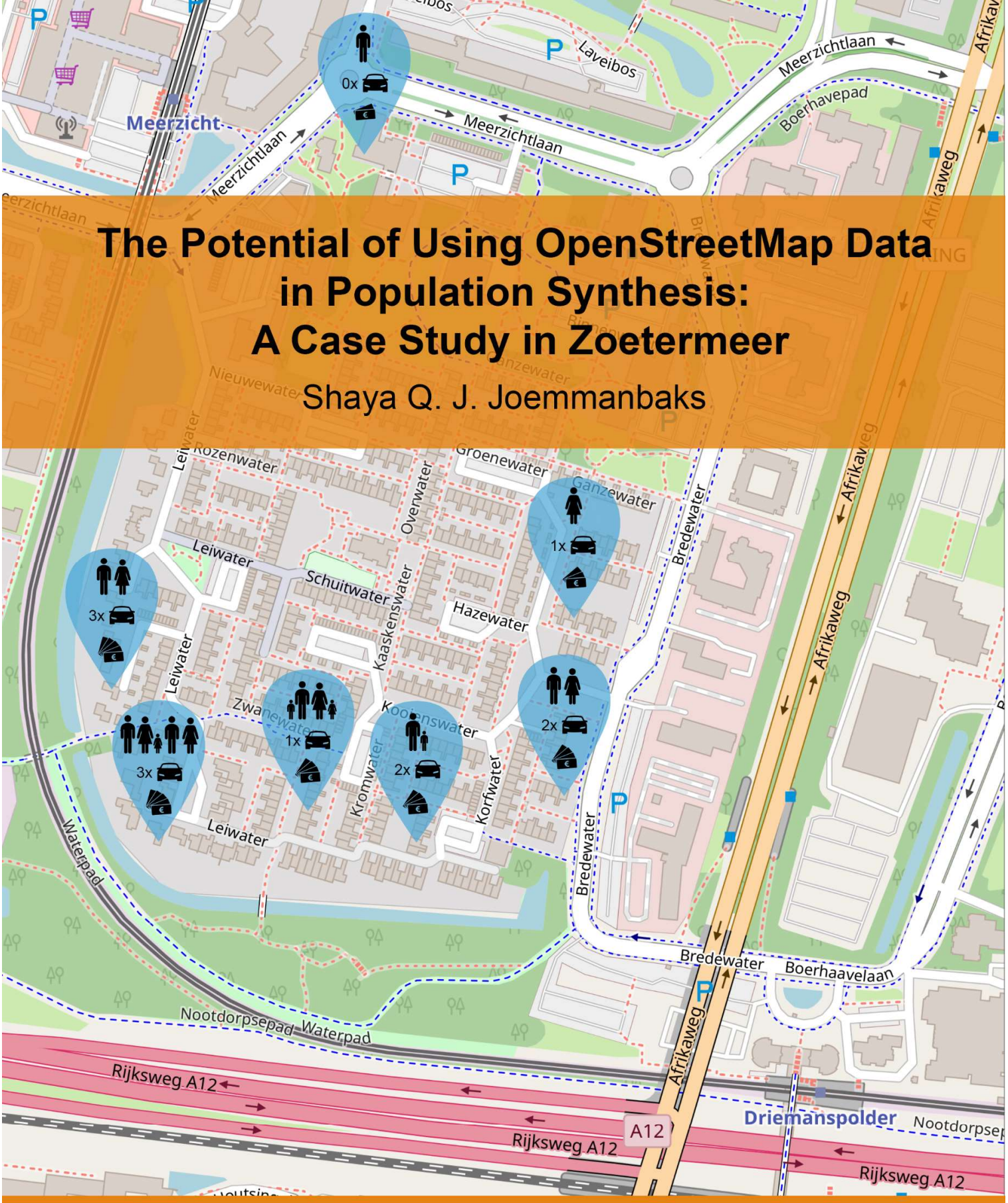

The Potential of Using OpenStreetMap Data in Population Synthesis: A Case Study in Zoetermeer

Shaya Q. J. Joemmanbaks

TU Delft Delft University of Technology

The Potential of Using OpenStreetMap Data in Population Synthesis: A Case Study in Zoetermeer

by

Shaya Querida Jaleela Joemmanbaks

to obtain the degree of

Master of Science Civil Engineering – Transport and Planning

At the Delft University of Technology

To be defended publicly on April 25, 2022

Student number: 4802292

Project duration: May 27, 2021 – April 25, 2022

Thesis committee:

Chair	Dr. Ir. R. Van Nes	Delft University of Technology
Daily Supervisor	Dr. Ir. A. J. Pel	Delft University of Technology
Second supervisor	Dr. Ir. W. Daamen	Delft University of Technology
Company Supervisor	J. Kiel, MSc.	Panteia B.V.

An electronic version of this thesis is available at http://repository.tudelft.nl/

TU Delft

Panteia
Research to Progress

PREFACE

This thesis project is the final phase of the master's program in Transport and Planning (Civil Engineering) at TU Delft and marks the end of my journey as a student in the Netherlands. This research aims at establishing a methodology for synthesizing a population and adding spatial units to this population through OpenStreetMap data. I hope that this report will be of value for future research and practice.

It has been a special adventure being an international student from Suriname and in times of COVID-19. This road was filled with ups and downs. I have encountered different academic and personal challenges that led to the growth of me as a researcher, an engineer, and a person. This journey would not have been possible without the support of the people around me.

First, I would like to take this opportunity to thank my thesis committee. I thank Rob van Nes, the chair of this committee, whose expertise was invaluable in formulating the research proposal and coordinating the process. I owe a debt of gratitude to my daily supervisor, Adam Pel, for this untiring support and guidance throughout this project. His insightful feedback pushed me to sharpen my thinking. I would like to thank my second supervisor, Winnie Daamen, for her time and careful attention to detail. Her feedback made me identify aspects that could have been easily overlooked. I am also grateful to Jan Kiel for helping in providing a research topic and scoping this research. His feedback has given me a lot of insights into the practical side of transport modelling and his guidance has kept me focused on the goal of this research instead of feeling overwhelmed by all the different components of the methodology.

I would also like to acknowledge colleagues from my internship at Panteia for their wonderful collaboration. I thank Arnaud Burgess for giving me the opportunity to do my thesis internship at Panteia. I am grateful to Ivo Hindriks and Roeland Houkes for their patience and guidance with helping me program the procedures in this research. I thank Yuko Kawabata for her support in delineation of the study area.

Finally, I would like to thank my family and friends without whom I could not have completed this thesis. I am thankful to my family for encouraging me in all my pursuits and inspiring me to follow my dreams. I am especially grateful to my parents, Jeffrey and Chandra, who believed in me and supported me emotionally and financially. I am grateful to my grandparents for helping me stay motivated and inspired. I thank my sister, Shary, for her unrelenting support and encouragement. I thank my Surinamese friends Gio, Pravish and Ita who together with my sister created a home away from home. I am also thankful to Ali for helping me finetune images and for all the fun conversations. I am grateful to my friends from the university. Sofia, Aswin, Raunaq, Mukil and Neeraj, I thank you for providing stimulating discussions as well as happy distractions to rest my mind outside of my research.

Shaya Joemmanbaks

Delft, April 2022

EXECUTIVE SUMMARY

Recent years have seen an increase in urbanization, spatial restructuring, and population growth. With the rise in computational power of computers and open-source data becoming more accessible, new opportunities arise for microsimulation models. Most of these models require a realistic synthetic population. However, using collected microdata causes privacy and confidentiality concerns and is often not available. Therefore, a process named population synthesis is adopted to generate a synthetic population that on aggregate levels adheres to the real population. This synthetic population contains attributes associated with households and/or individuals.

This population can be further enriched and ready to be implemented in transport models if they include a spatial distribution of the households as well. This results in a population for which the home end of trips/tours and in activity schedules is known. This distribution can be made realistic and accurate by taking attributes of households and houses into account when allocating households to houses (denoted as household allocation). Crowd-sourced OpenStreetMap data (OSM data) has shown potential to be a viable data source in literature and will be explored in this research to provide the spatial units for the synthesized population. This implies that the houses and residential units will be retrieved from OSM data to function as the spatial units by which the synthesized population is distributed.

This research contributes to the current body of literature by providing implementation details, transparency and modelling a synthetic population for small areas in detail which has not been done in studies before. It also proposes a methodology that outlines steps for generating a population and adding spatial units through OSM. Moreover, it provides empirical evidence on the application through a case study. Furthermore, this research aids in paving the way for using OSM data in microsimulation models by analysing the quality of OSM and its suitability. Lastly, the household allocation that combines population synthesis, OSM data and a statistical technique for allocation, also forms a contribution.

For the research, the following main research question was formulated:

How can population synthesis be carried out for neighbourhoods and to what extent can OpenStreetMap data be used to add a spatial distribution to the synthesized population?

A review of existing literature has led to identifying components that need to be part of the methodology that will be developed. These components included population synthesis type, input data, control variables, validation, OSM data quality assessment and choice of method for household allocation.

There are different population synthesis techniques found in literature with each having pros and cons. The most researched method is 'Synthetic Reconstruction'. This method adopts a statistical procedure named 'Iterative Proportional Fitting' (IPF) to estimate and reweight joint distributions from sample data (microdata) by setting population constraints. This method was chosen for this research because of its many advantages including robustness, computational ease, guarantee of convergence and flexibility of spatial units (Choupani & Mamdoohi, 2016).

This method differentiates between single-level fitting and multilevel fitting. The single-level fitting approach is only able to adhere to constraints at the household level or individual (person) level at a time. Whereas multilevel fitting approaches process constraints at the household level and individual level simultaneously. The variables used to reweight the sample data and for which the totals are used as constraints are called control variables.

There are usually two types of data used, namely aggregate data (constraints) and disaggregate sample data. Aggregate data are demographic summary tables from the fully enumerated population synthesis and are used as constraints in population synthesis. Aggregate data is often referred to as marginals or totals.

Disaggregate sample data is a representative sample file from unit records that are randomly drawn from a population census. Disaggregate data is often denoted as seed data or sample data.

Through application in a case study in the neighbourhood Meerzicht Oost in Zoetermeer, more components were identified and the entire methodology containing sequential steps was established. The methodology is presented in Figure 1. The solid lines represent the input that is needed, and the dashed lines illustrate input that is fed back to steps that have already been carried out. Steps 1, 2, 3 and steps 5 and 6 can be carried out simultaneously.

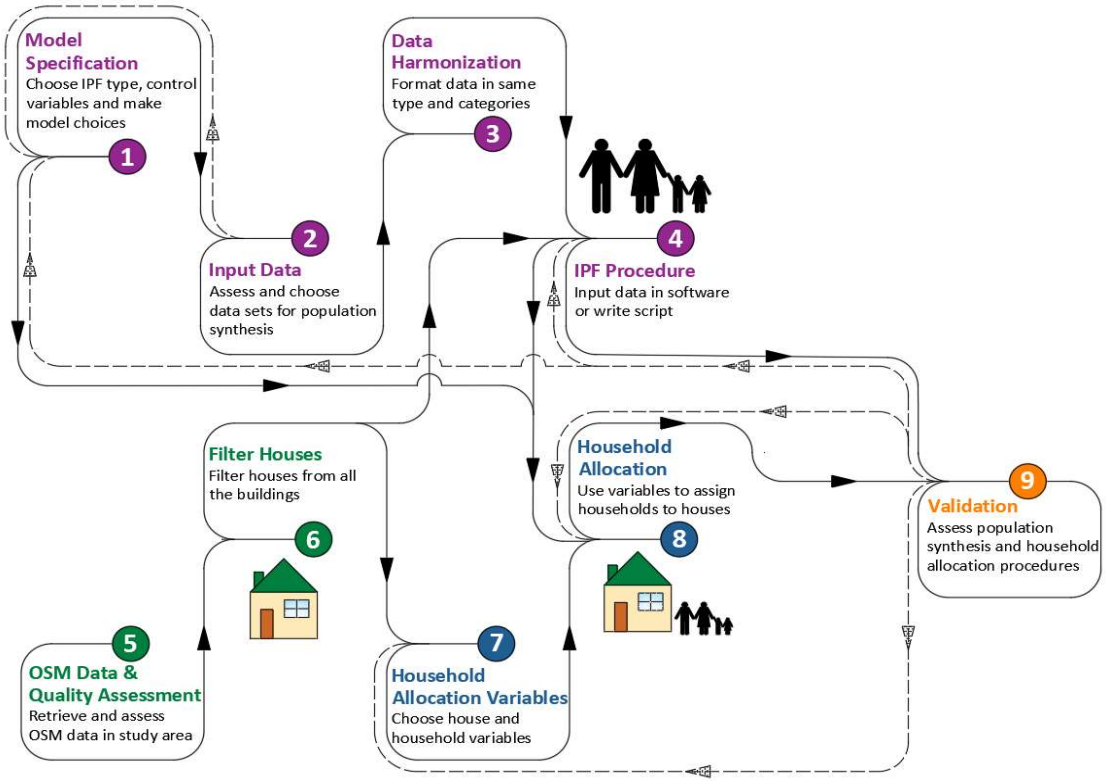

Figure 1 Developed methodology

The steps outlined in the methodology will be described below along with implementation in the case study. The steps are:

1. Model specification
   This entails making model choices such as the type of IPF that will be used, the control variables, assumptions, and simplifications. It was opted for a single-level fitting approach in the case study with the control variables household composition, household income and car availability (the number of cars in a household).
2. Input data
   This pertains to the data sets being chosen for population synthesis. For the case study, it was found that the data collected in the Netherlands is not available at the fine geographical scale needed and that microdata samples are not available as open-source data either. As a mitigation strategy, aggregate data was used from higher geographies (Zoetermeer) and then downscaled to the size of the study area (neighbourhood Meerzicht Oost). And for the microdata to be used as seed data, the OViN (Onderzoek Verplaatsingen in Nederland) data set that was available was used.

3. Data harmonization

   This step is the pre-processing of the data by converting it into the format needed for population synthesis. The data collection institutes in the Netherlands also do not harmonize the data sets to include general variables with the same definitions, categories, spatial scale and time and this leads to inconsistencies in the data sets limiting their usability.

4. IPF procedure

   This step focuses on the IPF procedure itself. The researcher can either use existing software for population synthesis or program the procedure. For the case study, the procedure was programmed in Python-based Jupyter Notebook. The population was generated for Zoetermeer and then downscaled to the neighbourhood using a multiplier. This multiplier was calculated through the total number of houses retrieved from step 6 and dividing this by the total number of households in Zoetermeer.

5. OSM data and quality assessment

   This entails retrieving the OSM data for the study area and analysing its suitability for providing the spatial units through indicators. These indicators include completeness (how complete the data in OSM is compared to validated data sets), positional accuracy (how much the position and geometry of OSM entities such as buildings and roads differ from validated data sets) and thematic accuracy (how well OSM entities are classified). In the case study, mainly the thematic accuracy was analysed as OSM data is imported data from the Automotive Navigation Data (a validated data set). OSM data has for each entity a set of keys and values specified. The keys and values together are called tags and provide characteristics of the entities. The thematic accuracy was assessed using these tags and checking the number of versions, sources and the richness. Field observations and a comparison between OSM and Google Maps were also used. It was found that the OSM data could not be used without corrections. These errors were corrected through field observations and the register for addresses and buildings (Basisregistratie Adressen en Gebouwen (BAG)).

6. Filter houses

   This step focuses on filtering the houses or residential units (flats and apartments) in the study area by using the tags. This was done for the case study after the corrections and the number of houses was fed back to step 4.

7. Household allocation variables

   This step seeks to find the household allocation variables. The household characteristics of the synthesized populations and the set of characteristics of the houses and apartment buildings serve as the household allocation variables. In the case study, the household composition, household income and living area of the house were used as household allocation variables. However, there was no open-source data available to establish a relationship between the chosen variables. Therefore, experts were consulted, and their expert judgement was elicited.

8. Household allocation

   This step describes the procedure for household allocation and rules that can be implemented to make the household allocation more accurate. The chosen household allocation technique was linear regression analysis, and this model was trained using a data set retrieved from the elicitation of expert judgement. The regression analysis then calculated a desired area for each generated household. The house in OSM which has a living area that satisfied this desired area of a certain household, was then allocated to it. Additional rules were formulated for allocation that were related to the household income, car availability and property valuations.

9. Validation

   This pertains validating the population synthesis internally and doing a (partial) external validation. The results of the household allocation also must be validated. For the case study, internal validation was used in the form of the Pearson's correlation coefficient and a perfect correlation was found. And for external validation, the results of the population synthesis were aggregated and

compared to external data sets. It was found that there were minor differences, but given the data restrictions, the synthesized population still came close to the real population. The household allocation was validated using the housing survey of the Netherlands (Woon Onderzoek Nederland (WoON)) to again distribute the households and compare the expert judgement distribution with the distribution of the WoON data set. It was concluded that these two distributions were similar indicating that the method functions well given the quality of the data and subjective data from expert judgement.

After the implementation of the case study, it was concluded that population synthesis can be carried out at the level of neighbourhoods but that there are limitations due to data availability and these limitations insert uncertainty in the model output. Although, OpenStreetMap did provide the spatial units for the synthesized population, it did need to be corrected by other geospatial data (BAG) and field observations. This indicates that OSM data does have potential but can currently only be applied in combination with other sources. This proof of concept does show that with relatively low data requirements, a plausible population synthesis with high spatial resolution and granularity can be obtained. And this is certainly valuable to the field of transport modelling and more specifically microsimulation.

This findings and analyses of this research resulted in the following recommendations for future research:

- Perform proper external validation by collecting microdata and marginals for the case study area and thus obtaining a ground truth.
- When collection of data is not possible, implement the entire protocol for structured elicitation of expert judgement.
- Conduct a sensitivity analysis for the household allocation by altering attributes of the synthesized population.
- Include more control variables in the population synthesis and more household allocation variables and assess how this influences the distribution of households in the study area.
- Further research also necessary for refining the household allocation and looking into more sophisticated allocation methods that also include stochasticity.
- Implement the methodology in areas other than residential neighbourhoods to analyse the transferability.
- Utilize the model output (spatially distributed synthesised population) by implementing this in a microsimulation model for transport to assess the value of having such a disaggregated population

.

Keywords: *population synthesis, OpenStreetMap data, spatial microsimulation, household allocation*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADAPTS	Agent-based Dynamic Activity Planning and Travel Scheduling
ALBATROSS	A Learning-based Transportation Oriented Simulation System
CEMDAP	Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns
C.I.	Conditional Independence
CO	Combinatorial Optimization
BAG	Basisregistratie Adressen en Gebouwen (Register for addresses and buildings)
HIPF	Hierarchical Iterative Proportional Fitting
ILUTE	Integrated Land Use, Transportation and Environment model
I.I.A.	Independence of Irrelevant Alternatives
I.I.D.	Independently and Identically Distributed
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
ODiN	Onderweg in Nederland (Research in Mobility of the Netherlands)
OLS	Ordinary Least Squares
OOP	Object-Oriented Programming
OSM	OpenStreetMap
OVG	Onderzoek Verplaatsingsgedrag (Dutch Travel Survey)
OViN	Onderzoek Verplaatsingen in Nederland (Research in Mobility of the Netherlands)
PCU/d	Personal Car Units per day
POI	Point of Interest
PopGen	Population Generator
PUMA	Public Use Microdata Area
PUMS	Public Use Microdata Sample
VGI	Volunteered Geographic Information
V-MRDH	Verkeersmodel Metropool Rotterdam – Den Haag (traffic model Rotterdam – The Hague metropolitan area)
SBM	Simulation-based method
SimTravel	Simulator of Transport, Routes, Activities, Emissions and Land model
SR	Synthetic Reconstruction
TIGER	Topologically Integrated Geographic Encoding and Referencing
TRANSIMS	TRansportation ANalysis SIMulation System
WoON	Woon Onderzoek Nederland (housing survey of the Netherlands)

1 INTRODUCTION

The past century has seen a surge in urbanization and spatial restructuring. This increase has demanded special attention from the transportation field. Transport and traffic systems play a vital role in shaping today's society. To get a handle on these systems, models are often utilized. With the increase of computation power, storage and more accessible public or open-source data, more opportunities arise for the development and application of microsimulation models.

Most of these microsimulation models need a realistic individual-level and/or household-level population. This requires detailed socioeconomic and socio-demographic data of the population for a geographical zone. Due to privacy, confidentiality, and data collection issues, data is mostly only available at aggregate levels and for big geographical scales. And disaggregate data is sometimes available but only as small samples. To overcome these limitations, a process called population synthesis or spatial microsimulation is used to generate a synthetic population.

It is progressively becoming more apparent that there is a geographical impact as a result of government policies, investments, and social networks (Ballas & Clarke, 2009). Researchers realize that spatial microsimulation (population synthesis) is a valuable tool in estimating these impacts. Spatial microsimulation models enable analysis with micro units and are increasingly being used in various applications (O'Donoghue, Morrissey, & Lennon, 2014).

Applications that profit from these disaggregated models can be found in different domains such as transport planning, policy, health care, economy, socio-demography, and many more. These applications demand more accuracy and granularity from the models and must represent reality accurately and at high geographic detail. The precision and realism with which changes in populations within geographical areas are represented may have significant implications for modelling, particularly when the results are used to advise policy (Harland, Heppenstall, Smith, & Birkin, 2012).

1.1 CONTEXT

In this paragraph, background information is given for this research. The first subsection focuses on population synthesis. Then OpenStreetMap data (an opensource geodata source) is introduced to add spatial detail to the results of the population synthesis. In doing so, the synthetic population is given a spatial distribution. Elaboration of this spatial distribution is given in the last subsection of this paragraph.

1.1.1 POPULATION SYNTHESIS

Population synthesis can be described as "a procedure that generally involves expanding a sample drawn from a population to a full set of synthetic population, such that the generated synthetic population conforms as much as possible to the actual population at various aggregation levels" (Lim & Gargett, 2013, p. 2). This synthetic population is then representative of the actual population because it is generated based on real population census data of the chosen area in such a way that it forms a match with the data on aggregate levels.

To illustrate this with an example, assume a study area in which aggregate data like income and age are available for a region. This data is shown as orange checks in Figure 2. The study area has some disaggregated data for income and age in the form of samples of households. This sample data can be presented as a cross-tabulation. To enumerate the sample to the size of the population of the entire area, the aggregate data is used as constraints. This data is often referred to as marginals or totals. The seed data (sample data) is then multiplied with weights iteratively until the sum of the rows and columns match

with the marginals. This results in the fitted values illustrated by stars in Figure 2. Income and gender are called control variables as they are used to reweight the sample data in this case. The result is an enumerated cross-tabulation that covers the study area.

