHome         | 4.4       |
| Without  | $\geq 1$ employed adult | 34             | 7    | HomeEducationHome           | 2.18      |
| Without  | $\geq 1$ employed adult | 34             | 8    | HomeWorkShoppingHome        | 2.05      |
| Without  | $\geq 1$ employed adult | 34             | 9    | HomeWorkHomeLeisureHome     | 1.64      |
| Without  | $\geq 1$ employed adult | 34             | 10   | Work                        | 1.14      |
| Without  | $\geq 1$ employed adult | 34             | 11   | HomeWorkLeisureHome         | 1.12      |
| Without  | $\geq 1$ employed adult | 34             | 12   | HomeWorkMaintenanceHome     | 0.99      |
| Without  | $\geq 1$ employed adult | 34             | 13   | HomeWorkEscortHome          | 0.96      |
| Without  | $\geq 1$ employed adult | 34             | 14   | HomeMaintenanceWorkHome     | 0.85      |
| Without  | $\geq 1$ employed adult | 34             | 15   | HomeEscortShoppingHome      | 0.78      |
| Without  | $\geq 1$ employed adult | 34             | 16   | HomeLeisureHomeLeisureHome  | 0.74      |
| Without  | $\geq 1$ employed adult | 34             | 17   | HomeShoppingLeisureHome     | 0.74      |
| Without  | $\geq 1$ employed adult | 34             | 18   | HomeLeisureShoppingHome     | 0.72      |
| Without  | $\geq 1$ employed adult | 34             | 19   | HomeShoppingHomeLeisureHome | 0.68      |
| Without  | $\geq 1$ employed adult | 34             | 20   | HomeEscortHomeLeisureHome   | 0.64      |
| Without  | $\geq 1$ employed adult | 5+             | 1    | HomeWorkHome                | 22.2      |
| Without  | $\geq 1$ employed adult | 5+             | 2    | Home                        | 15.88     |
| Without  | $\geq 1$ employed adult | 5+             | 3    | HomeLeisureHome             | 9.13      |
| Without  | $\geq 1$ employed adult | 5+             | 4    | HomeShoppingHome            | 8.16      |
| Without  | $\geq 1$ employed adult | 5+             | 5    | HomeEscortHome              | 7.68      |
| Without  | $\geq 1$ employed adult | 5+             | 6    | HomeMaintenanceHome         | 4.38      |
| Without  | $\geq 1$ employed adult | 5+             | 7    | HomeWorkShoppingHome        | 1.95      |
| Without  | $\geq 1$ employed adult | 5+             | 8    | HomeWorkHomeLeisureHome     | 1.89      |
| Without  | $\geq 1$ employed adult | 5+             | 9    | HomeEducationHome           | 1.71      |
| Without  | $\geq 1$ employed adult | 5+             | 10   | HomeWorkLeisureHome         | 1.33      |
| Without  | $\geq 1$ employed adult | 5+             | 11   | Work                        | 1.12      |
| Without  | $\geq 1$ employed adult | 5+             | 12   | HomeMaintenanceWorkHome     | 0.96      |
| Without  | $\geq 1$ employed adult | 5+             | 13   | HomeWorkMaintenanceHome     | 0.93      |
| Without  | $\geq 1$ employed adult | 5+             | 14   | HomeWorkEscortHome          | 0.88      |
| Without  | $\geq 1$ employed adult | 5+             | 15   | HomeShoppingLeisureHome     | 0.82      |
| Without  | $\geq 1$ employed adult | 5+             | 16   | HomeEscortShoppingHome      | 0.67      |
| Without  | $\geq 1$ employed adult | 5+             | 17   | HomeLeisureMaintenanceHome  | 0.66      |
| Without  | $\geq 1$ employed adult | 5+             | 18   | HomeMaintenanceLeisureHome  | 0.63      |
| Without  | $\geq 1$ employed adult | 5+             | 19   | HomeEscortHomeLeisureHome   | 0.6       |

| <b>Children</b> | <b>Employment</b>       | <b>Household<br/>Size</b> | <b>Rank</b> | <b>Activity chain</b>   | <b>Share<br/>(%)</b> |
|-----------------|-------------------------|---------------------------|-------------|-------------------------|----------------------|
| Without         | $\geq 1$ employed adult | 5+                        | 20          | HomeLeisureShoppingHome | 0.6                  |

Table 27: Twenty most frequent activity chains by household group