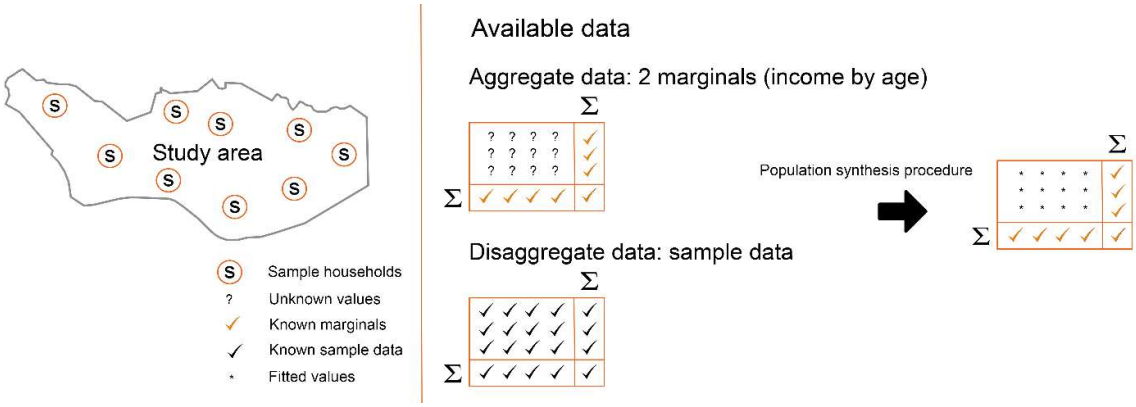

Figure 2 Working of common population synthesis adapted from Hobeika (2005)

Population synthesis has as input microdata apart from the aggregate data. This is usually population census data at the level of households or individuals providing socio-demographic characteristics. And even though population census data is becoming more accessible, this data is still lacking for small geographies (e.g. neighbourhoods). Spatial microdata that consists of demographic information is also limited and not available for small areas. This makes analysis of subgroups within the population difficult, if not impossible. With population synthesis, techniques are applied to generate new data with the demographic granularity of individual surveys and the spatial granularity of small area censuses and surveys (Lovelace, Ballas, & Watson, 2014).

The lack of access to accurate and available geocoded (spatial) microdata forms a fundamental barrier in population synthesis. Even when using more specific surveys to gain further insight into travel patterns at the individual level, high-resolution geographical information (such as street addresses or postal codes) is still omitted due to privacy issues. Many countries simulate spatial microdata because reliable secondary data sources are restricted to zonally aggregated census data and non-geographical, individual-level microdata (Lovelace, Ballas, & Watson, 2014).

In the transport domain, population synthesis is used in disaggregated travel demand models to estimate the travel demand. These models assume that every household and individual expresses different travel behaviour and can predict this behaviour. These models are not only able to model the population on an individual level, but they give the opportunity to accurately model locations for houses (often productions) and attractions (firms, facilities, and locations of activities) as well (Briem, Heilig, Klinkhardt, & Vortisch, 2019). Briem et al. (2019) even suggested that an automatic or semi-automatic process collecting the required data from geodata sources would speed up the creation of new travel demand models.

The population synthesis is only responsible for generating agents and/or households to gain a synthetic population for a geographical area. The specific location of where these agents and households reside (i.e. their houses), is removed from the microdata so this is not part of the synthetic population. The result is that the microdata is no longer geocoded. When using this type of synthetic population in microsimulation models, the agents are randomly assigned to houses. These houses are often not mapped accurately within the model and there are not many aspects considered such as the relationship between household composition and house size or the relationship between income and neighbourhood of the house when allocating households/agents to houses. This sparks the question of how the microdata can be geocoded again so that households and agents within a household can be allocated to houses more accurately. And

consequently, productions and attractions can be known at this fine geographical resolution. This paves the way for looking into the potential of a geodata source that can be combined with the microdata to attach geographical information.

1.1.2 OPENSTREETMAP DATA

In the last decade, a new phenomenon of crowdsourced geodata (Volunteered Geographic Information, VGI) has emerged. This data is collaboratively collected by users and shared on an online community platform. It consists of a special kind of user-generated content of which the geo-location of a distinct feature is an integral part of the collected data. It has been shown that VGI, especially OpenStreetMap (OSM), holds the potential to serve as a major data set in urban areas (Goetz & Zipf, 2012).

The goal of OpenStreetMap is to create a free digital map of the world. They achieve this through the engagement of participants in the OSM community (Haklay, 2010). OSM data is based on collected data from GPS tracks through digital tracing of aerial images such as Landsat and Yahoo! Imagery, or data from other free sources, personal photography, or maps. Some contributors provide OSM with geographic data by importing the geodata to OSM data. Some instances of this are Topologically Integrated Geographic Encoding and Referencing (TIGER) data for the US, GeoBase data from the Canadian government, and Automotive Navigation Data for the Netherlands (Kounadi, 2009).

The data in OSM also gives information on features such as the road category, building type, and amenity type. Even though there are resources that are from exclusive software that conforms to the conventional standards, open-source software is quite often comparable or even superior in quality. This is owing to its openness, the code to the software can be seen and modified by each user (McConchie, 2008 as cited in Kounadi, 2009).

1.1.3 SPATIAL DISTRIBUTION OF POPULATION

Using OpenStreetMap data to add a spatial distribution to a synthetic population can be a solution to not having geographical information (such as addresses) attached to the population. The geographical information that was previously omitted can be added to the synthetic population using rules or statistical techniques and information that is stored in OpenStreetMap. This leads to households being attached to houses and this process will be denoted as household allocation hereafter. This will not be an “exact” match since allocation will be done based on similar characteristics of houses and households and the generated population is synthetic.

The network as mapped in OSM contains not only roads but also buildings and information about these buildings. This information can be utilized to identify houses or residential units. The synthesized households can be appropriately allocated by making use of statistical techniques that take attributes of households and houses into account along with the relationship these attributes have with each other. The houses/residential units would then be incorporated as the spatial units by which the population is allocated resulting in a more accurate spatial distribution. And since this concerns a synthetic population, the addresses or houses assigned to the generated households should not pose any privacy issues. This leads to synthetic geocoded microdata with the demographic granularity of individual surveys and spatial granularity of small area censuses that can be used as input in microsimulation models to get a more refined travel demand. The synthetic population by itself can also be used to explore behaviour because of changes in policies or demography.

In this household allocation procedure, it is important to know three distinctions of the household population in the study area. The first one is the actual study area population, which gives the number of households that are present in the study area. The second distinction is the number of households in the study area based on the number of houses according to OpenStreetMap data with the underlying assumption that

only one household is present in each house. And the third distinction is the synthesized population that is equal to the number of households resulting from either the first or second distinction. The researcher can choose whether the first or second distinction holds for the synthesized population. This is of course dependent on whether the assumption of one household per house is enforced or not. For this research, the spatial units (i.e., the houses) that are attained from OSM will function as the number of households in the study area. And this implies that the real number of households in the study area population might not be equal to the number of households in the synthetic population.

1.2 PROBLEM DEFINITION

There are several methods for synthesizing a population. These methods will be described in chapter 2. The most used method is named Iterative Proportional Fitting (IPF) (Choupani & Mamdoohi, 2016). This procedure is known to be robust. It requires less census data than other population synthesis techniques and can generate a large amount of synthetic microdata that can be used as input for traffic models. However, most literature on IPF is insufficiently described to reproduce the procedure for non-specialists (Lomax & Norman, 2016). Also, in most papers, the IPF procedure was not used at such a high geographical resolution as neighbourhoods (it was mostly used for cities), which leads to two gaps in IPF that require further investigation for this research.

Another problem is that in many cases sociodemographic microdata for small areas is not available or not collected. Hence, the microdata that is needed for neighbourhoods as the disaggregate input data for population synthesis is not readily available. This requires an additional step before implementing the population synthesis and it needs to be investigated whether microdata from bigger geographies can be used and scaled down for smaller areas and to what extent this results in a representable synthetic population for the smaller areas.

Lastly, the use of OpenStreetMap data in travel demand models is still in its infancy. There are many questions regarding the quality and accuracy of this data, and this is owed to its crowdsourcing nature. In the context of population synthesis, OSM data can provide the spatial units (houses/residential units) by which the population can be distributed. This research will therefore also investigate the suitability of OSM data for allocation of the generated population.

1.3 FOCUS AND SCOPE

The research will focus on population synthesis at a fine geographical scale (i.e. neighbourhoods) and will add a spatial element to the generated synthetic population by attaching houses to the households within this area through OpenStreetMap data and a statistical technique. In doing so, it is explored whether OpenStreetMap data is equipped with sufficient information to allocate households. The envisioned result is a population with the home locations of residents accurately known in the network of OpenStreetMap. This result can then be used to implement scenarios in different research domains or can be used as input in transport models.

To scope the research, only open-source data and in-house data from Panteia B.V. was used. For the population synthesis procedure, the focus will be to outline all the steps, describe the data requirements and these houses provide implementation details through a case study. For allocation of households to houses, only house and household characteristics are considered. This means that no social factors such as community cohesion or built environmental factors such as proximity to schools are considered. For this allocation process, it is also assumed that one household resides in each house. OSM data will be the only

geodata source. The population synthesis and household allocation will be programmed in Python Jupyter Notebook to make the script easily accessible.

1.4 RELEVANCE AND IMPORTANCE OF RESEARCH

The current state of knowledge in the field of population synthesis lacks implementation and transparency and modelling a population for small areas in detail (Choupani & Mamdoohi, 2016; Lomax & Norman, 2016; Rich, 2018; Lim, 2020). It is thus essential to gain an in-depth understanding of population synthesis in all its steps to generate a representable synthetic population for small geographies. Population synthesis and specifically IPF is usually not performed at this geographical resolution. The first contribution of this research is methodological by providing a framework for generating a population at the fine scale of neighbourhoods. Another contribution stems from the empirical evidence on the implementation and application of the proposed method in a case study. OpenStreetMap has shown potential in being a reliable data source but has not been used often in spatial microsimulation. Therefore, this thesis also contributes to the less explored realm of OpenStreetMap data in population synthesis by showcasing the extent of the suitability of OSM data for attaching geographical locations to the synthesized population. Furthermore, the development and proof of concept of the household allocation by combining population synthesis with OpenStreetMap data to gain a spatial distribution of households in a study area also forms a contribution of this research.

1.5 RESEARCH OBJECTIVE AND RESEARCH QUESTIONS

The research objective is to create a proof of concept of a method in which OpenStreetMap data can be utilized to add a spatial distribution to synthetic microdata. The synthetic microdata is generated through population synthesis at the fine spatial scale of neighbourhoods. The methodology will be developed through literature and a case study. The case study also demonstrates how practical the methodology is given the available data.

From the research objective, the main research question is formulated as:

How can population synthesis be carried out for neighbourhoods and to what extent can OpenStreetMap data be used to add a spatial distribution to the synthesized population?

This main question is supported by sub-questions. These sub-questions are grouped in categories to highlight the part of the methodology the question relates to. The sub-questions are:

1. *Methodology:*
    a. What population synthesis technique can be selected for this research?
    b. What steps need to be outlined in the methodology?
    c. Which statistical technique can be used to allocate households to houses?
    d. How can the generated synthetic population be validated?
2. *Data:*
    a. What are the data requirements for the chosen population synthesis technique?
    b. What data in OpenStreetMap can be used for the allocation of households to houses?
    c. How can the quality of OpenStreetMap data be assessed?
    d. How can input data still be derived when confronted with a lack of (micro)data for small areas?

3. *Case study:*
    a. Which control variables should be used in population synthesis to get a representative population?
    b. Which variables from the available OpenStreetMap data for the study area can be used to allocate houses to the generated households?

1.6 RESEARCH METHODOLOGY AND REPORT OUTLINE

The research methodology used to conduct the research is shown in Figure 3. This shows all the steps taken to answer the research questions and consequently developing the methodology for population synthesis in small areas and adding a spatial distribution to the synthesized population using OSM data.

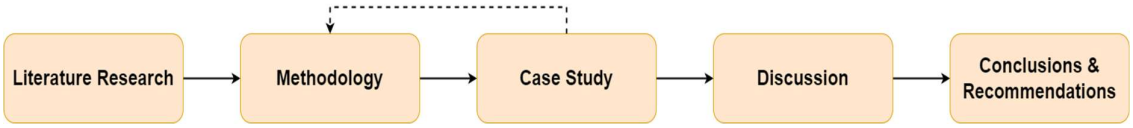

Figure 3 Research methodology

The literature research is done first and is described in Chapter 2. The literature research looks at population synthesis, OSM data and candidate methods for household allocation. The literature research will aid in answering research questions in the category of the methodology and data. The second step is the development of the methodology based on findings in the literature research and this will be presented in Chapter 3. The following step is the implementation of the developed methodology in a case study which will seek to answer questions listed in all three categories. The findings and implementation details in the case study help refine the methodology. Hence the feedback loop from the case study to the methodology. Chapter 4 focuses on the setup of the case study and the results. Then the results are analysed and discussed in Chapter 5. Finally, the findings are described and linked to the research questions in the conclusions and recommendations, which is Chapter 6. Limitations are also mentioned in this chapter along with suggested topics for future research.

The appendices in this report give additional explanations and data. This research also has supplementary files that contain the Jupyter Notebook with the code and instructions and files with the data used in this Notebook.

2 LITERATURE STUDY

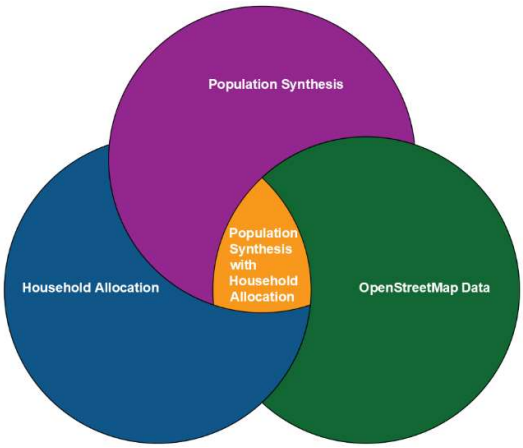

Figure 4 Themes of literature review

This chapter presents the literature review. The purpose of this review is to showcase the existing literature, identify literature gaps and finally to help build a conceptual framework for the methodology. Implementation details that are not part of scientific research but are useful to this research are also discussed in this chapter. To structure the literature review, themes were identified, namely population synthesis, OpenStreetMap data, and household allocation. These themes are all components of the methodology. The population synthesis is responsible for generating a synthetic population. One of the population synthesis techniques was already briefly described in Figure 2. OpenStreetMap data is used to identify and retrieve houses or residential units with attributes that are specified in OSM for a study area. And the household allocation uses the retrieved data from OSM, characteristics of the synthetic population and a statistical technique to allocate the households to houses. All these themes come together to answer the main research question which focuses on population synthesis and OSM data for household allocation. The themes are illustrated in Figure 4. This chapter will go through all the themes and the figure will be extended to include all the subthemes too. Finally, the research gaps and conclusions are summarized.

2.1 POPULATION SYNTHESIS

The first of the themes that is covered is population synthesis. The structure for this theme is illustrated in Figure 5 on Page 22. The population synthesis methods found in literature are synthetic reconstruction, combinatorial optimization, the simulation-based method, and sample-free methods. Of all these methods, synthetic reconstruction is found most often and has several variations whereas combinatorial optimization only has one variation, and the sample-free methods have three variations. Each of these methods will be described and are compared to each other in a qualitative manner.

Since literature on synthetic reconstruction is more common and more applied, recurring components could be identified. These components include input data, decisions for control variables, the multiple variations that exist for the methods and validation. Implementation details is lacking in literature and was more evident in non-scientific literature in the form of software development platforms as GitHub and question and answer websites such as Stack Overflow.

According to Bowman (2009), all population synthesis procedures have two common stages: fitting and allocation. Fitting is the stage where an aggregate representation of the target population is computed for the base year. And allocation is the stage where disaggregation is performed (Müller & Axhausen, 2010). During this allocation stage, the groups or households are computed for each agent based on a population distribution (Ye, Wang, Chen, Lin, & Wang, 2016). This allocation stage should not be confused with the household allocation stage. The difference is that the allocation stage mentioned by Bowman (2009) and Müller & Axhausen (2010) describes the process of redistributing the population over different categories and zones. And the household allocation entails assigning the generated households from the allocation stage by Bowman (2009) and Müller & Axhausen (2010) to physical houses in OpenStreetMap.

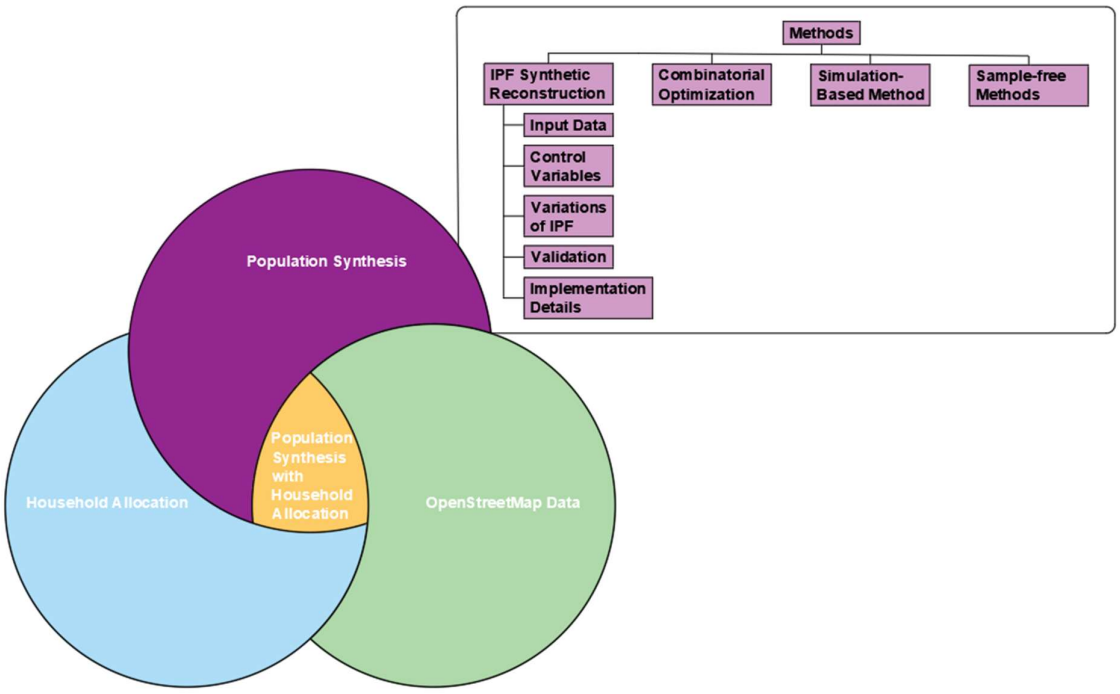

Figure 5 Population synthesis subthemes

2.1.1 POPULATION SYNTHESIS TECHNIQUES

There are two main approaches in literature for population synthesis, these are sample based (meaning they require a sample data set):

- Synthetic reconstruction:
  These approaches are mostly used when the only available data are small area crosstabulations and when suitable microdata is not available. Most of these approaches use a statistical method known as Iterative Proportional Fitting (IPF) and often in combination with Monte Carlo sampling to combine joint-probability distributions from small area census tables (Lovelace, Ballas, & Watson, 2014). This method is the most popular method for population synthesis because the IPF procedure has many advantages (Choupani & Mamdoohi, 2016). These advantages will be discussed later.
- Combinatorial optimization:
  These approaches work by searching for the optimal combination of individuals and households from microdata to match aggregated count data. Thus, they create a synthetic population by randomly allocating individuals from a microdata sample into every geographical zone (Lovelace, Ballas, & Watson, 2014). The goal is to maximize the goodness of fit by replacing households or individuals in the synthetic population with households or individuals from sample data (Voas & Williamson, 2000).

Apart from these two main approaches, there have also been two other approaches that are found less frequent in literature but have been emerging in recent years. These two are:

- The Simulation-based method (SBM) by Farooq et al. (2013):
  This method starts drawing agents from the sample and generates a zone population without the need of fitting tables using the Markov Chain Monte Carlo Simulation.
- Sample-free methods:

These approaches do not require a disaggregate sample but only use aggregate data to synthesize a population. Based on optimization procedures, the sample free fitting assigns appropriate household members to households and has lower data requirements (Ye, Wang, Chen, Lin, & Wang, 2016). Examples of these methods can be found in Gargiulo et al. (2010), Barthelemy & Toint (2013), and Ye et al. (2017).

2.1.2 COMPARISON OF POPULATION SYNTHESIS METHODS

To be able to pick a method for population synthesis in this research and answer research question 1a, the methods that already exist, will be discussed along with their advantages and disadvantages. When choosing the synthesizing method, the application should also be considered.

Synthetic Reconstruction (SR)

These IPF-based approaches are the most used approach and researched. It can generate a large set of disaggregate data, it requires little census data, it has computational ease and speed (Beckman et al., 1996; Bowman, 2004; Lovelace and Ballas, 2013 as cited in Choupani & Mamdoohi, 2016), guarantee of convergence (for the classic IPF) (Pukelsheim, 2013, as cited in Choupani & Mamdoohi, 2016) and flexibility of spatial units (Rahman et al., 2010, as cited in Choupani & Mamdoohi, 2016).

The procedure uses an n-dimensional table with n attributes. The (multiway) table contains the full population scale with all possible combinations of the attributes (Ye, Wang, Chen, Lin, & Wang, 2016). The IPF procedure estimates and reweights the joint distributions of a microdata sample by setting population constraints. Then households or individuals are randomly sampled to make a synthetic population that forms the best match to the estimated joint distributions (Beckman, Baggerly, & McKay, 1996).

Although SR methods also come with its challenges. One of these is the zero-cell problem and it occurs when dealing with small geographies because of a non-zero marginal for a category that has no representative in the reference sample. The IPF algorithm would have a division by zero and thus the outcome is not defined. A solution to this is to replace the false zero-cells with arbitrarily small values (Müller & Axhausen, 2011). Even though research to overcome this zero-cell problem is done often, there is still no unbiased technique to cope with it (Choupani & Mamdoohi, 2016).

Another challenge is the memory requirements for the contingency tables and this requirement grows exponentially with the number of attributes. A large contingency table is inherently sparse so storage of contingency tables should be more efficient. It is recommended to therefore use a list-based representation, and this can be easily implemented because the reference sample is normally given as a list of attributes anyway. An example of this is the list-based version of IPF that will be discussed in the next section.

Another problem is the reduction in categorization detail and number of control variables to keep memory consumption efficient. A solution to this is the list-based IPF (Müller & Axhausen, 2010). Also, according to Choupani and Mamdoohi (2016), the vast majority of IPF-based synthesizers have trouble converging when four or more control variables are used. The same research reported that several of the population synthesizers lack implementation details and have validation issues which leads to problems concerning the reuse of such synthesizers (Choupani & Mamdoohi, 2016).

When using the IPF-based approaches in applications such as agent-based models, it comes as a disadvantage that the procedure leads to non-integer weights. Therefore, fractions of agents can be the result of the population synthesis instead of whole individuals. There is a solution to this by using 'integerisation' but this can lead to breaking correlation structures (Lovelace, Ballas, & Watson, 2014).

In research by Pukelsheim and Simeone (2009), proof of convergence was given for when a contingency table can converge under the classic IPF procedure. During practical applications of this procedure, there are only convergence issues when entire rows and columns are zero and the concerned marginal total is nonzero (Müller & Axhausen, 2010). This is seen as an advantage of the IPF procedure.

Another major benefit is that the IPF-based models are deterministic and generate the same results with each model run. It is also known to be a robust and reliable technique (Mosteller, 1968; Fienberg, 1970; Wong, 1992 as cited in Lovelace, Ballas & Watson, 2014) and its speed and simplicity are among other benefits as well (Pritchard & Miller, 2012; Lovelace and Ballas, 2013 as cited in Lovelace, Ballas & Watson, 2014).

Apart from the advantages, there is also a lack of literature that specifies the population synthesis framework and all its stages (Rich, 2018). The majority of literature focused on IPF is presented in a manner that a non-specialist cannot easily reproduce the procedure (Lomax & Norman, 2016). Most of the current population synthesizers are concealed in computer codes and inaccessible language as well which causes a lack of transparency (Lim, 2020). Furthermore, to this day no literature proposes a well-established validation framework for IPF. There has been no evaluation of fitting, spatial units, integer conversion, and selection stages according to Choupani and Mamdoohi (2016).

Combinatorial Optimization (CO)

Combinatorial optimization approaches work by searching for the optimal combination of individuals and households from microdata to match aggregated count data. Thus, they create a synthetic population by randomly allocating individuals from a microdata sample into every geographical zone. The goodness of fit of these allocations is calculated after every draw by comparing the allocations to the known crosstabulation of selected variables in the zone. There is a predetermined threshold for the goodness of fit and the algorithm will keep iterating until this threshold is reached (Huang & Williamson, 2001; Voas & Williamson, 2000 as cited in Lim, 2020). This method requires more social survey microdata and is computationally more expensive than Synthetic reconstruction (Lovelace, Ballas, & Watson, 2014).

There is less research conducted on combinatorial optimization than on synthetic reconstruction in the field of transport research (Lim, 2020). This is because of the reproducibility and the computational efficiency of Synthetic reconstruction. Both of these techniques are capable of generating reliable micro-demographic data with high accuracy. However, in terms of having a smaller deviation from the real population, combinatorial optimization has a bigger advantage. In this contrast, the aspect of the scale change in the input sample and aggregate data was not taken into account (Ye, Wang, Chen, Lin, & Wang, 2016). When comparing the results of population synthesis with Combinatorial optimization and Synthetic reconstruction, it was reported by Huang and Williamson (2001) that the Combinatorial optimization generated populations that were far more accurate than the IPF method for a small population data set. For large population data sets, the IPF was more accurate. Pritchard and Miller (2012) also highlighted some conceptual problems when using Combinatorial optimization for population synthesis like the inability to directly maintain the correlation structure of the control variables in the seed data when compared to the IPF method (Ma & Srinivasan, 2015).

Simulation-based method (SBM)

In recent years, the Simulation-based method (SBM) by Farooq et al. (2013) has been on the rise and is receiving attention as it can overcome some shortcomings of IPF such as the poor scalability when

increasing attributes or control variables, only being able to fit one contingency table while there can be other solutions that match the data and combinatorial issues when apart from joint distributions from sample attributes and marginal distributions from population attributes are being used. The SBM also captures the heterogeneity that may not be properly present in the microdata because, unlike the IPF procedure, the SBM carries out true synthesis instead of cloning agents from the sample data (Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013)

This method starts drawing agents from the sample and generates a zone population without the need for fitting tables. It does not have the zero-cell problem or table sparsity and accommodates more variables in the synthesis. It was shown that the synthesized populations from SBM resulted in more accurate results, and it preserved the correlation structure of the attributes better than the IPF (Ma & Srinivasan, 2015; Choupani & Mamdoohi, 2016).

Farooq et al. (2013) has also presented that even when the SBM and IPF have the same amount of data, the IPF cannot fully take advantage of conditions because it transforms them into marginals and the IPF relies heavily on the sample to maintain the correlation and only fits the marginals. The SBM utilizes information form the sample to a lesser extent and managed to outperform the IPF in terms of making joint distributions.

Sample-free methods

These approaches have the major advantage of not requiring a sample data set or disaggregate data. The sample-free method proposed by Ye et al. (2017) is explained by creating an individual pool that is the size of the target population from the most disaggregate data source. After, the missing characteristics are initialized by drawing at random from the respective value sets. In doing so, the individual pool will be established containing all relevant attributes. Ideally, this individual pool would meet all the conditionals and marginals determined by the target population's joint distribution. Yet, this does not always occur because of conflicts between conditionals/marginals from various data sources. In a situation where this happens, an attribute shift of some persons is required. Even though the sample-free method has relaxed data requirements, the approach is still time-consuming and memory expensive because of the synthesis of the individual pool and attribute shifts (Ye, Hu, Yuan, & Wang, 2017).

The method also does not ensure simultaneous matching at both household and agent levels and lacks application details. It was applied in Belgium with only a few attributes, it is thus unknown whether it can be successfully expanded to other cases. When compared to synthetic reconstruction and combinatorial optimization, the sample-free method gave small errors, so the SR and CO approaches had better performance. An important aspect is that the bias of an input sample can be overcome to some extent when using sample-free fitting because these techniques use the complete population attributes as the initiation points. On the other hand, sample-free methods do not have the benefit that a disaggregate sample brings in terms of the associations between attributes (Ye, Hu, Yuan, & Wang, 2017).

Conclusion

To give an overview of the methods and their strengths and weaknesses, a Harris profile in Table 1 was created. Harris profiles are graphic representations of design concepts and give evaluations of these concepts by defining criteria and grading the concepts based on these criteria (van Boeijen, Daalhuizen, & Zijlstra, 2020). This allows for qualitative analysis and shows the best method by simply counting the green fields. Criteria that are generally deemed important for picking a method were formulated and the methods

were graded based on the body of literature and interpretations based on the literature. The description of these criteria is given in Appendix A.

Table 1 Harris profile for population synthesis methods

Criterion	Synthetic Reconstruction				Combinatorial Optimization				Simulation-based Method				Sample-free Methods			
	-2	-1	+1	+2	-2	-1	+1	+2	-2	-1	+1	+2	-2	-1	+1	+2
Computation efficiency and memory																
Data requirements																
Convergence																
Flexibility																
Transferability																
Performance																

Due to a lack of research on some of the methods and the differing contexts in which the methods have been applied, it is not possible to fill in the Harris profile in a definitive manner. If more research is done, the verdict might change on these criteria. In literature, limited comparisons have been done between all the methods as well, making it particularly difficult to properly compare the methods to each other. This is also the reason for not including important criteria such as reliability and robustness as this has not been properly researched and compared for all the methods.

It is also important to realize that the technique chosen is case-specific and dependent on the type of data that is available in different locations. It should therefore be noted that there is no definitive superior method that will always work in population synthesis. The methods are a collection of techniques each aimed to solve a particular problem (Fournier, Christofa, Akkinepally, & Azevedo, 2018).

Since there is so little research and application of the simulation-based method and sample-free methods, these approaches are not desirable for this research. From Table 1, it is also evident that these methods score lower on convergence and transferability. And this is allotted to the limited applications of these two methods.

The goal of this research is to add spatial granularity to population synthesis by adding physical houses as units through OSM data. So, the population synthesis method should be able to reliably produce an accurate synthetic population that is reproducible and robust. The method should also be easy to implement and not have high computational complexity or lack of transparency. Looking at the reviewed literature and Table 1, the synthetic reconstruction approach seems fitting and scores well on most of the criteria in Table 1. From the reviewed literature and Table 1, it can be concluded that Synthetic reconstruction methods are also preferred over Combinatorial optimization methods for now.

2.1.3 INPUT DATA

To answer research question 2a, it is important to shed light on the data required for population synthesis. All the reviewed Synthetic reconstruction methods start by assessing the data that is available for synthesis. This also helps to answer research question 2a. It identifies steps that should be taken in the methodology developed in this research. There are two types of data typically used for this, namely:

- *Disaggregate sample data*
  This is a representative sample file from unit records that are drawn randomly from a population census. The selected attributes in the sample file that will be used for synthesizing the population are termed control variables. The sample file enables the construction of joint-probability

distributions and these distributions create multidimensional contingency tables that are often called seed data.

- *Aggregate constraints*
  These are demographic summary tables from the fully enumerated population census or other sources of known aggregate data. These tables are one-dimensional and each table gives univariate distributions in small geographical areas (collection zones). For each of the control variables, aggregate constraints are made with respect to geographical areas. These constraints are named control marginal totals (Lim, 2020).

Primary data sources in population synthesis include the population census, traffic surveys, labour force surveys, tax records from revenue agencies, real estate cadastre data, and household registration information. These form suitable data sources even though a few of these are rarely implemented in applications. Of all data sources, population census data is most used as it directly represents the state of the target population (Ye, Hu, Yuan, & Wang, 2017).

2.1.4 CONTROL VARIABLES

The next step most reviewed literature takes is deciding on what variables to include. How representative the synthesized population is dependent on the number of control variables being used in the IPF procedure. In general, the more control variables used, the more accurate the synthetic population (Auld, Mohammadian, & Wies, 2009).

On the other hand, synthesizing with more control variables increases the computational complexity and can cause convergence issues due to the increased likelihood of false zero cells and the added dimensions that result from adding more control variables. Furthermore, adding more control variables leads to an increase in the cells of the contingency tables leading to more computer memory requirements (Lim, 2020). In a review of 15 synthesizers, Choupani and Mamdoohi (2016) stated that when synthesizing more than four control variables, convergence can be troublesome. This can be overcome by using more efficient algorithms such as the sparse list-based IPF by Pritchard and Miller (2012).

The control variables chosen are dependent on for what purpose the synthesized population will be used and in what way it must be representative of the true population, so the context and objective of the application will be important in answering research question 3a. The control variables are split into household and individual level control variables. Household control variables represent control variables at the household level, meaning that the synthesized population that consists of households will have household attributes such as the household size and household income. The individual level control variables are the control variables at the level of individuals or persons. In this case, the synthesized population consists of persons with attributes such as age and gender.

The single-level fitting approaches result in synthetic populations that either consist of households or individuals with their respective attributes. The generated populations in multi-level approaches consist of households and individuals with each of these levels having distinctive characteristics. In the reviewed literature, there were several control variables at both household and individual level that occurred in four and more studies in which a population was synthesized for use in transport models. These variables were seen as commonly found in literature and are presented in Figure 6 on Page 28. These variables can give inspiration for choosing control variables in this research.

Figure 6 Commonly used control variables

2.1.4 VARIATIONS OF IPF

The majority of the population synthesizers are based on the IPF procedures proposed by Beckman, Baggerly, and McKay (1996) (Auld & Mohammadian, 2010). They were able to generate individual records at fine geographical levels to reconstruct a synthetic population (Müller & Axhausen, 2010). The first application of a generated synthetic population in a model was in the TRansportation ANalysis SIMulation System (TRANSIMS) in 1996. This was an activity-based travel forecasting microsimulation model that covered the travel behaviour of a person over 24 hours. Survey data was used to derive the activities for this (Hobeika, 2005). Since this application, there have been several types of research on how to make IPF more efficient and more accurate. Most of these developments were applied in models and were aimed to solve different types of problems. The modifications are shown in Figure 7. Synthetic reconstruction approaches can be split into single-level fitting and multi-level fitting procedures. Both of these procedures can be zone by zone or multizone hence the link between single-level fitting and multi-level fitting.

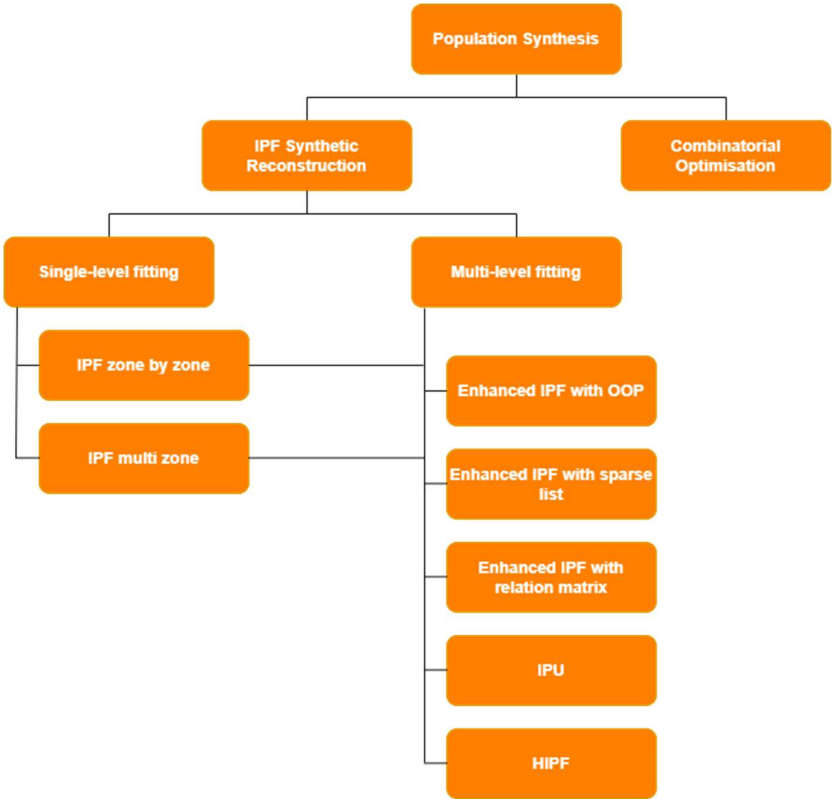

Figure 7 Main methods for population synthesis as given by Lim (2020)

IPF zone by zone and IPF multizone

IPF uses spatial hierarchy geography to have interrelated hierarchies of zones and regions. The data sources used for IPF often have different spatial resolutions even though the geographical classification is the same. The classic IPF uses the Deming-Stephan algorithm, which can process only one geographical zone at a time and is therefore termed as IPF zone-by-zone. The two-step IPF algorithm used by Beckman, Baggerly, and McKay (1996) can fit all geographical zones simultaneously by adding a zone dimension to each control marginal total and is thereby known as IPF multizone (Müller & Axhausen, 2010). Pritchard and Miller (2012) have compared these two approaches and found that the multizone approach had better fit but additional computational resources were needed. Predominantly, population synthesizers use either the zone-by-zone or multizone approach in the fitting stage, but they are constrained to one specific geographical resolution.

The classic IPF is referred to as single-level fitting as it is only able to adhere to constraints at the household level or individual level at a time. IPF procedures that can process constraints at the household level and individual level simultaneously, use multilevel fitting. It is important to note that there are interdependencies between individuals and households and that these influence the decision-making process of the travel demand. The individual-level attributes correlate with the household level attributes and the true population may deviate from the synthesized results if this is not accounted for (such as in single-level fitting) (Lim, 2020).

Enhanced IPF with Object-oriented programming

Object-oriented programming IPF (OOP IPF) was proposed by Guo and Bhat (2007). This recursive IPF procedure makes it possible to combine two contingency tables with the same variables during the iterative process and therefore controls statistical distributions at household and individual levels. At random, individuals are selected from the seed data and will be added to the synthetic population if this addition satisfies a pre-defined threshold value for its homogeneous group. Values in the contingency tables are iteratively filled in, based on the reduced desirability of selecting a certain household and individuals belonging to the household. The synthesized individuals comply with constraints on the household and individual levels while the households only comply with the household constraints. This is more efficient and helps to monitor the number of households and persons in the same homogeneous group. This was applied in an operational software system named Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns (CEMDAP) (Guo & Bhat, 2007). The population synthesizer requires census data. CEMDAP uses a tolerance that cannot be violated to ensure that target values of household and individual level tables are achieved (Choupani & Mamdoohi, 2016).

Enhanced IPF with a sparse list

Sparse list-based IPF by Pritchard and Miller (2012). This is an addition to the classic IPF by using a sparse list structure to handle more control variables and to generate household-person relationships. This procedure is more efficient and requires less computer storage. It also allows data aggregations and different levels and easily links data to different sources. The multidimensional contingency table is constructed from many records with household and individual attributes taken from unit records in the reference sample. A weight is assigned to each record, and these are then used as expansion factors to produce the unit records to a full population. This was implemented in the Integrated Land Use, Transportation and Environment model (ILUTE). The reference sample file consists of Canadian census data. A major problem with this data is that it does not have links between households and persons so the distributions at household and individual levels had to be estimated and fitted against each other to be consistent. A Conditional Monte Carlo Simulation is used at the generation stage to allocate individuals to households (Lim, 2020).

Enhanced IPF with relation matrix

Multilevel fitting with relation matrices developed by Arentze et al. (2007). The relation matrices define the distribution of households over the attributes of household members. The relation matrices are used to transform marginal distributions of persons to marginal distributions of households for a chosen set of control variables. This relation matrix is an altered contingency table obtained from seed data (disaggregate sample data). The cell values of the table are filled in by allocating individuals to household positions for each predefined household structure type to such a degree that the distributions of individuals are in accordance with the distributions of households. The relation matrices and aggregate constraints from known demographic data are then used to perform the IPF procedure on the relation matrices. The result of this is household distributions which are then applied as aggregate constraints for a second IPF procedure on the original seed data at the household level (Müller, 2017 as cited in Lim, 2020). This (two-step IPF) procedure ensures consistencies between person and household level counts, but the results of the household synthesis are not connected to the synthesis of personal data (Pritchard & Miller, 2012). This procedure was applied in a rule-based and activity-based model of travel demand that is named A Learning-based Transportation Oriented Simulation System (ALBATROSS). The Dutch Travel Survey (OVG) and

demographic data were used as the sample data and for the construction of relation matrices for age group and work status (Choupani & Mamdoohi, 2016).

Iterative Proportional Updating (IPU)

The Iterative Proportional Updating (IPU) by Ye et al. (2009). The IPU can carry out iterations with the distributions at the household and individual levels simultaneously. The mechanism of IPU adjusts household weights to the extent that household and individual level distributions can be best matched. When using IPF to get distributions at the individual level, the procedure would have to be done twice. It is first applied at the household level and then to the individual level and this would result in two separate and independent sets of weights (Lim & Gargett, 2013). The IPU algorithm was used in a standalone and open-source population synthesizer named PopGen. It was part of the integrated modelling system Simulator of Transport, Routes, Activities, Emissions and Land model (SimTravel) by the Arizona State University. The input for this model is census data in the US (Ye, Konduri, Ram, Sana, & Waddell, 2009). Another standalone open-source software is PopSynWin developed by Auld, Mohammadian, and Wies (2009). This software was implemented to gain synthetic populations for the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model (Auld & Mohammadian, 2010). PopGen and PopSynWin have in common that they were specifically developed for census data in the US and they both tend to underestimate the generated persons (Lim, 2020).

Hierarchical IPF (HIPF)

The Hierarchical Iterative Proportional Fitting (HIPF) by Müller and Axhausen (2011). This multi-fitting algorithm performs IPF on household and individual levels by initiating an entropy optimizing fitting step to alternate between the two levels. In the entropy optimizing step, an entropy function is defined that can simultaneously match the household distributions and the individual distributions. This algorithm was used for population synthesis in Switzerland and was compared to Swiss census data. It was found that the run time and convergence are similar to other approaches, and it outperforms IPU based on the analysis of the goodness of fit of the synthetic population. The algorithm can be adapted for usage in other populations (Müller & Axhausen, 2011). A similar approach was implemented by Bar-Gera et al. (2009). Both studies use an entropy-based model where a weight is attached to each household that is later determined and the entropy function is defined in terms of these weights (Choupani & Mamdoohi, 2016).

In literature, there is still limited research that compares these variations of the IPF procedure to each other. This is because the variations were developed in models to cope with the limitations from classic IPF procedure and researchers tend to 'start from scratch'. This leads to plethora of expedient variations and little information on the relative benefits of the various techniques. And even though population synthesizers have been stated mathematically well, written scripts in current programming languages are still scarce (Lovelace, Birkin, Ballas, & van Leeuwen, 2015). This also makes it hard for researchers to easily set them side by side.

Variations of the IPF procedure seek to do fitting for multiple zones at a time, capture interactions between household level attributes and individual level attributes and fitting simultaneously between these levels, increase the computational efficiency by using sparse lists or relation matrices or an entropy optimizing step. All methods form candidate methods for this research and which method is best to use depends on the application, level of detail and the available data. Picking a suitable IPF procedure should also be a step to add in the methodology and will help to answer question 1a and 1b.

Furthermore, to allocate households to houses, the population should have household characteristics that can be linked to attributes of houses or residential units. This implies that for this research, it would not be suitable to do the population synthesis at the level of individuals only. Therefore, the fitting should be done at the household level only or a multilevel approach can be used in which fitting is done at the household and individual level.

2.1.5 VALIDATION OF IPF

Validation important for every developed method of population synthesis as it helps to analyze the appropriateness and usefulness of the developed method and can give information about biases. Validation of the IPF procedure remains a difficult task. The population synthesis output is often individual-level and detailed. So, to validate this, such disaggregate microdata must be available for small geographies. However, if this data were available, the population synthesis would serve no purpose (Lovelace, Ballas, & Watson, 2014). There are still techniques that can be applied to overcome this. There are two types of validation:

- *Internal validation*
  This entails comparing the aggregate constraint variables with the aggregated results of the population synthesis using the same variables (Lovelace, Ballas, & Watson, A spatial Microsimulation Approach for the Analysis of Commuter Patterns: from Individual to Regional Levels, 2014).
- *External validation*
  This shows how well the synthesized population fits the true population by using other variables than the aggregate constraint variables. It illustrates whether the synthesized population is reliable (Choupani & Mamdoohi, 2016). This often requires the use of external data.

Internal validation

The purpose of internal validation is to find the magnitude of the errors introduced by the IPF procedure in its two stages (Choupani & Mamdoohi, 2016). Because the IPF always converges towards the optimal result of the known control variables when the zero-cell problem does not occur, internal validation is often seen as less important (Lovelace, Ballas, & Watson, 2014). For internal validation, categories and subcategories of the marginals, cells, and tables are compared to corresponding estimates in other corresponding tables, cells, and marginals. When using multilevel fitting, multiple tables are also checked. This is done by constructing a vector that consists of cells of both household and individual levels. This is then compared to its counterpart (Choupani & Mamdoohi, 2016).

External validation

For external validation, four methods were identified by Lovelace, Ballas, and Watson (2014) for validating the results of the IPF procedure with external data:

1. Use real spatial microdata as a comparison to the synthesized population data.
2. Use surveys to collect primary data for the study area and then test the synthesized population with this.
3. Make comparisons on aggregate levels with the synthesized data and an external data set (data that was not used for the population synthesis as seed data or marginal totals).

4. Sum and accumulate the small area synthesized population to a larger area population and then compare the results with real data from higher geographies (Lovelace, Ballas, & Watson, 2014).

The errors measured through external validations are small as long as the distribution of the uncontrolled variables remains the same throughout the sample and each zone. There should also be a strong correlation between the controlled and uncontrolled variables (Voas and Williamson, 2000 as cited in Choupani & Mamdoohi, 2016).

Validating population synthesis is also referred to as a non-trivial problem. It is comparable to the validation of the goodness of fit for a model for high dimensional probability distributions. This is an active research topic in statistics. The biggest issue to overcome is that usual indicators like the root mean square error do not give much insight when it is analysed across multiple dimensions and cannot indicate how well the model is performing at the cell level (Rich, 2018). The following indicators were found in literature:

- Uncertainty in the household simulation stage and uncertainty in the final output (Rich, 2018).
- The prediction performance at zone level when simulating for 5 years ahead from the base year (Rich, 2018).
- A δ-value that indicates the average absolute deviation between the synthesized weighted sample and the marginal totals (Lim & Gargett, 2013).
- Comparison of the distributions of the household and individual level attributes of the synthesized population and the true population (Lim & Gargett, 2013).
- The household and person-level attributes were also benchmarked against the actual number. The distributions of these attributes were compared to the actual data (Lim & Gargett, 2013).
- The coefficient of determination (R^2-values) between the controlled and uncontrolled variables (Lovelace, Ballas, & Watson, 2014).

According to Choupani and Mamdoohi (2016), the majority of population synthesizers that they have evaluated focus more on internal validation. Internal validation leads to a lot fewer errors than that of external validation. It was therefore also stated that there is a lack of literature that gives a validation framework for IPF focusing on the different stages (fitting and allocation), spatial units, zero-cell problem, and selection stages.

Research question 1d focuses on the validation of the generated population. And this section has provided ways in doing so for internal as well as external validation. Since a validation framework is yet to be established for all the stages of population synthesis and the IPF procedure has not been fully validated yet, it remains relevant to include a validation step in the methodology developed for this research.

2.1.6 IMPLEMENTATION DETAILS

Now that the population synthesis techniques, input data, control variables and validation have been explored, implementation details were sought. In literature, there seems to be a lack of an outlined stepwise methodology that focuses on the IPF algorithm and can provide researchers with tools and tips on how to best generate a synthetic population. The zero-cell problem, scalability, and 'integerisation' are all named to be aspects to consider but when implementing the procedure other aspects also arise that are not addressed directly or at all in literature. These include:

- Details on how to specifically prepare data sets for the IPF procedure. The more control variables there are, the more challenging this becomes.
- Harmonizing aggregate data and disaggregate sample data. These two data sets are not always collected in the same manner and the variables in the data are also not defined in the same way either, so it becomes important how to cope with this. Also, the format of the data plays an important role when synthesizing with more control variables.

- What to consider when choosing appropriate control variables and categories for these control variables. Auld, Mohammadian and Wies (2009) do mention in their research that to mitigate the zero-cell problem, a reduction in categories and increase in control variables is advised and that their population synthesizer (PopSynWin) does this through a formula. The downside is that this does not lead to a better synthesis as households are generated using fewer data and this also leads to a coarser depiction of household data (Choupani & Mamdoohi, 2016).
- In cases where the only data that is available does not have the spatial granularity needed and how to cope with these instances.

Non-scientific literature (forums, blogs, websites) does provide additional information about the implementation of the IPF procedure. Hunsinger (2008) outlined in a document how the IPF procedure works and has visualized this to make the format of the data used clearer. He also discussed how the control variables dictate the dimensions needed for the data sets in the IPF procedure and has specified this for a two-dimensional, three-dimensional, and four-dimensional IPF (single-level fitting) procedure. Based on the example given by Hunsinger (2008),

Forthomme and Ballis (2021) wrote a script in Python using the ipfn package that Python offers and outlines that there are two versions for the procedure. The first one is a Numpy version which is the quickest approach, and the second version is a Panda version. The Panda version is much slower but is easier to use. The script also helps to understand the data format and the working of the IPF procedure. For this research, the ipfn package with the Numpy version will be used in the case study and the examples given by Hunsinger (2008) and Forthomme and Ballis (2021) will serve as the basis for the implementation.

2.2 OSM DATA

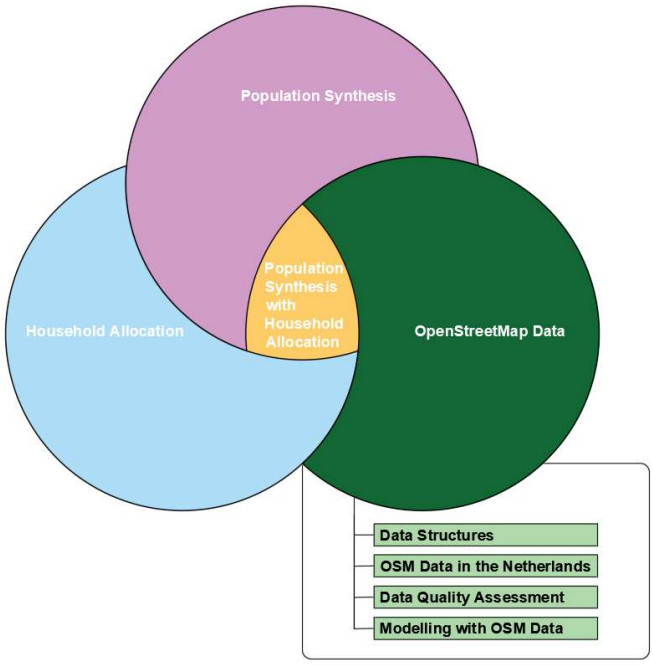

Figure 8 OSM data subthemes

The focus of this section will be OpenStreetMap data and what the data structures look like. The quality of OSM data in the Netherlands was also explored and methods to assess the data quality of OSM were found. Lastly, the existing studies where OSM data has been used in population synthesis or transport modelling are discussed. This overview is shown in Figure 8.

For the intended usage of OSM data in this research, it is important that in terms of quality, that the locations of houses and residential units is accurately specified. It is also important that these houses and residential units are distinguished as individual entities and have attributes specified such as the usage, address, surface area and number of floors.

2.2.1 DATA STRUCTURES

In OSM, community members can specify road networks, buildings, parks and more to map the world. Data structures are needed to specify these elements. OpenStreetMap uses basic data structures such as nodes, ways, and relations. Nodes are points that are specified in space, ways specify the linear features and area boundaries, and relations specify how other elements are connected.

Tags are attached to these structures to represent physical features on the ground like roads and buildings. Each tag consists of a key and a value. An unlimited number of tags can be used to describe a feature. The community decides the key and value combinations and more tags can be created to improve the style and enable better analysis of the map over time (OpenStreetMap Wiki, 2021). The tags that can be useful for carrying out population synthesis (if they are available) are shown in Table 2. Only a subset of the values will be specified to give an idea of what can be specified within a key that can be used in population synthesis.

Table 2 Example of useful tags in OSM for household allocation

Key	Values
Amenity	Restaurant, college, kindergarten, library, charging_station, fuel, parking, bank, dentist, hospital, pharmacy, nursing_home, cinema, courthouse, fire_station, police, post_office, etc.
Building	Apartments, farm, hotel, house, houseboat, residential, commercial, industrial, kiosk, office, retail, supermarket, warehouse, church, government, public, hospital, train_station, barn, school, college, university, stadium, parking, etc.
Height	Number (height of a building)
Building:flats	Number (of residential units)
Building:levels	Number (of floors/levels)
Addr:housenumber	User-defined (house number may contain letters, dashes, or other characters.
Addr:flats	Number (of flats or apartments located behind a single entrance door)
Addr:postcode	User-defined (postal code of the building)

2.2.2 OSM DATA IN THE NETHERLANDS

Since data in OSM is collected by volunteers, the data is very heterogeneous and may provide different levels of accuracy and completeness depending on the country or city. In Roick et al. (2011), the data quality of OSM in The Netherlands was assessed and compared to Germany, Spain, Portugal, and Poland. It was concluded that The Netherlands seemed to be mapped almost completely and that the sum of features and number of attributes were the highest compared to the other countries. Furthermore, these numbers seem to be consistent throughout the whole country. The average number of objects modified per user is also amongst the highest in Europe. It was concluded that the buildings are mapped almost completely throughout the country as well. However, the completeness of the map in The Netherlands is not due to its active community but rather due to data imports covering the whole country (the Automotive Navigation Data import) (Roick, Hagenauer, & Zipf, 2011).

These findings indicate that the quality of OSM data is reasonably well and can be used as a source for population synthesis because the buildings and road network are mapped to a great level of detail. Both studies are from 2011, so the expectation is that the data would have gotten better over the years.

The accuracy of GPS receivers is usually 6-10 meters from the true location, and it is estimated that with OSM the accuracy is around 20 meters from the true location (Haklay, 2010). But if data is imported from validated sources, the accuracy in OSM will be equal to the accuracy of the imported data. This relates to geometric and positional accuracy, the thematic accuracy might not be as good and should be assessed for the Netherlands.

2.2.3 QUALITY ASSESSMENT OF OSM DATA

Maps are only useful if they are reliable, complete, and accurate. Through OSM, geographical data is readily available online but there is still a question about how reliable this data is. To answer this, over the years many researchers have focused on this topic and explored the quality of OSM data using different indicators. It was found that often data in OSM is comparable or even better than conventional mapping software (Kounadi, 2009). The important indicators that could be used to assess the completeness and accuracy of OSM data are:

- *Completeness:* this is done by calculating the length of every attribute and summing this in OSM. The same is done for validated geographic data and then the completeness can be calculated in a percentage using ($length\ completeness = 100 \times \frac{OSM\ lengt}{Validated\ lengt}$). Name completeness could also be checked similarly by comparing the number of road names in OSM with road names in a validated geodata set.
- *Thematic accuracy:* this indicator gives the percentage of attributes that are correctly classified. To do this, OSM is divided into grid squares of 1 km^2. For every grid, it was analysed whether the OSM type matches the type of attribute in a validated data set. Then the length of the correct attributes was calculated and divided with the total length multiplied by 100 to get the percentage.
- *Positional accuracy:* this can be done through a buffer analysis. During the analysis lines and boundaries are identified in a validated data set and compared to where they are in OSM. There is a buffer zone created around the reference line. The percentage of each line/boundary that overlaps with the buffer zone of the reference line from the validated data set is calculated (Kounadi, 2009; Haklay, 2010). There is great positional accuracy if the deviation is no more than 1-2 meters off from the true location (Haklay, 2010).

It is also important to note that the OSM community has a variety of tools for quality assurance at their disposal when they are mapping in OSM. These tools report bugs and errors and help track and visualize the added data in OSM. When working with OSM data it is helpful to check whether there are reported errors in the area of interest through the use of these tools. Some of the tools also give an overview of areas in OSM that are mapped in detail. A few of these tools and (some of) their functions are:

- *OSMantic:* this tool helps with the tags in OSM by suggesting relevant tags for mag features. It also has a system for reporting bugs and labelling them once they are fixed.
- *Osmose:* this carries out data consistency checks.
- *iOSMAnalyzer:* this focuses completeness of OSM features and tags, how recent the data is, the positional accuracy of features. It also checks the user profiles and activity and assists in geocoding (adding addresses, postal codes, and house numbers). It checks the logical consistency and geometry of polygons and helps with the development of POIs.
- *OSM inspector:* this is a tool for error debugging. When tags or not filled in or incorrectly filled in, the tool indicates this with a "FIXME" tag (Almendros-Jiménez & Becerra-Terón, 2018).

The tagging quality is also of importance in OSM because the open tagging system in OSM can lead to many features being wrongly or incompletely classified. The tags that are used in OSM to provide thematic information about OSM entities are referred to as folksonomy. Various websites list all these tags and their statistics. Examples of such websites are TagInfo and TagFinder (Almendros-Jiménez & Becerra-Terón,

2018). TagInfo can give the most popular keys, a combination of keys, and the most common values for keys. In a paper by Almendros-Jiménez & Becerra-Terón (2018), TagInfo was used to assess the quality of tagging in OSM. They developed a webtool named QXOSM that can carry out the analysis of tagging quality for any area in OSM. The webtool can be accessed through the link http://xosm.ual.es:8080/qxosm. The analysis uses TagInfo as a reference and several quality indicators. Accuracy is not included in these indicators. The indicators are summed and explained in Appendix B.

2.2.4 MODELLING WITH OSM DATA

Throughout travel demand modelling literature, the use of OpenStreetMap data as a source is still in its infancy. In previous years, a few studies have looked into the quality of OSM data in general concerning road networks. Girres and Touya (2010) found that OSM data is very good in terms of responsiveness and flexibility in France. They have also pointed out that the heterogeneity of OSM data greatly limits the possible application. For Germany, Arsanjani et al. (2015) explored the process of contribution in OSM in the spatial and temporal realms for the years 2007 and 2012. They showed that once the basics of mapping are in place, a densification process starts. This means that the quality of the OSM data improves over time. At the start of OSM in 2004, the areas that were mapped did not contain the level of details they contain today (Arsanjani, Helbich, Bakillah, & Loos, 2015).

Another study by Mashhadi et al. (2012, as cited in Briem et al., 2019) analysed the quality of Points Of Interest (POIs) in OSM data by setting it side by side with commercial data from Navteq and Yelp in London (UK) and Rome (Italy). They concluded that for urban areas, the accuracy through the geographic position of POIs is very high. This was supported by Neis et al. (2012) and they showed that in densely populated urban areas, OSM data can be alternative to commercial data sets. Nevertheless, the quality of the OSM network was lacking in rural areas.

The geographical data stored in OSM can also be used for the setup of agent-based travel simulations. The data gives information about the size, locations of elements, and details about the shops, offices, or companies. These details can help provide attributes for destinations in destination choice models (Zilske et al., 2011, as cited in Briem et al., 2019). It should be noted that data from online sources have the tendency to be incomplete and may require synthetic methods to generate missing data. A possible synthetic way is to do this is to approximate the number of employees, students, or persons in a building by making inferences about the size of the building including all floors of the building (Briem, Heilig, Klinkhardt, & Vortisch, 2019).

When looking at the applicability of OSM data in the context of spatial allocation of population synthesis, there was little found in literature. Long & Shen (2015) discussed a method where OSM is used for quick and robust delineation of parcels and therefore giving the basic spatial units for allocating populations at a fine scale. They indicate that the POIs that are available in most online mapping services, can be coupled with OSM to map the population in high resolution. The road network provides the identification and delineation of parcel geometries, and the crowd-sourced POIs are used to distinguish urban parcels with a vector cellular automata model. Housing-related online check-in records or POIs are then used for filtering the residential parcels from all identified urban parcels. Then population census data and residential POI density are used for the population synthesis (Long & Shen, 2015).

Another approach was described by Balac & Hörl (2021) where multiple sources including OSM data were used to create a synthetic population. They used OSM data to obtain the location of residential, work, shopping, or leisure places. The first step is to create the synthetic population is to create synthetic persons and households and through IPF attach socio-demographic and mobility tool ownership attributes. Then daily activity chains are attached to individuals by using data from household travel surveys and hot-deck matching. Afterwards, the home location has to be assigned to every household. This is done using OSM

data to get the locations and then through random sampling the households are assigned to their locations. The locations for work, education and, non-mandatory activities are assigned using household travel surveys and community surveys that contain information about commuting patterns and commuting distance. the simulation can then be run and this results in an OD-matrix for the synthetic population (Balac & Hörl, 2021).

2.2.5 CONCLUSION

From the literature study, it can be concluded that OSM data is a viable open-source data source. Several studies have shown that the accuracy and completeness of OSM are sufficient to be used as a geographic data source. Furthermore, two studies have focused on using OSM data to provide spatial details for allocating the population (Long & Shen, 2015) or identifying locations for activities such as work, leisure, and education (Balac & Hörl, 2021) and have demonstrated that OSM is suited for these purposes. Tools that ensure the quality of OSM data have also been explored. The Netherlands itself seems to be mapped out almost completely due to data imports from validated geographic data (Automotive Navigation Data) so it does not seem fitting to check positional, geometric accuracy, or length completeness for OSM data. Buildings and streets will be mapped with accuracy compared to conventional GPS receivers in The Netherlands. However, the same cannot be said for the thematic accuracy of entities in OSM. Therefore, it does bode well to check using indicators introduced in this chapter or the webtool developed by Almendros-Jiménez & Becerra-Terón (2018) for this research. This helps to answer research question 2c.

2.3 CANDIDATE METHODS FOR HOUSEHOLD ALLOCATION

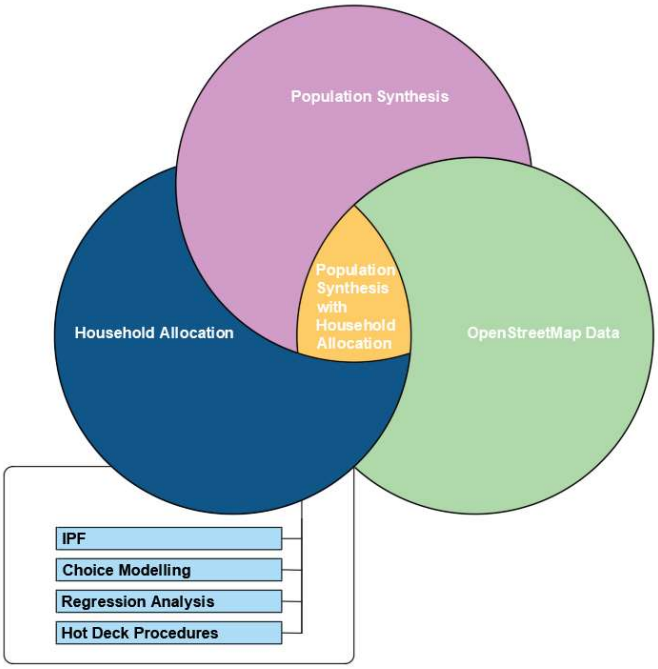

Figure 9 Household allocation subthemes

This section focuses on the household allocation and consists of four subthemes as shown in Figure 9. The purpose of this section is to find candidate methods for the household allocation and answer research question 1c.

When the population synthesis has been carried out, the households or agents with their respective attributes are generated to meet the marginals of a certain geographical area. OpenStreetMap data can outline the houses or residential units that exist in a geographical area. The question then becomes how the generated households can be linked to the houses in an area in an unarbitrary manner. To do this, there are several statistical methods available. Some of the candidate methods are:

- *IPF procedure*
  This would entail that attributes of the house are added as control variables and when generating the population, the households will immediately be assigned to a house as well. The problem is that this would require a data set that contains all household and house attributes and how these are observed together.
- *Choice modelling*
  In this case, the households are allocated based on revealed or stated preferences. The preferences are used to adjust the taste parameters and the household and house variables are included in the alternatives. The utility for each alternative is calculated and the decision rule (random utility maximization, random regret minimization, or taboo aversion) dictates to which house a household is allocated. The residential location choice in Frenkel et al. (2013) was modelled similarly.
- *Regression analysis*
  With regression analysis, the regression coefficients can also be estimated by using revealed or stated preference data and then this leads to establishing relationships between the variables and can predict in which houses households with certain characteristics will reside. The Hedonic price method by Rosen (1974) is a form of regression analysis and is focused on what individuals are willing to pay for different attributes of a house and its surroundings (van Duijn & Rouwendal, 2012).
- *Hot deck procedures (Statistical matching)*
  This is a form of statistical matching and involves imputing missing values. This is used when variables needed for modelling are not jointly observed in the same data set. A donor and recipient data set are then used to match observations by considering the common variables (also called the matching variables) between the data sets (D'Orazio, 2017). This method was used to assign activity chains to individuals in research by Balac & Hörl (2021). The approach could potentially also be used to assign households to houses.

Each of these methods have advantages and disadvantages. These have already been discussed for the IPF procedure in Section 2.1. Choice modelling introduces assumptions that must hold for implementation. The first of these assumptions is that the random components of the utility functions are independently and identically distributed (I.I.D.). Another assumption is that the choices being made satisfy the independence of irrelevant alternatives (I.I.A.) property (Navrud & Bråten, 2007). These assumptions can limit the use when they do not hold, this forms a disadvantage. A benefit of choice modelling is that it more accurately models consumer behaviour and gives insight into the implicit trade-offs people make. Another benefit is that choice modelling is an appropriate method for situations involving ethical or moral considerations. It also allows for disaggregating the utility associated with particular goods and makes extrapolation possible (Rolfe & Bennett, 1996).

Like choice modelling, regression analysis also has model assumptions that have to be enforced. All assumptions are pertaining the error terms. The first assumption is that the error terms are independent of each other (independence). The second assumption is that the error terms are normally distributed (normality). The third assumption is that the error terms have equal variance (homoscedasticity). If it concerns linear regression, then there is also an assumption of linearity between the dependent variables and independent variables. These assumptions can also be seen as disadvantages of this method. Regression is also susceptible to noise and overfitting and sensitive to outliers. There is also the issue of multicollinearity, and this occurs when predictors are highly correlated. This technique also has many

advantages including being easily interpretable and implementable, being able to cope with a small sample size and relatively weak signal and it allows for extrapolation beyond the data set (Su, Yan, & Tsai, 2012).

Hot deck procedures, like choice modelling, also have the limiting I.I.D. assumption. This is difficult to preserve when matching data from complex sample surveys with more than one stage of selection of sample units. This forms a disadvantage. Another model assumption made in statistical matching is conditional independence (C.I.) of the target variables (the variables that are distinctly observed in the donor and recipient data set) given the common variables. This assumption rarely holds in practice. Another disadvantage is that to have sufficient accuracy and consistency, many data requirements must be met including data sets having the same data collection technique and the same definition of common variables. The uncertainty associated with statistical matching is also an important aspect and reducing this accuracy still requires more research (Donatiello, et al., 2014). This technique is significantly less researched than the other proposed methods which can also be seen as a disadvantage. This method does have the benefit that it is able to create synthetic data and input values in data sets where necessary variables are not jointly observed. It can also aid in data sets where the necessary variables are present but not reliable. Statistical matching can then be used to replace these values with values from a more reliable data set (D'Orazio, 2017).

Whether these methods can be used, is ultimately dependent on the available data, the variability in the data and the desired goodness of fit. The suggested methods have their advantages and disadvantages. Some of the methods are better in handling great variability (e.g. hot deck procedures) and are flexible while others methods require more data. This makes choosing a method case dependent. The researcher essentially has the freedom to use any of these methods given that it is properly implemented. The research methodology will be designed in such a way that any of these methods could be used.

2.4 SUMMARY

In this section, the focus will be on which research gaps were apparent after doing the literature study and the main findings.

2.4.1 RESULTING RESEARCH GAPS

From the reviewed literature, gaps were identified that require further research and would be beneficial to population synthesis and OSM literature. These gaps are:

1. Population synthesizers using OSM data (or other open-source VGI data) are scant.
2. There is a lack of literature that specifies the population synthesis framework and all its stages (Rich, 2018).
3. The vast majority of IPF based synthesizers have trouble to converge when four or more control variables are used (Choupani & Mamdoohi, 2016).
4. Most of the current population synthesizers are concealed in computer codes and inaccessible language which causes a lack of transparency (Lim, 2020). This makes proper comparison between different population synthesis techniques difficult as well.
5. Several of the population synthesizers also lack implementation details and have validation issues which leads to problems concerning the reuse of such synthesizers (Choupani & Mamdoohi, 2016).
6. No literature proposes a well-established validation framework for IPF. There has been no evaluation of fitting, spatial units, integer conversion and selection stages according to Choupani and Mamdoohi (2016).

7. Even though the zero-cell problem has been the topic of a handful of research, there still is no unbiased technique to resolve the problem (Choupani & Mamdoohi, 2016).

This research will focus on research gaps 1, 2 and 4. Research gap 5 will also be addressed partially. By providing a method for spatial distribution in population synthesis with OpenStreetMap, this research can add to the literature and explore how much potential OSM data has in the field of spatial microsimulation and thus addresses research gap 1. By developing this methodology and implementing it in a case study, research gaps 2 and 4 can be bridged. Through the case study, implementation details can be formulated, and this helps to address research gap 5. Although, there are many applications for population synthesis, the research will be carried out more from a transport perspective. This precludes research gaps 3, and 7 as these are more focused on mathematics and computational efficiency. Research gap 6 is also not focused on as this deviates from the research objective.

2.4.2 SUMMARY

In this chapter, the various population techniques have been described and compared to each other to the extent that was possible given the current literature. From this comparison, it was concluded that Synthetic reconstruction methods are much better researched and used in models. Consequently, this also makes synthetic reconstruction methods more desirable for this research.

Furthermore, the variations of the IPF procedure used in synthetic reconstruction have been discussed and it became evident that there are not many studies that compare the variations to each other. This makes it difficult to choose the approach upfront. However, the availability of data, desired level of detail and model purpose can help to narrow down the options and ultimately make a choice.

For this research, it is only possible to choose population synthesis with single-level fitting at the level of households or multilevel fitting with both individual and household levels. The reason for this is that the population needs household characteristics for the household allocation.

The reviewed literature on OSM data indicated that OSM data is of reasonable quality specifically in the Netherlands. However, because this is largely imported data, the thematic accuracy should still be analysed. Tools to assess the accuracy and completeness were also mentioned and explained. Moreover, two studies were presented in which OSM data was used and proved that OSM data has potential to serve as a data source.

Four candidate methods were proposed for the household allocation. The choice for the best method here also depends on the available data and the desired goodness of fit. And just like with the variation of the chosen IPF procedure, the chosen household allocation method will also be included in the methodology and will be designed that it can accommodate all candidate methods.

The literature review also aided in identifying steps that must be part of the methodology. These steps include finding input data, picking the IPF type and the control variables, validation of the generated synthetic population and OSM data quality assessment and the choice for the household allocation.

3. METHODOLOGY

This chapter outlines the proposed framework for population synthesis with OSM data to add spatial units. This framework is developed through literature and by using the case study for implementation. The methodology developed will be encompassing both single-level fitting and multilevel fitting and different house allocation methods. It can serve as a helpful guide on how to generate households and assign these households to houses. It consists of several steps and each of the steps will be elaborated. The focus, in the beginning, is to generate the population and at the end, the focus is to attach the population to houses.

The literature review provided components that have to be part of the methodology. These were input data, IPF type, control variables, validation for population synthesis. The components for OSM were information stored in OSM and data quality assessment. For the household allocation, the components were choosing a method and variables for allocation. For all these components, there was no indication of which order these should be carried out and no instructions on pre-processing of data and proper description of the data to be used. Therefore, the entire methodology presented here is a contribution as it places all components found in literature in a specific order, adds steps for data preparation, gives description of the data and sheds light on implementation details. These implementation details are more directed to implementation in Python. Explanations for implementations may contain specific technical detail but this is to enhance transparency and reproducibility of the method.

The colour scheme of Figure 4 representing the subthemes is carried through in the methodology given in Figure 10. The population synthesis is marked in purple, the OSM data is marked in green, and the household allocation is marked in blue. Validation is orange as it concerns the validation of both the population synthesis and household allocation. At each step, there is a choice to be made and this choice is dependent on the research goal.

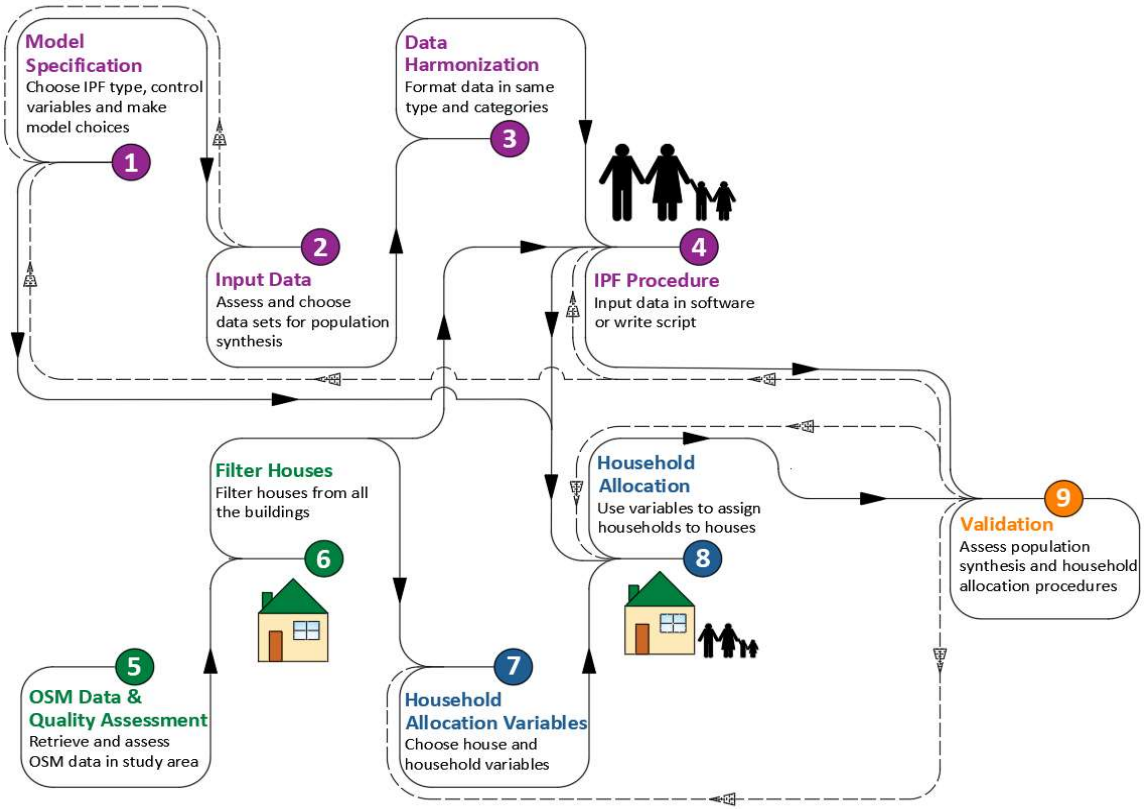

Figure 10 Developed methodology

The solid lines represent the input that is needed, and the dashed lines illustrate input that is fed back to steps that have already been carried out. Steps 1, 2, 3 and steps 5 and 6 can be carried out simultaneously. The first step results in the IPF type and control variables chosen. This then goes into the second step and outlines what disaggregate, and aggregate data is needed. If the necessary data sets are not available and that becomes apparent at step 2, then the researcher must go back to step 1 and change the control variables or make compromises on the accuracy. Step 3 concerns pre-processing of the data sets and the result of this are data sets that are ready for implementation in step 4. Step 4 is the IPF procedure itself and the programming (if applicable) and verification of the procedure.

The following step pertains OSM data and checking the quality of this data. Inaccuracies found in the data quality assessment may be adjusted. This then flows into step 6 where the focus is to filter houses and residential units from other buildings. This gives the total amount of houses and residential units that need to be assigned a household. This number of houses is also used as input for the IPF procedure as it dictates the number of households that need to be generated.

The seventh step is choosing the household allocation variables and this flows into the next step which is the house allocation procedure. Choices made in the model specification influence the household allocation procedure (simplifications and assumptions that have been made). Therefore, there is a connection between the first step and the eighth step. The generated population is also input for the household allocation hence the solid line from the IPF procedure to the household allocation. The result of this step is a population consisting of households attached to houses in OSM.

Step 9 concerns the validation of the synthesized population and the household allocation. The validation may lead to changes in model specification, the IPF procedure itself, the household allocation variables, and the household allocation procedure. As a consequence of this, there are feedback loops to these steps.

3.1 MODEL SPECIFICATION

The methodology starts with the model specification where choices such as IPF type, control variables and other choices such as simplifications are made. These choices will be further described in the following sections.

3.1.1 IPF TYPE

There are several IPF procedures as discussed in the literature study. If the goal of the research is to generate populations for more than one zone, then the IPF multizone might be ideal. The choice of single-level fitting or multilevel fitting should also be made here. This depends on the intended use of the generated population and the desired accuracy.

If the population should be accurate and needs agents and households along with the correlation between these levels in the simulation, then the choice falls on multilevel fitting. And if only agents or households are needed, the choice falls on single-level fitting. Depending on the size of the study area (thus the size of the population that needs to be generated) and whether it concerns multilevel fitting, the choice can be made for more efficient algorithms (such as the sparse-list IPF or HIPF) if this is required.

As stated by Lomax and Norman (2016), the choice for the method of synthesis is greatly influenced by the problem being researched, the preferences in place for the research and the resources in terms of time, software, and skills. The last part of this statement is also a constraint because if the researcher has limited time, software or skills, the synthesis procedure might have to be simplified (for example by using single-level fitting) and can lead to using software that impose biases and limitations. This is the case when using

PopGen and PopSynWin. These synthesis software can only handle census data properly and underestimates the generated agents at the individual level (Lim, 2020).

3.1.2 CONTROL VARIABLES

After the IPF type has been decided upon, the control variables can be chosen. This depends on the goal and context of the research. It might be helpful to first perform an analysis of what variables influence the behaviour that is being researched through the generated population. Statistical tests for correlations might be a tool to find relevant variables. Two examples will be given to illustrate how the choice for control variables is influenced by the intended model purpose.

For example, if it is desired to see the influence of different subsidies for public transportation on the trips per household, the sensitivity to these subsidies should be embedded in the household or rather the attributes of the household. This can be done by including the household income and the availability of motor vehicles and bicycles in the household.

If the synthesized population will be used in a transport model to for example research commuting patterns, then variables such as possession of driver's license or commuting distance (trip length) or the main mode to travel to work are useful to include as control variables. These would be control variables at the individual level. At the household level, it might be insightful to include control variables such as number of cars or bicycles within a household or the household income. Research by Lovelace et al. (2014) was done on commuter patterns and they suggested the control variables age, gender, mode, travel distance, employment relations and conditions of occupations (National Statistics Socio-economic Classification), household income, type of car and telecommuting potential. It is of essence that the control variables chosen influence the behaviour that is being researched and studies that explore these influences in literature can be a guidance to finding the probable control variables.

Since it is also known that when synthesizing with four or more control variables, the IPF procedure may have trouble converging (Choupani & Mamdoohi, 2016), it is advised to check if the IPF procedure chosen, can handle the number of control variables. If not, the number of control variables should be restricted to four. If there is a situation where it is necessary to add more control variables, each control variable can be added in a stepwise manner while checking for every additional variable whether the IPF converges. It might also be helpful to loosen the convergence criterion (the order at which the difference between the best solution and the estimates converges to zero) or tolerance rate (a specified value for the difference between two consecutive iterations). An alternative solution can be to restrict the number of categories for a control variable. Another option would be to use more efficient algorithms or software packages that can converge while using more control variables.

3.1.3 SIMPLIFICATIONS AND OTHER MODEL CHOICES

Lastly, other model choices should also be made at this stage. The model choices can include assumptions or simplifications that are introduced to make the procedure computationally more efficient or to simplify the problem. The model choices can also be the result of restrictions introduced by the data being used. Time resources and skills of the researcher may also elicit simplifications and assumptions.

3.2 INPUT DATA

The amount of control variables and levels (single or multilevel fitting) dictates the dimension of the IPF procedure. When using the single-level fitting, n-control variables will lead to n-dimensional IPF requiring (n-1)-dimensional marginals and n-dimensional seed data. So, for example, if there are four control

variables, then it concerns a four-dimensional IPF procedure, the marginals will have to be three-dimensional, and the seed data will have to be four-dimensional. Multi-dimensional data may be harder to obtain because data on aggregate levels is usually collected as one-dimensional. Microdata used as disaggregate data is often available as multiway tables. So, each addition of a control variable increases the data requirements. Each addition also requires the IPF procedure to iterate over an additional dimension increasing the computational effort.

The type of data that is most often used in IPF is population census data. This data should be able to provide the number of agents or households belonging to every homogeneous group. The homogeneous groups are all possible combinations of the categories of the control variables. The goal is to be able to make crosstabulations from this type of data. Multilevel fitting requires more data than single-level fitting because data is needed at both household and individual levels and the data must be collected in such a way that the relationship between the household variables and individual variables is captured.

The marginals and seed data need to have the same control variable i.e. the data for that variable needs to be collected in preferably the same manner and the control variable should have the same definition in both the marginal as the seed data set. If the control variable is categorical, the categories should also be the same in both data sets.

Data availability is one of the biggest constraints when using IPF. Sometimes, the data needed may just not be publicly available and belongs to a paid set. Other times, the data that is required is not available because it is not collected at the scale at which it is needed. Some ways to still acquire input data despite this, are:

- Consider purchasing the data if it is collected but not publicly available.
- Consider doing a survey in the study area that enables insight into the control variables. This is especially attractive if it concerns small geographies.
- If the previous options are too costly, data from zones that are similar to the study area can be used as a means of imputation. To do this, it must be assessed if the zones are alike in ways that are important for the research. This could be in terms of household types, activities, workforce, etc.
- Another alternative would be to use data of bigger geographies and scale them down to the study area. The assumption here would then have to be that the characteristics of the bigger geography have the same or similar distributions as the characteristics of the smaller geography.
- If there is some data available but the data set is not statistically significant, data from previous years can be used to enrich the data set.
- Another possibility if there is no data available for a certain control variable would be to use another control variable that is correlated with this certain control variable. For example, household size and household composition are variables that can be correlated. The same holds for the household income and car availability.

The quest to acquire input data may thus change the control variables used and even lead to control variables being left out and therefore changing the dimensions of the IPF procedure. As a consequence of this, there must be a feedback loop from this step to the previous step.

3.3 DATA HARMONIZATION

The next step is to put the data in the format that is needed for the IPF procedure. The format was discussed for single-level fitting in the previous paragraph. The control variables should be present in both the seed data and the marginal constraints with the same categories. The categories of the variables can be altered to make sure that each category (bin) has a proper size and thus enough observations.

Although, it should be stated that the more categories that are grouped, the more detail will be lost. So, at a certain stage, the researcher must decide if the categories can capture the differences in the population and that they have been sufficiently grouped. This is a trade-off between the number of categories and the detail of the generated population.

All the homogeneous cells that, after choosing the categories, still have zero observations should be altered. This is done by replacing the zeros with arbitrarily small values to avoid the zero-cell problem. How small these values must be are dependent on the values within the cells of the contingency table. It should be small enough that for marginals that have zero observations, the fitted seed data will also sum up to zero when rounded. There are other techniques to avoid this problem and it is ultimately up to the researcher's preference. It should be noted that there is no unbiased technique to avoid this zero-cell problem (Choupani & Mamdoohi, 2016).

When the data has been harmonized, the crosstabulations can be made. When doing this for the marginal totals, the sum of the row constraints must be equal to the sum of the column constraints in all dimensions. If this is not the case, the observations have to be reweighted by multiplying with a factor to correct for this. If the sum of the rows is not equal to the sum of the columns, the IPF procedure will not be able to converge (Lomax & Norman, 2016).

3.4 IPF PROCEDURE

Based on the IPF type chosen, at this stage, the researcher would either need to input all the data into software and then run the pre-programmed IPF procedure or the researcher would need to write a script for this using a programming console like R Studio, Jupyter Notebook, Spyder or many other available options. The researcher's own skills and knowledge play a big role in how efficient the script is. If the researcher is opting for programming the script, the pseudocode by Ye et al. (2016) may be helpful. This starts with the overall table (aggregate data with marginals) and sample table (disaggregate data) given in respectfully Figure 11 and Figure 12. The procedure itself is given in Figure 13.

Table: The overall table to be estimated

Attribute j / Attribute i	j = 1	…	j	…	j = s	Marginal sum
i = 1	m_{11}	…	m_{1j}	…	m_{1s}	$N_{1\cdot}$
…	…	…	…	…	…	…
i	m_{i1}	…	m_{ij}	…	m_{is}	$N_{i\cdot}$
…	…	…	…	…	…	…
i = r	m_{r1}	…	m_{rj}	…	m_{rs}	$N_{r\cdot}$
Marginal sum	$N_{\cdot 1}$	…	$N_{\cdot j}$	…	$N_{\cdot s}$	N

Figure 11 Overall table (Ye, Wang, Chen, Lin, & Wang, 2016)

Table: The sample table

Attribute j / Attribute i	j = 1	…	j	…	j = s	Sum
i = 1	n_{11}	…	n_{1j}	…	n_{1s}	$n_{1\cdot}$
…	…	…	…	…	…	…
i	n_{i1}	…	n_{ij}	…	n_{is}	$n_{i\cdot}$
…	…	…	…	…	…	…
i = r	n_{r1}	…	n_{rj}	…	n_{rs}	$n_{r\cdot}$
Sum	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot s}$	n

Figure 12 Sample table (Ye, Wang, Chen, Lin, & Wang, 2016)

Algorithm: Algorithm for Iterative proportional fitting

Input: Sample table and overall table with marginals (totals)

Output: Fitted overall table

1: Repeat

2: Update the elements in the overall table by row according to

$m'_{ij} = n_{ij}(^{m_{i\cdot}}/_{n_{i\cdot}})$ # value of m_{ij} after adjustment to rows

3: Update the elements in the overall table by column according to

$m''_{ij} = m'_{ij}(^{m_{\cdot j}}/_{m'_{\cdot j}})$ # value of m_{ij} after adjustment to rows and columns

4: Until iteration stops

Figure 13 IPF algorithm for single-level fitting (Ye, Wang, Chen, Lin, & Wang, 2016)

The aforementioned algorithm and tables are for single-level fitting approaches and concerns a two-dimensional IPF with attributes i and j (control variables). When adding more control variables, there will be additional steps after step 3 in the algorithm that will update the elements in the overall table by slices (with three control variables), stacks (with four control variables) and so on. In multilevel approaches, attributes are added at household and agent level along with an extra step that allows for alternating between the two levels.

Some consoles already have downloadable packages for the IPF procedure. An example of this is the ipfn package available for Python-based consoles. This package comes with documentation and examples to help program the IPF procedure. The documentation can be accessed through the web link https://pypi.org/project/ipfn/. It should be noted that this package is only for single-level fitting. An example of the usage of this package will be given in Chapter 4.

After running the procedure, the next step is to assess whether the algorithm was performed correctly through verification. This entails checking whether the marginals and seed data have been programmed correctly, whether the total number of observations from the marginals is equal to the total amount of agents or households generated. And lastly to check if the IPF has converged. If the IPF procedure has not converged, the researcher must check whether the sum of the rows is equal to the sum of the columns in each dimension. Another alternative can be to check if there are zeros that have not been replaced by an arbitrarily small value. The cause of not converging could also be that the marginals are wrongly programmed and thus do not constrain over the right dimensions.

3.5 OSM DATA & DATA QUALITY ASSESSMENT

This step concerns retrieval the OSM data for the study area. This can be done by delineating the study area in webtools, software or packages that enable modelling and analysis of OpenStreetMap data. The Python package named OSMnx was developed by Boeing (2017) for this purpose.

After retrieving the OSM data, this needs to be assessed in terms of quality. The aforementioned indicators for completeness, thematic accuracy and positional accuracy in Section 2.2.3 can be used to assess the OSM data. If it concerns imported geodata, then then validating the completeness and positional accuracy may not have added value as this geodata comes from a source that is already validated.

As part of assessing the thematic accuracy. The tagging quality can be analysed through the QXOSM webtool developed by Almendros-Jiménez & Becerra-Terón (2018) (http://xosm.ual.es:8080/qxosm/#!QXOSM). This tool analyses the tagging quality of OSM data will be analysed on completeness, compliance, consistency, granularity, richness and trust. High numbers of contributors, versions and confirmations are seen as positive indicators of trust and revisions and

corrections are seen as negative (Keßler & De Groot, 2013). TagInfo (https://taginfo.openstreetmap.org) can also be used for different analyses of the tags in OSM.

Significant errors in the study area can also be analysed by using the error detector tools available for OSM such as Osmose (http://osmose.openstreetmap.fr). These tools detect potential errors, inaccuracies, and sparsely mapped areas. Analysing these potential errors and assessing the impact that they can have on the research can also help with the validation process.

Another alternative would be to compare OSM data with other mapping platforms such as Google Maps or Google Earth and to analyse the differences. If there are significant differences, the mistakes in OSM data can be corrected (if needed) or assumptions can be made about the OSM data. Field research and observations can aid in validating the data as well.

3.6 FILTER HOUSES

In OSM data, the buildings can be filtered using tags. All the buildings that are labelled as houses or apartments should be retrieved for the study area. These houses will be used as the unit for distributing the households. When retrieving the houses, all the tags (such as levels, height, flats, address, house number, etc.) that are listed for the houses should be saved as well. These tags are needed because they give information about the houses and can therefore be seen as house variables that may be used in the next step to help assign households to houses. It also comes in handy to calculate the surface area of these houses and add that as a house variable as well if the area is not specified in a tag already.

3.7 HOUSEHOLD ALLOCATION VARIABLES

The variables that can be used to allocate houses to households are termed the household allocation variables. This includes household and house variables and what is especially important is the correlation that these household and house variables have with each other. The household variables are the control variables at the household level that were used for the population synthesis. If there is data available between agents and the house that they live in, then control variables at the agent level may be used as well. Albeit these relationships between agent variables and house variables are often difficult to find and, in some cases, not collected.

It is more common to find data on correlations between household characteristics and house characteristics. The house variables are variables that are available in OSM data such as surface area, location, height and levels. House variables may also be extracted from additional data sources such as house prices or property valuations (WOZ value). An example of house allocation variables are the household income, the housing costs and then the data available for the house price-to-income ratio. This ratio is used as indicator of overvaluation of housing costs and affordability of housing (Chen & Cheng, 2017). For the allocation, this same indicator can be used to see which households can afford which houses in the study area and then assign these houses to the households if it is affordable to them. If there is no data available for the study area, assumptions can be made, or the households can be randomly distributed over the houses.

3.8 HOUSEHOLD ALLOCATION

All of the methods discussed in the literature review form candidate models for this part of the framework when there is data available for attributes covering both households and houses. Each of the methods will seek to couple the household with its desired or most probable house attributes house based on the correlations between household and house attributes present in the input data.

Possible options when using the methods mentioned in Section 2.3 are:

- When using IPF, the attributes can be added as control variables and are part of the fitting process. Meaning that for every household generated in the IPF procedure, desired or probable attributes of the house are specified as well. Some examples of these attributes are living area, house type (single dwelling, rowhouse, duplex, flat, etc.) and property valuation.
- When using choice models, the utility of different houses is calculated and in the utility function household and house attributes are included. The taste parameters dictate whether the attribute introduces a disutility. A decision rule, such as random utility maximization, calculates the choice for the house for each household.
- The approach for regression analysis is like the approach for choice modelling. The household characteristics are then used as predictors to estimate characteristics of the house. For each household, a house attribute such as living area can be predicted and then this value can be used for the allocation.
- The hot deck procedure uses observed data that captures household variables and house variables to assign each household with the most probable house variables that would fit given the observed response from similar households in the data set that is used as donor.

After imputing probable or desired house attributes for the households generated in the population synthesis, the house variables need to be matched to the actual houses in OSM. Since the house variables used are also the same variables specified in the tags available in OSM for buildings, this can be done by making rules that the house matched with the household at least satisfies the specified house attributes determined by the methods listed above.

After the potential house is found, the household should be allocated to it. This can be done in two ways. Either the house gets an attribute named ‘household ID’ in which the unique household ID is imputed of the allocated house, or the household gets an attributed named ‘OSM ID’ in which the OSM ID of the matched house or residential building is given. After the house is allocated, it needs to be taken out of the set of available houses. This is similar to draws without replacement. The algorithm ends when all households have been allocated. In instances where there are no houses that match with the households, the next best option will be allocated. An example of this algorithm is described in pseudocode in Figure 14.

Algorithm: Proposed algorithm for matching in household allocation

Input: DataFrame with households and DataFrame with houses from OSM
Output: DataFrame with assigned households and DataFrame with assigned houses

```
For each unallocated household do:
    For each unallocated house do:
        If household attributes are smaller or equal to house attributes then:
        Make sorted DataFrame of houses in ascending order of house attributes
        Assign household to minimum item (house) of the sorted DataFrame
        Update status of house and household from unallocated to allocated
        Else:
        Make sorted DataFrame of all unallocated houses in ascending order
        Assign household to maximum item (house) of the sorted DataFrame
        Update status of house and household from unallocated to allocated
        End if
    End for
End for
```

Figure 14 Pseudocode for household allocation algorithm

3.9 VALIDATION

The validation of the IPF procedure and household allocation is described in this section. As mentioned in section 2.4.1, there is no well-established validation framework for IPF. The uncertainty associated with the methodology is also described in this section.

3.9.1 IPF VALIDATION

When the verification is done, the validation is next. Seeing as the IPF procedure itself is frequently researched, robust and well-established technique, the procedure itself does not need to be validated for each application. The control variables on the other hand should be validated to confirm that the resulting generated population is trustworthy. This can be done in different ways. Commonly used are methods to calculate the correlations between the disaggregate and aggregate data. The quality of the fit can also be calculated with the R^2 values. Another data set can be gathered to test whether the IPF procedure has good predicting power. The verification and validation step can lead to changes in the IPF procedure and control variables, therefore there are feedback loops to these two parts in the framework.

3.9.2 HOUSEHOLD ALLOCATION VALIDATION

The validation of household allocation requires a data set that is external to the model and contains the same variables for houses. Correlation coefficients can then be calculated between the model results and the external data set to assess the goodness of fit. The houses in OSM can also be plotted using colour coding for each homogenous household type which results in maps containing the spatial distributions of the households. The spatial distribution of the households of the test data set and validation data set can be compared to each other afterwards to assess how well the household allocation presents reality.

3.9.3 UNCERTAINTY

It is important to note that the IPF procedure itself is deterministic. So, if there is uncertainty in the data used for the IPF procedure, the errors will also propagate through the entire model. Another source of uncertainty is the OSM data. If house locations or the number of houses in a certain area are not accurate, the distribution of the households over the houses will also contain errors. Lastly, the household allocation method might also introduce uncertainty because the model might not fit the data precisely. In these cases,

multiple synthetic populations can be generated and then sampled using for example the Monte Carlo Simulation or Halton draws.

To gain better insight into the uncertainty of the household allocation, a sensitivity analysis can be performed. The number of homogeneous households placed in houses having certain characteristics when changing the inputs slightly, can be an indicator used in the sensitivity analysis. For a more complex sensitivity analysis, the maps containing the spatial distributions of the households can also be used. The global sensitivity analysis for spatially dependent outputs described by Marrel et al. (2011) describes how sensitivity analysis in this case can be performed.

4. IMPLEMENTATION IN CASE STUDY

In this chapter, the results and findings from implementation of the framework in Chapter 3 are presented. The challenges met during the case studies are also discussed along with mitigation strategies. The chapter will start by describing the context for the case study and then a paragraph is devoted to each of the steps of the proposed methodology presented in Chapter 3.

4.1 CASE STUDY CONTEXT

The physical environment entails buildings, infrastructure, water, soil, landscapes and the natural environment. Activities are carried out in this physical environment by individuals, businesses, and authorities and each of these entities have their own interests. Rules and regulations were drawn up in the form of an Environment and Planning Act to structure all these interests and conflicts that arise from them along with controlling the effects on the physical environment (Teekens, 2017).

In July 2022, the Dutch Government will implement a new Environment and Planning Act that decentralizes the decision making and gives more policy space and policy discretion to municipalities (Vereniging van Nederlandse Gemeenten (VNG), 2020). The new Act will require every municipality to submit their noise emissions annually starting from 2021. This applies to roads with an intensity higher than 4,500 personal car units per day (PCU/d). For the year 2026, the noise emissions have to be reported for roads with an intensity higher than 1,000 PCU/d (Van Der Honing & Henckel, 2021).

Since neighbourhood access roads can already meet this requirement, it is essential that traffic can be simulated at the level of neighbourhoods. To aid in this, a travel demand model is needed to estimate the travel demand for neighbourhoods. The population synthesis and household allocation methodology can be applied within a travel demand model for this purpose.

The study area for this research should be the size of a neighbourhood. The neighbourhood should be small enough so that field research can be conducted. The area must also have a variety of activities such as living, working and education. Having a combination of activities in the study area makes the study area heterogeneous and comparable to many neighbourhoods. The V-MRDH (Verkeersmodel Metropoolregio Rotterdam Den Haag) is a traffic model for the region Rotterdam and The Hague and uses its procedure to gather data such as number of residents, houses and average cars per households for its zones. Panteia B.V. has access to this data, so it was decided to use these zones as a guideline to demarcate the study area.

The study area as chosen is shown in Figure 15. The different data zones from the V-MRDH can also be seen in this figure. It concerns part of the Meerzicht Oost neighbourhood in Zoetermeer and is bounded by the Meerzichtlaan, the Africaweg and the railway tracks of the passenger railway operator named Nederlandse Spoorwegen (NS). For this study area a thorough assessment has been made as motivation that the study area can be used within this research. The assessment is presented in Appendix C. The analysis showed that the variability that has been measured qualitatively and quantitively is sufficient and makes the case study transferable to other neighbourhoods.

For the present research, it is important to have an algorithm that is programmed in such a way that it can be executed in an open-source environment and be transparent. Essentially, any of the methods can be chosen for population synthesis but to fit the application within the time budget of this research, it was opted for the classic IPF that is zone-by-zone and uses single-level fitting. The single-level fitting approach provides more flexibility in the data requirements but results in a less detailed population as there is only a population synthesized at the household level. For the study area there happened to be more data on the

household level than the agent level as well, so fitting at the agent level would have been problematic too. To enable the household allocation, the single-level fitting will have to be carried out at the level of households.

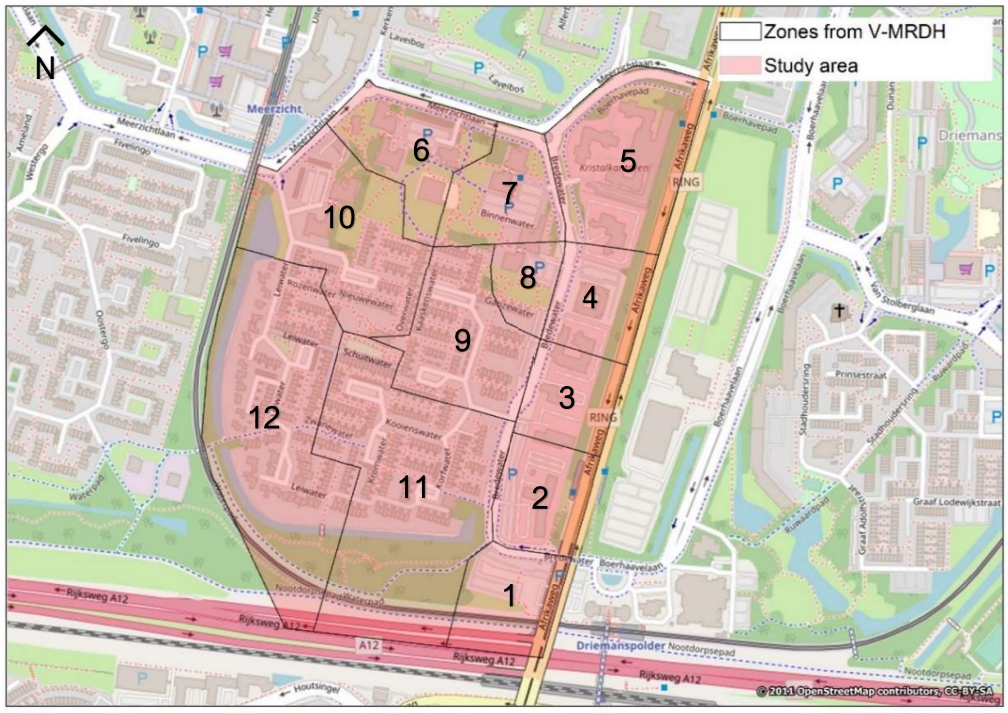

Figure 15 Chosen study area

4.2 MODEL SPECIFICATION

Simplifications and assumptions are introduced in this case study. This is needed to make results attainable within the timeline of this research. The goal is also to provide a simple proof-of-concept. Therefore, the following assumptions are made:

- Households are allocated to houses based on household attributes and housing unit characteristics. No other social circumstances (e.g. crime rate) or environmental attributes (e.g. proximity to shopping areas).
- One household resides in each house or residential unit.

And the following simplifications and constraints are made:

- IPF will only be done at household level. This requires less data and a relatively simple single-level fitting approach, which will be better as it is required to write an own script that does not need paid software packages.
- The amount of control variables to be used should not exceed four to prevent convergence issues (should they occur) and to not complicate the data requirements.
- Only integer households will be generated (some correlation structures might get broken this way).

On the decision of what household variables to use as control variables, two aspects are important:

- *Whether the variables chosen are in line with the model purpose.* In this case, travel patterns need to be modelled to eventually get the flows on roads of interest. The effect of changes (such as pricing schemes or changes in the infrastructure) that may occur because of the Environmental and Planning Act, will need to be captured in some way in the chosen control variables.
- *Whether the control variables are available in the data sets for the seed data and the marginal totals.* This means that there need to be observations of households that have all the control variables in the same data set; this is the disaggregate seed data. And for the marginal totals, disaggregate household data of these same variables should be available to show the relationship that these control variables have amongst each other.

From the literature review in commonly used household variables, inspiration was drawn for the choice of household variables in the case study. The household size, household composition, household income and car availability were identified as potential control variables that give insight to mobility behaviour and were chosen for this case study.

After reviewing data from the Central Bureau for Statistics (CBS) and the Databank of the Municipality of Zoetermeer, it became evident that the data type and format of any household variables needed for the study area are not available or not collected at such a fine geographical scale. The data from the collection zones of the V-MRDH also had only one household variable available and this variable was not in the right format as it were the average cars per household per zone whereas total amount of households that have 0, 1, 2 or more cars available (totals per homogenous group) is needed. There was no seed data available for this study area either.

It became clear that averages of household variables can be found on the level of neighbourhoods but without the standard deviation or variance, the totals per homogenous group were not able to be extracted. On the other hand, for the entire municipality of Zoetermeer, there was seed data available, this was in the form of the OViN (Onderzoek Verplaatsingen in Nederland, which is research of mobility in the Netherlands) data set from 2015 and 2016. And for the marginal totals, data was sought at CBS. The potential control variables based on the OViN data set are listed in Table 3 on Page 55.

Based on this table, the only options left for the control variables are the household composition, the standardized disposable income and car availability. It should be noted that much of the data used, was not directly available and this required using data from other years and using data at a lower spatial resolution. There were also no readily available constraints, the data from CBS for these variables was mostly univariate. This means that for each variable, frequencies were known for the categories but no crosstabulations with the other control variables can directly be made from this type of data. If there are three control variables and a single-level fitting approach is being used, then the IPF will be three-dimensional and thus require a three-dimensional seed data (OViN) and two-dimensional constraints.

Table 3 Overview of potential control variables

Potential control variable	Motivation	All marginal totals available
Household size	The household size influences the number of trips being made. The more members there are in a household, the more trips will be made.	✗
Household composition	The household composition is closely related to the household size. This variable also includes whether there are children in the household. And children in the household influence the complexity of trip chains (Strathman, Dueker, & Davis, 1994).	✓
Disposable household income	Income influences the amount of trips as well and with this variable in the procedure, the influences of pricing schemes can be assessed.	✗
Standardized disposable household income	This is the same as for the disposable household income. * The data was not available for Zoetermeer, but for the Netherlands.	✓*
Car availability	Car availability influences the length and frequency of trips and also the mode choice. This variable is also related to the income. *The data was not available for Zoetermeer, but percentages of crosstabulations from CBS were available for the Netherlands.	✓*

As stated before, these constraints were not available for all marginals. CBS has only one crosstabulation available which was for household composition by standardized household income for the Netherlands and in the year 2015. By adding another two two-dimensional IPF procedures for each of the remaining two marginal totals to get two-dimensional constraints, the issue with the univariate variables can be avoided. So, two extra IPF procedures were carried out for the household composition by car availability and car availability by standardized disposable household income. This influences the accuracy of the generated population because converting univariate variables to two-dimensional variables using a sample is not as reliable as two-dimensional data that is directly collected. The data to be used in these two-dimensional IPF procedures are two-dimensional seed data (OViN) and one-dimensional constraints (the univariate distributions for the control variables from CBS). This will be further elaborated in the next paragraphs.

4.3 INPUT DATA

This section explains the data sets that are used as input for the IPF procedure. Mitigation strategies are also described for cases where data is not specifically available for the geographical area.

4.3.1 DATA FOR IPF

The three-dimensional seed data and two-dimensional seed data for respectfully the three-dimensional and two-dimensional IPF procedure will be retrieved from the OViN data set of 2015 and 2016. The year 2016 is chosen, as the data set from the V-MRDH is also from 2016. In this data set, households are observed, and information is given for each household in the form of the three control variables and other variables that will not be utilized for this research. The important aspect to note here is that these observations of the three attributes all occur in the same data set. The data set of 2015 was also added as most of the marginal data was only completely available for the year 2015 at CBS. The two years were stacked together.

The two-dimensional marginals needed for the three-dimensional IPF will be calculated through the two-dimensional IPF. For each of the two control variables (the household composition by car availability and

car availability by standardized disposable household income), one two-dimensional IPF will be needed. The one-dimensional marginals needed for the two two-dimensional IPF, are from data from CBS and the Databank of the municipality of Zoetermeer.

For the household composition, the univariate distribution was available from the Databank of Zoetermeer. The data for 2016 was incomplete but the data for 2015 was available. The data set of 2015 will therefore be used and it is assumed that the distribution amongst the categories is the same. The data can be found in Appendix D. For the standardized disposable household income, data from 2015 was used from CBS. It was from a report in 2019 that investigated the wealth in the Netherlands (Centraal Bureau voor de Statistiek, 2019). Again, it must be assumed that the distributions remain the same for the year 2016 as well. The data is shown in Appendix E. For car availability, CBS had data on the percentages of households in the Netherlands that have cars available according to their household composition (Centraal Bureau voor de Statistiek, 2017). This data is presented in Appendix F.

4.3.2 CAVEATS FOR DATA

Seeing as the data that is required, is not available and other ways are deployed to still get estimates for the study area, there should be awareness that this leads to assumptions and in turn this leads to uncertainties in the data. Therefore, the population that is generated may not be the only population that fits the data. There can be multiple populations that will fit the data equally well because of these introduced uncertainties. The data used for the marginal totals are for the Netherlands and are from the year 2015. Whereas the seed data is from the year 2016 and from Zoetermeer. The idea is to downscale the marginal constraints to Zoetermeer and after generating the population for Zoetermeer to downscale this population to the study area.

This is a longer road to get to the population synthesis for the study area and is the result of constraints introduced by data availability issues. In ideal circumstances, the data that is required would be collected and available. if this is not the case, the researcher can still adjust the year that is being used or the geographical scale but must know the uncertainties that comes with this. If there are resources available and the study area is reasonably sized, then a survey can be done to gather the seed data for the IPF procedure. A potential survey is outlined in Appendix G.

The sample size can roughly be estimated by assuming a sample standard deviation (s) as a maximum of 0.5 (this is the value of the standard deviation that will lead to the biggest sample size), a margin of error (MOE) of 5% and a confidence interval of 95%. Filling in the formula for the sample size calculation gives:

$$n = \frac{\frac{Z_{\frac{\alpha}{2}}^2 \times s \times (1-s)}{MOE^2}}{} = \frac{1.96^2 \times 0.5 \times 0.5}{0.05^2} \approx 385 \text{ observations (households needed)} \quad (1)$$

Thus, there are 385 observations needed for the seed data. The OViN data set for the year 2016, only has 282 observations and the set of 2015 has 277 observations. As mentioned before, the years 2015 and 2016 are stacked together for the case study so this sample is big enough. Increasing the margin of error and decreasing the confidence interval can also decrease the sample size needed if there is no possibility of stacking data sets from multiple years.

If the years are not stacked, it should be mentioned that the amount of observations differs from year to year which can ultimately lead to whether or not the sample size is statistically significant. Table 4 demonstrates the observations of the years for Zoetermeer from OViN and from ODiN (Onderweg in Nederland). ODiN is the successor of OViN and is essentially just a name change when regarding this research. The seed data sample size is only sufficient for ODiN when using a single year and the sample size calculation from above.

Table 4 OViN and ODiN observations for Zoetermeer

Data collection	Year	Sample size (n)
OViN	2015	277
OViN	2016	282
OViN	2017	286
ODiN	2019	510
ODiN	2020	500

Although, it should be noted that the disaggregate sample file is usually a small file containing representative households with real existing combinations of attributes from the population and is less reliable. The aggregate data (marginals) are the more reliable data set. The IPF uses these marginal totals to enumerate the unreliable seed data. So, the sample file itself does not have to be statistically significant. A smaller sample size could still be used, but this has some repercussions. The sampling error (standardized absolute error between synthesized and true population) was estimated to be between 5% and 11% when using a sample size that is 5% of the real population. And this sampling error reduces when the sample size increases (Choupani & Mamdoohi, 2016).

4.4 DATA HARMONIZATION

The seed data for the three-dimensional IPF is a table that has 3 sides and can be visualized as a cube. To fill in all the sides, there are three two-dimensional variables needed. The three dimensions can be seen as the rows, the columns, and slices. The table will be household composition by household income by car availability, meaning the following two-dimensional margin totals are needed:

- Household composition by household income
- Household composition by car availability
- Car availability by household income

The order of these variables may be switched as long as the dimensions are consistent with the seed data. The seed data defined here is of the format (rows, columns, slices). Therefore, the rows have index 0, the columns have index 1 and the slices have index 2 for the case of a three-dimensional seed matrix. For the case study, the format for the seed matrix is Household Composition x Household Income x Car Availability. To gain the marginal totals, the two two-dimensional IPF procedures have to be executed first. To do this, the variables have to be further specified and the categories have to be decided. This is referred to as data harmonization. In order to do this, the OViN data set has to be prepared.

4.4.1 OVIN DATA SET

The OViN data set has entries that are based on the trips a person (OP) makes. It also includes attributes of the household that this person belongs to as mentioned before. For this person, all the trips recorded appear, meaning that without filtering duplicates out, it will appear as if there is a lot of household information available when in reality there is not. So, all duplicates should be taken out by only using data for which the OP is specified as 1 (new person) and thus discarding all rows in which the OP is specified as 0 (not a new person). The data set should also be filtered for the municipality code. The population synthesis only requires data from Zoetermeer. Afterwards, the columns with the control variables should be taken. So, the data set will only have new households (no duplicates) that are all from Zoetermeer and the household

composition, household income and car availability are specified for this sample. This procedure is carried out for OViN 2015 and 2016 and then the years are stacked together to make one data set. Each of the variables will be explained in the next sections.

4.4.2 HOUSEHOLD COMPOSITION

The variable that resembles the rows and has index zero is the household composition this gives information of the composition of house. The OViN data had 8 (one-person household, couple, couple with kids, couple with kids and others, couple with others, one parent and kids, one parent and kids and others, other composition) categories but these were reduced to five types. These five types are:

- Type 1: one person household. It is the same as the type 1 (one-person household) of the OViN data set.
- Type 2: Couple without kids and is the same as type 2 of OViN.
- Type 3: Couple with kids and is the same as type 3 of OViN
- Type 4: other multiple person households which consists of more than one person and can not be classified in the other categories. For this, the type 4, 5, 7 and 8 of OViN were grouped together.
- Type 5: One parent household and this is the same as type 6 from the OViN data.

The marginal data that was available from the Databank of the Municipality of Zoetermeer was also grouped in terms of these categories. The categories of the seed data have to match the categories of the marginals. In all the cases, the OViN data categories were adjusted to match the categories of the same variables in the marginals. Normally, the groups should be created in such a way that there are observations for each homogenous group. This was not possible for the case study because if the categories would be grouped even more, lots of detail would be lost. To avoid the zero-cell problem, all the categories that have zero observations were replaced with the number 0.0001.

4.4.3 HOUSEHOLD INCOME

The variable that resembles the columns and has as its index 1 is household income and this gives the yearly disposable standardized income of an household in euros. The Central Bureau for Statistics (CBS) (2021) defines the disposable standardized income as the net income that has been corrected due to differences in household size and household composition. The correction is implemented through the use of equivalence factors that are established every year by CBS. This is needed to capture all the advantages and even out the scales when looking at collective or joint households. The equivalence factors reduce the income of all households to that of a one person household to make comparisons possible (Centraal Bureau voor de Statistiek, 2021).

In OViN, it consisted of 7 categories (<€10,000, €10,000-€20,000, €20,000-€30,000, €30,000-€40,000, €40,000-€50,000, >€50,000 and income unknown). The marginals used were not directly derivable from the data available at CBS or at the municipality of Zoetermeer. Data from CBS had a crosstabulation with the household composition and standardized disposable household income for the Netherlands for the year 2016 (CBS, 2021). This table was scaled to the household composition values of Zoetermeer and scaled again to ensure that the sum of the rows were equal to the sum of the columns. The result of this table is given in Appendix E. Lastly, the categories were limited to 5 to reduce the amount of zero cells. These categories are also applied to the income groups of OViN by taking the following categories:

- <€10,000 which is the same as the first income category of the OViN data set
- €10,000-€20,000 this is also the same as the second income category of OViN
- €20,000-€30,000 this is also the same as the third income group of OViN
- €30,000-€40,000 this is also the same as the fourth income group of OViN

- >€40,000 which consists of the fifth and sixth income groups of the OViN data set. So, inlcuded here are the group €40,000-€50,000 and >€50,000.

The six categories for which the income was known of OViN have been reduced to these five categories. The OViN data set also has income group 7 which are the households that had an unknown income. There were seven households that fell under this category. These households were placed in an income group on the basis of their household composition and the probability of them belonging to a certain household group (the income group with the most observations for the same type of household composition). For example, there is one household of type 4 (other multiple person household) for which the income was unknown. Two other multiple person households for which the income is specified, belong to income group 2. So, the household with the unknown income was therefore also added to income group 2. This method was used for the marginals.

For the seed data, similar method was used. In the seed data, which is three-dimensional, car availability was also taken into account. The household of type 4 with unknown income had one car available to them. So, the amount of observations of households of type 4 with one car was looked up and it was evident that they all belonged to the income group 2, so this household was also added to income group 2. For the six other households, there were two options, meaning that there were two homegenous groups with an equal amount of observations. In this case, three households were added to the first group and three were added to the second group. To make it more clear: all the six households had household composition type 3 and 2 cars available. From the rest of the data, it was seen that 6 households had these attributes, belonged to income group 3 and another 6 households that had these attributes belonged to income group 4. Thus, of the six unknowns, three households were added to income group 3 and three were added to income group 4.

4.4.4 CAR AVAILABILITY

Car availability is the third control variable and has the number 2 as its index. It denotes the number of cars that are available to a household. In OViN these were integers and the data set for Zoetermeer had 0, 1, 2, 3, 4, 5 and 7 cars available in the observations. This was brought back to 0, 1, 2 and 3+ cars to match the data set of the marginals. The marginals were again not directly available. So, in other to get the data, percentages were used. The table was available through CBS and gave the percentages of the whole population in terms of household composition and car availability for the entirety of the Netherlands (Centraal Bureau voor de Statistiek, 2017). This table is given in Appendix F. These percentages were then multiplied with the number of cars in Zoetermeer which was also obtained from CBS.

4.4.5 VISUALIZATION OF DATA

To make the IPF procedures clearer, visualizations were made and are shown in Figure 16 and Figure 17 on 60 and Page 62. On the left side of Figure 16 are the univariate distributions that will be used as the marginals for the two-dimensional IPF procedures. The crosstabulations on the right side of this figure are the seed data for the two-dimensional IPF. The household composition x household income is placed here to show the format but does not require a two-dimensional IPF procedure as the crosstabulation could already be derived from data from CBS. OViN data was filtered and counted to fill in these tables. In the crosstabulations, the seed data is indicated in orange. The first crosstabulation is household composition and car availability and is given in Table 5 on Page 60.

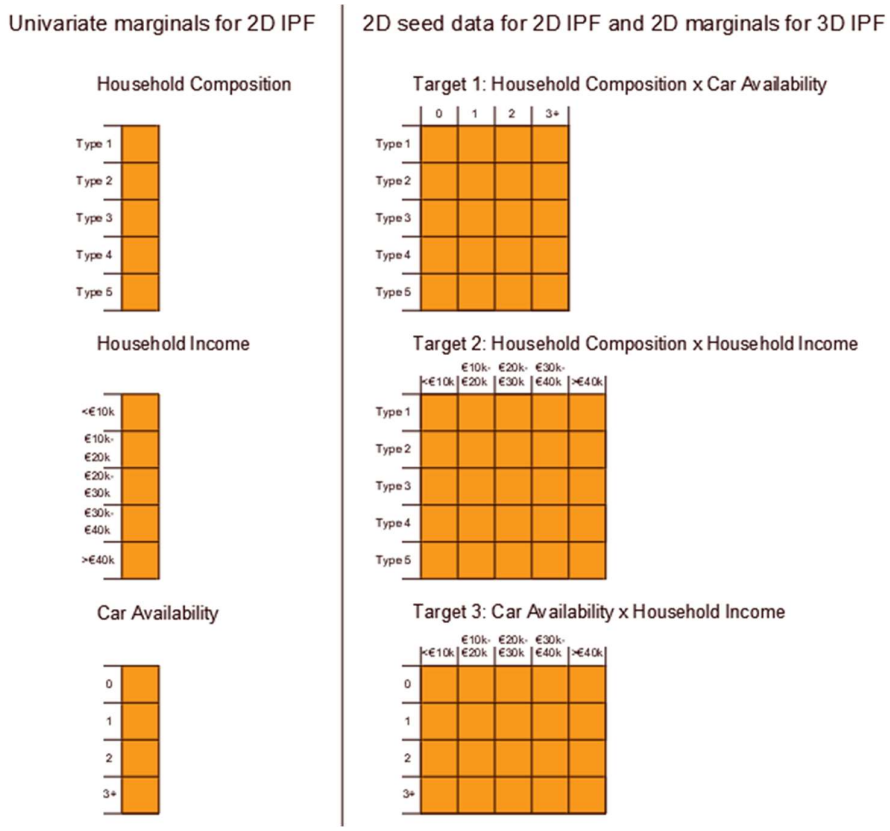

Figure 16 Visualization of marginals and seed data for the 2D and 3D IPF

Table 5 Crosstabulation household composition by car availability (uncorrected)

	Car availability					
Household composition	0	1	2	3+	Total	Target
Type 1	36	39	4	0.0001	79	17,400
Type 2	8	98	54	2	162	15,100
Type 3	7	117	126	18	268	14,800
Type 4	2	3	2	0.0001	7	1,000
Type 5	11	28	13	1	53	5,400
Total	64	285	199	21	569	53,700
Target	16,132	25,334	10,446	2,663	54,575	

The sum of the rows should be equal to the sum of the columns. There are 53,700 households in Zoetermeer according to the data set of the household composition. The number of households of the car availability was not directly reported but rather calculated through percentages and the total amount of cars in Zoetermeer. This led to the 54,575 households. However, this was not directly reported so it is regarded as being less reliable than the total for the households reported by the household composition data set. Therefore, it was decided to match the sum of the rows to the sum of the columns by scaling it using a factor of 54575/53700. The result is presented in Table 6.

Table 6 Corrected crosstabulation household composition by car availability

	Car Availability					
Household composition	0	1	2	3+	Total	Target
Type 1	36	39	4	0.0001	79	17,400
Type 2	8	98	54	2	162	15,100
Type 3	7	117	126	18	268	14,800
Type 4	2	3	2	0.0001	7	1,000
Type 5	11	28	13	1	53	5,400
Total	64	285	199	21	569	53,700
Target	15,873	24,928	10,279	2,620	53,700	

The crosstabulation for the second control variable was for the Netherlands. This was then scaled down to the size of Zoetermeer. This was done by multiplying with the factor that is the total amount of households in Zoetermeer divided by the total amount of households in the Netherlands. The result is shown in the Table 7. Table 7 leads to other totals for the household composition than the ones given in Table 6. This needs to be corrected again and to do this the table is multiplied with a factor. The resulting crosstabulation is shown in Table 8.

Table 7 Uncorrected crosstabulation of household composition by standardized disposable household income

Household	Standardized disposable household income						
Composition	<€10k	€10k-€20k	€20k-€30k	€30k-€40k	>€40k	Total	Target
Type 1	2,109.03	7,884.57	5,947.41	2,571.84	1,353.89	19,866.74	17,400
Type 2	208.51	2,920.52	5,002.08	3,771.46	3,307.95	15,210.53	15,100
Type 3	179.63	2,078.74	4,817.52	3,969.40	2,830.36	13,875.65	14,800
Type 4	51.42	198.65	313.47	256.41	185.26	1,005.21	1,000
Type 5	162.02	1,624.39	1,261.62	474.78	216.07	3,741.88	5,400
Total	2,710.61	14,706.87	17,342.1	11043.89	7,896.54	53,700	53,700

Table 8 Corrected crosstabulation of household composition by standardized disposable household income

Household	Standardized disposable household income						
Composition	<€10k	€10k-€20k	€20k-€30k	€30k-€40k	>€40k	Total	Target
Type 1	1,847.17	6,905.58	5,208.96	2,252.51	1,185.79	17,400	17,400
Type 2	206.99	2,899.30	4,965.73	3,744.06	3,283.92	15,100	15,100
Type 3	191.59	2,217.22	5,138.45	4,233.83	3,018.91	14,800	14,800
Type 4	51.16	197.62	311.84	255.08	184.30	1,000	1,000
Type 5	233.81	2,344.20	1,820.67	685.17	316.15	5,400	5,400
Total	2,530.72	14,563.92	17,445.65	11,170.64	7,989.07	53,700	53,700

The crosstabulation for the second two-dimensional IPF is car availability and standardized disposable household income. And the result is illustrated in Table 9.

Table 9 Crosstabulation car availability by standardized disposable household income

Car availability	Standardized disposable household income						
	<€10k	€10k-€20k	€20k-€30k	€30k-€40k	>€40k	Total	Target
0	4	37	17	6	0.0001	64	15,873
1	5	68	120	64	28	285	24,928
2	4	18	67	62	38	189	10,279
3+	0.0001	1	6	6	8	21	2,620
Total	13	124	210	138	74	282	53,700
Target	2,530.72	14,563.92	17,445.65	11,170.64	7,989.07	53,700	

The seed data for the three-dimensional IPF is visualized as a cube in Figure 17. This seed data was filled in a matrix in the format of m[i, j, k] with m being the name of the matrix and i, j and k being respectfully the rows, columns and slices. The table for this seed data is shown in Appendix H.

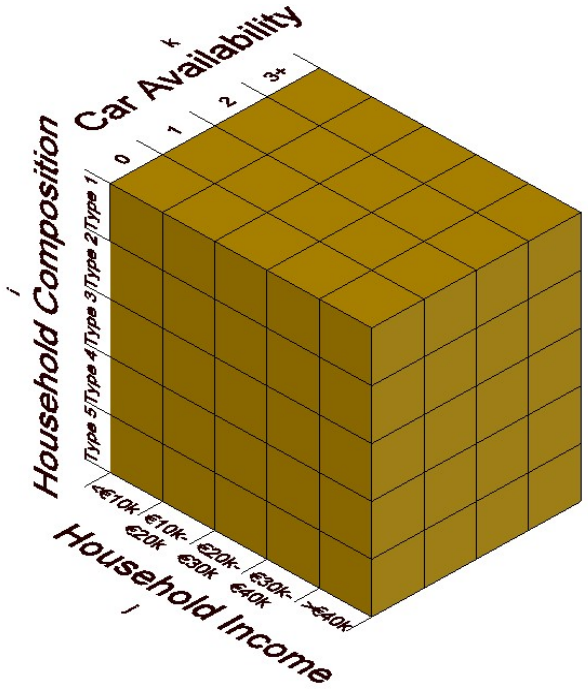

Figure 17 Seed data for 3D IPF adapted from Deming and Stephan (1940)

4.5 IPF PROCEDURE

In this section, the programming and verification of the IPF procedure are presented. The resulting fitted tables are also shown.

4.5.1 PROGRAMMING OF IPF

For the programming of the IPF procedures, a Python package named ipfn will be used in Jupyter Notebook. This package has documentation and examples available for the IPF procedure. The example codes from Forthomme and Ballis (2021) were used for this script. According to them, there are two ways to program the ipfn function. The first approach is using Numpy, which is another Python package that makes the

procedure fast and efficient. The second approach is using Pandas, which is also a Python package. The Pandas version is not as fast but is easier to understand and use.

For the case study, it was chosen for the Numpy version to make the procedure computationally efficient and have the possibility of scaling up to bigger study areas. The script for the three two-dimensional procedures can be found in Appendix I. And the script for the three-dimensional IPF procedure can be found in Appendix J.

Verification was done after the script was finished. This entails stepwise checks of the procedure to make sure that every step is carried out correctly. This can be done by summing the rows and the columns of the fitted tables and checking whether these matches the marginal totals. Also, it can be checked whether the sum of the whole matrix is equal to the total amount of households in Zoetermeer before scaling it down to the study area. This check can be seen in the script in Appendix K. When running the IPF procedure, the message that the IPF has converged or reached its maximum iterations should also be spotted. For the procedures, the sum of the rows, columns and slices were all equal. There were no zero-cells found and the procedures have converged. It was concluded that the procedure was carried out correctly.

4.5.2 RESULTS OF IPF

After running the script for the two-dimensional IPF procedures, the results were obtained, and these are shown in the tables below. The numbers have been rounded to two decimal places. These numbers were not rounded in the script to prevent rounding errors as much as possible. The households will be rounded after downscaling has happened.

Table 10 Fitted crosstabulation household composition by car availability

Household composition	Car availability				
	0	1	2	3+	Total
Type 1	10,950.97	5,975.40	473.71	0.03	17,400
Type 2	1,502.09	9,267.96	3,947.31	382.59	15,100
Type 3	777.06	6,541.74	5,445.36	2,035.75	14,800
Type 4	466.23	352.24	181.15	0.02	1,000
Type 5	2,176.65	2,790.65	231.11	0.02	5,400
Total	15,873	24,928	10,279	2,620	53,700

Table 11 Fitted crosstabulation car availability by standardized disposable household income

Car availability	Standardized disposable household income					
	<€10k	€10k-€20k	€20k-€30k	€30k-€40k	>€40k	Total
0	1,580.58	8,524.43	4,141.45	1,626.41	0.04	15,873.0
1	651.16	5,163.33	9,634.79	5,717.64	3,761.10	24,928
2	298.97	784.41	3,087.32	3,178.89	2,929.46	10,279
3+	0.02	91.75	582.10	647.70	1,298.46	2,620
Total	2,530.72	14,563.92	17,445.65	11,170.64	7989.07	53,700

The totals are not precisely equal to the marginal totals, this is because of rounding issues. According to the ipfn function, the IPF has converged for all these tables. These tables will be used for the marginal totals of the next IPF procedure. They must be in the right order to do this. The dimensions must match with the format of the seed data. For this reason, the transpose must be taken of Table 10 as input for the three-

dimensional IPF procedure. This can also be seen in the script in Appendix J. The result of running this IPF is shown in Appendix K.

The next step is to scale down the generated population of Zoetermeer to that of the study area. This population has to be scaled down to the size of the Meerzicht Oost neighbourhood. For this, the number of houses or residential units have to be counted. This is done in Section 4.8 and resulted in 1,122 houses. There are 53,700 households in Zoetermeer. Therefore, to scale the population down, the generated population matrix is multiplied with a factor of 1,122/53,700. The result of this is given in the last column of the fitted matrix in Appendix K. By scaling down the population this way, the underlying assumption is made that the distribution for Zoetermeer is the same as the distribution for the study area.

Finally, the households generated are rounded off to integers to make sure that there are no fractions of households that need to be allocated in a later stage. A sum preserving rounding method is used for this to make sure that the total number of generated households stays 1,122 even after rounding. The script is also in Appendix K.

4.6 OSM DATA & DATA QUALITY ASSESSMENT

4.6.1 RETRIEVING OSM DATA

To retrieve the data for the study area, a bounding polygon was used in the OSMnx package in Jupyter notebook. The coordinates for the data collection zones were available through the V-MRDH. The outside boundary from this cluster of zones was taken and was altered to a more simplified polygon that did not require the specification of many points and was not going through any buildings. The chosen coordinates were then programmed in Python to be the bounding polygon from which to retrieve data. Python takes the coordinates in the format of (longitude, latitude). The script is shown in Appendix L.

The script also includes plotting the study area with the different labels that are specified for the tag ‘building’. The street network is included but when plotting, there was no distinction made in roads, cycle paths or pedestrian paths. The specification in the command was network=’all’. The map is shown in Figure 18 on Page 65. For the buildings specified with the label ‘yes’, it is not sure what type of buildings these concerns. This will require further inspection either through comparison with other data sources such as Google Maps and/or field research.

Figure 18 Study area with building labels

4.6.2 QUALITY ASSESSMENT OF OSM DATA

The tagging quality will be checked first by using the webtool from Almendros-Jiménez & Becerra-Terón (2018). The tool uses an area that is larger than the study area, because it is not possible to manually select an enclosed area. This bigger area will be referred to as webtool area for clarity. However, if the tagging quality is good for the webtool area, this should also hold for the study area as this study area lies within the webtool area and is considered as well.

The tagging quality of the buildings will only be checked as this is what needs to be accurate for the house allocation. Within this webtool area, there are 1354 OSM entities. According to Keßler & De Groot (2013), a large number of contributors, versions and confirmations are positive indicators of trust while revisions and corrections are seen as negative. The versions of tags will be explored first.

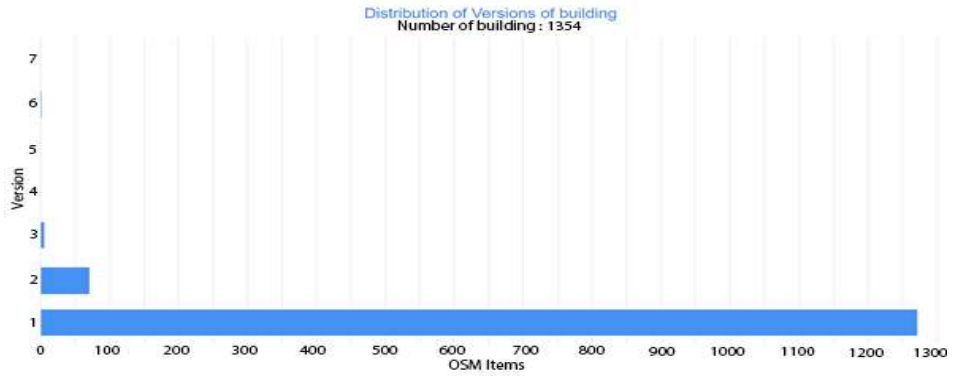

Figure 19 Versions of OSM entities

In Figure 19, it can be seen that most of the OSM entities with the tag of building, have one version. The average of versions for tags that are associated with buildings is presented in Figure 20. The average number of versions is equal to 1.88. This could indicate that the tags are not trustworthy. However, the majority of OSM data has been imported from the Automotive Navigation Data, this can explain why most entities have only one to two versions.

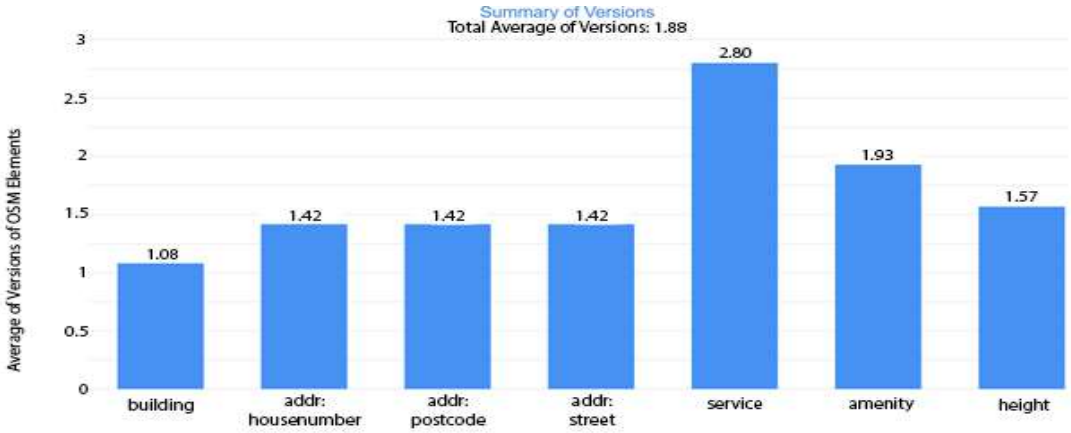

Figure 20 Summary of versions

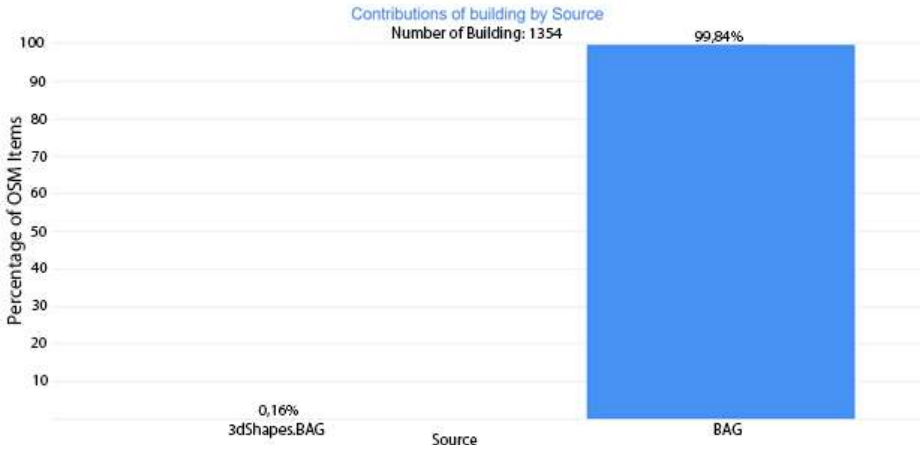

Figure 21 Sources of buildings

In Figure 21, the sources can be seen from which the buildings have been mapped. The buildings are all extracted from BAG (Basisregistratie Adressen en Gebouwen) which is the registration for addresses and buildings in the Netherlands. In the browser editor used to map in OSM, there is a layer for the BAG in which the outlines of all buildings can be seen. OSM members use this layer to trace buildings and specify attributes of buildings. This is the reason that all OSM elements have BAG as its source in the Netherlands.

The different values for the key 'building' is illustrated in Figure 22. The key 'building' has eight values for the webtool area. It can also be seen that the majority of buildings are houses. This indicates that the webtool area is a residential area. Having the eight distinguished values leads to this data being considered as rich in the context of this research when just taking the number of values into account. However, there is a significant amount of buildings with the value 'yes'. This is a general value and states that the element is a building but does not provide further details. So, in this group of buildings marked with 'yes' there can still be houses and apartments that are not classified as such and when counting the housing units for the study area, these buildings will not be detected because of the general value for the tag 'building'. Regardless of having eight values, the data cannot be seen as rich for this reason.

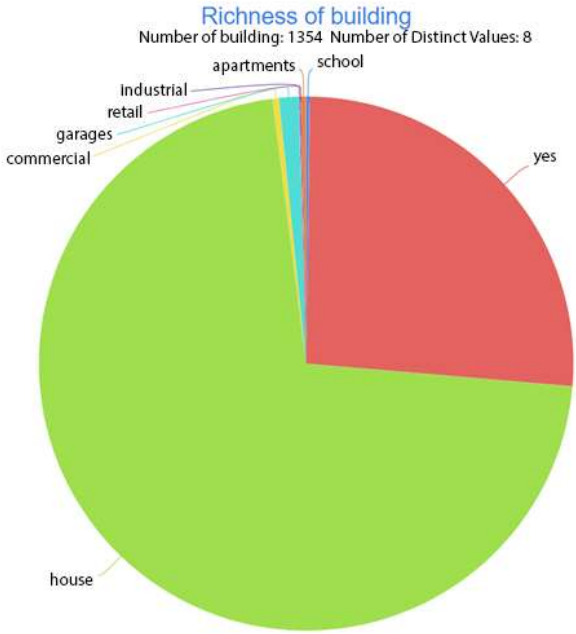

Figure 22 Richness of building

In order to validate the OSM data, it was important to find out significant errors that have been detected by error detector tools available for OSM for the study area. The tool used for this is Osmose (http://osmose.openstreetmap.fr) and this can detect potential data errors, inaccuracies or sparsely mapped places. These errors also include minor precision errors and errors that may little impact for the use. All the errors found in the study area can be found in Appendix M. The errors are numbered and explored in the Appendix along with the impact that the potential errors have on this case study.

From the errors and potential errors detected by Osmose, it can be concluded that there are no issues with the tags that will be of great influence on this research. If the data from OpenStreetMap would be used for geometries or model networks for traffic flow modelling, this can be problematic especially the bad turn lanes and missing access links.

To get details for the specific study area, the OpenStreetMap data was loaded in Jupyter Notebook. Then the buildings were filtered as this type of element is needed from OSM for household allocation. The filtered

list is saved in an Excel sheet. The script for this is outlined in Appendix N. When filtering for buildings, there were 763 OSM items. The variables present in the data set are summed in Appendix O. In Appendix O, the occurrences of the different tags and names of buildings is also explored. This is done to show the information that is stored in OSM for the study area. From the data set, it was apparent that the surface area of the buildings was not stored in OSM.

Furthermore, an overview of the comparison between OpenStreetMap and Google Maps is presented in Appendix P. Field research was also conducted and held in contrast to OpenStreetMap. The field research comparison is also given in Appendix P. From the findings presented in Appendix P, it can be concluded that the companies that were not found in the study area, are self-employed freelancers that work from home and in this case, the building in which they work from should still be counted as a house. When using OpenStreetMap, this will be the case seeing as these companies are not mapped in OSM. Another option can be that these companies have relocated or closed their doors and it has not been updated in Google Maps yet. In both cases, OSM data will be sufficient to use. Moreover, it can be concluded that for mapping of the schools, the data from OSM is correct and that of Google Maps is wrongly mapped.

As stated in Section 4.6.1, the buildings that have been specified with the value 'yes' for the tag 'building' have to be further investigated. This was also done during the field research through observations of these buildings. An overview of the findings is shown in Appendix Q. From the results presented here, it can be concluded that the OSM data needs corrections. There were five residential buildings found that contain a total of 603 residential units (apartments). Before using OSM in the household allocation, these apartment buildings need to be manually adjusted. The next section will focus on correcting this.

4.7 FILTER HOUSES

The houses and residential units must be filtered from the list of buildings and for this these elements need to be specified correctly. In the previous section, it was concluded that there are corrections to be made concerning five residential buildings. The corrections were made in OSM itself, but because these changes have to be checked by multiple members of the OSM community, these changes cannot be immediately seen and loaded in Jupyter Notebook. Therefore, these corrections were also made to the DataFrame of buildings in Jupyter Notebook. To account for the multiple flats in the apartment buildings the tag 'building:flats' is added. This value is based on the counted house numbers for each of the apartment buildings. The value 'apartments' is commonly used in OSM for indicating buildings that consists of individual dwellings and the key 'building:flats' is commonly used to provide the number of individual dwellings in a building (OpenStreetMap Wiki, 2021).

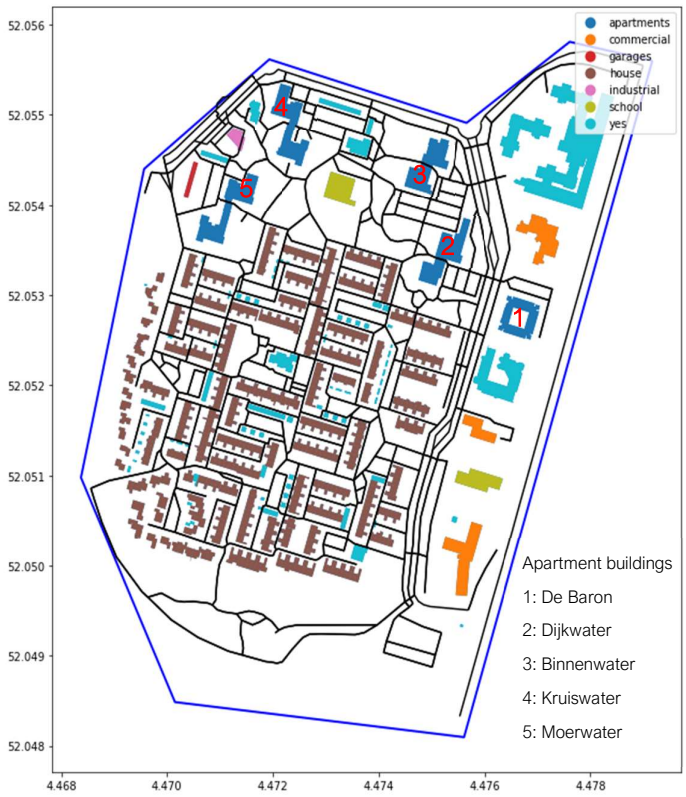

Figure 23 Adjusted graph with new classification of buildings

To change the value of the keys for these buildings, the osmid (OpenStreetMap ID) was used. These ID's can be found though 'Query Features on the OpenStreetMap website. Using this ID as index, the values can be changed. And after changing the values, a new column was added to the DataFrame to provide the number of flats (building:flats).

The corrected buildings are plotted again in Figure 23. The apartment buildings have been numbered too to make identification easier. The script for changing the building classification is shown in Appendix R. These adjustments were also added through the browser editor for OSM data, so these features are now accurately present in OSM.

4.8 HOUSEHOLD ALLOCATION VARIABLES

For the study area, there were a limited number of variables available. The possible variables to choose from are:

- The household variables:
  - Household composition
  - Household income
- The house variables:
  - Surface area (this can either be specified in the tags or can be calculated through the geometry)
  - Number of flats

It was decided to use all the variables apart from the number of flats. The number of flats will help to determine the number of residential units that need to be filled. But this variable will not directly be a part of the allocation procedure itself. For these house allocation variables, a data set containing these variables that may provide insight into the relationship between the variables is needed.

During this research, it was found that there is a lack of open-source data that provides insights into household characteristics and house characteristics and the relationships that exist between these two. For the Netherlands, this type of data is collected in the Housing Survey of the Netherlands (Woon Onderzoek Nederland). This survey provides insights into the housing situation of Dutch households. However, this data set is not open source and can therefore not be used in the household allocation. On the other hand, this data set can be used as the validation data set.

4.9 HOUSEHOLD ALLOCATION

The surface area is chosen as a household allocation variable, however for the study area the surface area of the buildings was not saved. The first section in this paragraph will outline a method to get the surface area of the buildings. Then because there is no open-source data available for the household allocation, the following section will provide a means in still getting data to allocate households. Then the household allocation method will be chosen and described along with the results from implementing the household allocation procedure.

4.9.1 SURFACE AREA CALCULATIONS

It should be noted that the living area is not the same as the surface area given in OSM data. The area given in OSM is the area of the polygon used to draw a house, but this area does not include the multiple levels in a house. So, this must be corrected. There are two ways to correct this:

1. Link the houses with the BAG administration to find the actual living area (this data is available in BAG).
2. Compare a sample of the area of the houses with the living area reported in BAG and calculate a factor that can be multiplied with all houses to get an estimate of the living area of the houses.

The focus of the research is to utilize OSM data as much as possible within the framework. Therefore, the second option is chosen. Additionally, the first option brings complications with it because linking the two data sources together requires an address, postal code, or house number. And there are errors in these tags in OSM as well as BAG. This may lead to wrong couplings and even losing data.

For 52 houses that were randomly selected in the study area, the surface area according to BAG was looked up and compared to the surface area in OSM. The BAG reference tag (ref:bag) was used to find the houses in the BAG. Then a factor was calculated by dividing the surface area from BAG by the service area from OSM. It was found that on average the living area is 1.9 times the area given by the polygons in OSM. So, to calculate the living area of each of the houses, the surface area from the polygons in OSM was multiplied with a factor of 1.9. The calculation is shown in Appendix S.

As with the houses, when calculating the surface of the flats, a similar problem occurs. So, an assumption is needed before calculating the area of the flats. It is assumed that all flats in a given residential building are equal in size. Then the surface area of the flats can be calculated by distributing the number of flats over all the floors of the apartment building. The script for this is outlined in Appendix S.

4.9.2 EXPERT JUDGMENT

As stated before, there is no open-source data set that can be used for the household allocation. The WoON data set of the year 2015 could be used if the limitation of using open-source data was not imposed. However, this survey did have trend breaks and gaps by not including all the categories of the household composition and containing slightly different variables like the household income instead of the standardized household income. The researcher can also use dummy values in the household allocation based on their expertise. It was decided to use expert judgement instead of using dummy values because to obtain reasonable dummy values, some sort of justification would still be needed, and this would also come from experts as well as there was no familiarity with the households and housing relationships for Zoetermeer and this could then lead to an unrealistic allocation.

An alternative approach is using expert judgement. Expert judgment is often utilized in prediction and decision-making when data is not available. To make expert judgment reproducible and reliable, the

elicitation of expert judgment should be structured. According to Hanea et al. (2017), there are three approaches to a structured protocol for the collection and combination of expert judgement:

1. The classical approach with behavioural aggregation
   Experts discuss and consensus is sought after. The major benefit is that experts interact and share their knowledge and ensure that all experts have the same understanding of the questions being asked. On the other hand, this could also lead to biases and in cases of strong disagreements, seeking consensus does not reveal the variety in the opinions of the group of experts.
2. Mathematical methods to aggregate judgement
   The interactions between experts are limited in these approaches because it is often assumed that this may mislead mathematical aggregation by bringing about dependence in their elicited judgements. The upside here is that the aggregation is explicit and verifiable. The downside is that choosing the aggregation rule is challenging and each rule leads to different properties.
3. Mixed methods
   These methods are a combination of behavioural aggregation and mathematical aggregation. The most popular approach here is the Delphi protocol. This protocol supplies the experts with feedback from other experts over consecutive question rounds. The feedback stays anonymous and there are no interactions between the experts. The method seeks consensus; however, it was also shown that this does not have to lead to increased accuracy (Hanea, et al., 2017).

Hanea et al. (2017) also highlights a protocol that is a combination of all the above methods. The method is named IDEA and stands for Investigate, Discuss, Estimate and Aggregate. Just like with the Delphi protocol, experts give their judgement and get feedback from other experts over consecutive rounds. A big difference between Delphi and the IDEA approach is that the goal of IDEA is not to get to an agreement among all experts. In IDEA, experts first answer questions and then get the judgements of other experts. Afterwards, the experts engage in discussion to address differing opinions. This allows solving issues of definitions and context. The individual judgements given in the first round remain anonymous.

Even though expert judgement is still subjective and can deviate from reality, using this approach still means that the household allocation can be carried out when confronted with a lack of real data. However, it does introduce uncertainty and should be validated when data becomes available. Due to COVID-19 restrictions and time constraints, it was not possible to conduct a full structured protocol for expert judgement but any of the approaches mentioned in this section would suffice and is a matter of preference for the researcher.

Depending on the researcher's preference for whether to include anonymity, interactions between experts or consensus, the approach can be chosen. The question sessions require surveys that can be set up like stated preference surveys and should include the confidence levels that the experts have in their answers. This helps with aggregation of the expert judgements. After the elicitation of expert judgement, the researcher can choose the statistical procedure to use for the household allocation.

A survey was set up for the case study. This survey is constructed in the same way as surveys used for stated preference experiments. The survey is presented in Appendix T. An expert is defined in this case as a person who is familiar with the existing relationship between the household and house variables at hand for Zoetermeer. The survey contains questions that infer the relationship between each chosen household allocation variable.

While constructing the survey, the household allocation method was also chosen to ensure that the questions would lead to solving the envisioned model. It was opted for the regression analysis approach. This approach is transparent and has a straightforward implementation. It also allows for the flexibility needed in this case since there is no data ready for this. The regression model will be discussed in the following section.

Due to COVID-19 restrictions, conducting the survey in a manner suggested for structured elicitation of expert judgement was not possible. The response rate for the survey was also very low. There were only 4 responses. Due to time constraints, it was decided to continue with these 4 responses. The participants gave the feedback that they found it challenging to fill in the survey and were not always sure of the answers given.

4.9.3 REGRESSION MODEL

The regression analysis will use the household variables household composition and household income as predictors for the living area of a house that would be linked to such a household. In the regression analysis interactions between the predictor variables are not considered as this would be challenging to estimate using expert judgement. With the responses of the experts, the envisioned or desired area can be determined for each household type. The formula for the regression analysis is given below.

$$Desired\ area = \beta_0 + \beta_1 HHComp_{typ} + \beta_2 HHComp_{type3} + \beta_3 HHComp_{type4} + \beta_4 HHComp_{type5} + \beta_5 HHIncome_2 + \beta_6 HHIncome_3 + \beta_7 HHIncome_4 + \beta_8 HHIncome_5 \quad (2)$$

Where,

$HHComp_{type\ 2}$: Couple without kids

$HHComp_{type\ 3}$: Couple with kids

$HHComp_{type\ 4}$: Multiple person household (other)

$HHComp_{type\ 5}$: One parent household

$HHIncome_2$: income class with €10,000 - €20,000

$HHIncome_3$: income class with €20,000 - €30,000

$HHIncome_4$: income class with €30,000 - €40,000

$HHIncome_5$: income class with >€40,000

The reference for this regression is households consisting of one person with a household income of less than €10,000 and this is captured in the β_0 coefficient. These categories (one person household and household income <€10,000) had to be removed to avoid multicollinearity issues. The regression analysis was programmed in Jupyter Notebook using Ordinary Least Squares (OLS) regression from the Python package Statmodels. The script for this is given in Appendix U.

The coefficient of determination for the regression analysis was 0.414. This may be seen as an indication of a bad model fit, but since this concerns human behaviour and there is a lot of variability in this, the model fit can be regarded as good. The data from their responses were used and the coefficients were calculated for the regression analysis. This resulted in the following equation for the desired living area:

$$Desired\ Area = 41.63 + (28.00 * HHComp_{type2}) + (15.75 * HHComp_{type3}) + (16.12 * HHComp_{type4}) + (17.00 * HHComp_{type}) + (7.75 * HHIncome_2) + (8.4483 * HHIncome_3) + (23.9747 * HHIncome_4) + (43.4483 * HHIncome_5) \quad (3)$$

4.9.4 DIAGNOSTICS OF REGRESSION ANALYSIS

In this section, the generated results of the regression analysis will be explored along with diagnostic plots. When the household composition and income are both 0, the intercept is 41.63 meaning that the desired living area for households consisting of one person and with an income of lower than €10,000 is equal to 41.63 m^2. As expected with the household income, the coefficient increases when the household income increases. For the household composition, the expectation is that the one-person household will have the smallest desired area and the couple with and without kids' households are expected to be larger. This is also what the regression coefficients show. The R-squared value is 0.414 meaning that 41.4% of the data can be described by these coefficients.

There are three important aspects to be considered here are:

- The model assumptions could be wrong. The relationship might be nonlinear or there are more factors that explain the desired living area apart from the household composition and household income. If the relationship is nonlinear and there are only categorical variables, it would be helpful to add the categories by using weighted effect coding instead of dummy coding. Weighted effect coding uses more values than just zeros and ones. It thus allows assigning different weights at various levels of a categorical variable. This can be translated to a nonlinear model then. Interactions or more predictor variables can also be included to gain a better model fit.
- There is a high variability in the behaviour that the regression analysis is trying to predict. The living area from households can vary a lot. There are instances where a one-person household is due to inheritance is living in a big house. The same can be stated for low-income households. There are also high-income households who have chosen small houses. Due to this phenomenon, the regression analysis can only predict the living area in a limited way.
- The experts were independent of each other and from different institutions (TU Delft, Panteia B.V. and the Databank of Zoetermeer). The surveys were not done in an iterative manner as it was supposed to and there is a high variation in the experts their opinions.

In Table 12, a model is generated for each expert and then the experts their opinions are added together and the influence of this can be seen in the adjusted R-squared value and log-likelihood. Due to the diverging opinions among the experts, the adjusted R-squared value and log-likelihood value decrease with addition of every expert to the model with the full model having the worst fit. In this case, IDEA could have been used to seek consensus and gather opinions in an iterative manner this could lead to a lower variance in the opinions given.

Table 12 Comparison of regression models of experts

Model	No. observations	Adjusted R-squared	Log-likelihood
Model Expert No. 1	25	0.971	-66.700
Model Expert No. 2	25	0.931	-69.621
Model Expert No. 3	25	0.945	-65.937
Model Expert No. 4	22	0.973	-71.291
Model Expert No. 1 and 2	50	0.830	-176.71
Model Expert No. 1, 2 and 3	75	0.628	-298.94
Model Full (all experts)	97	0.361	-454.95

It should also be mentioned that the setup of the survey required the experts to fill in a crosstabulation of the household composition by household income. This leads to experts choosing higher surface areas as with increasing income as they compare each of the composition and income combinations. This results in

high R-squared values for the individual models. If the survey did not have a crosstabulation and used questions in a random order, this relationship might not have been so clear. The experts their estimates differ in what they use as a baseline (the lowest living area and this most likely is associated with one-person household and an income of less than €10,000). This leads to a decreasing R-squared value when combining the opinions of the experts together in one model.

For the regression analysis, several diagnostic plots were created to gain more insight into the regression procedure and the data. First, the data retrieved from the experts and used as input is plotted in Figure 24. The estimated living area based on each household composition type and household income group is plotted in the graph. There is consensus among the experts when the points for a given household composition type and income group are close together. There is variability in the opinions of the experts and not a lot of instances of agreement.

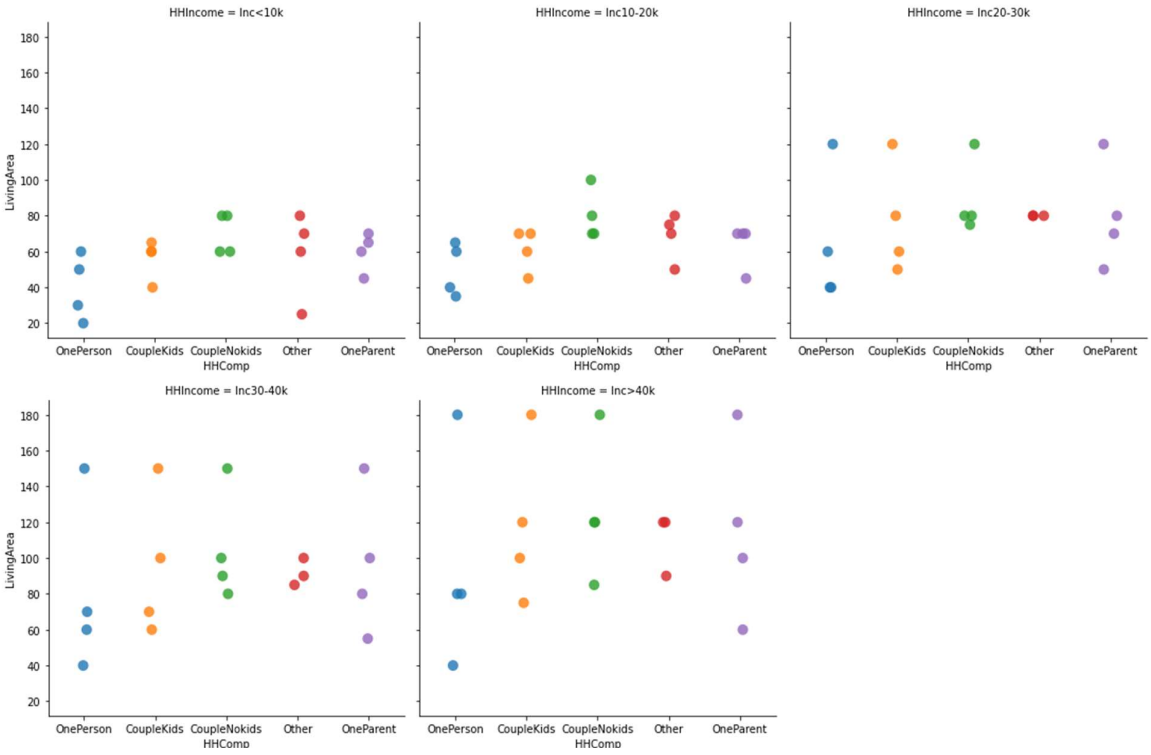

Figure 24 Input data for regression analysis

The standardized residuals are also plotted against the predicted (fitted) values for the living area in Figure 25 using the estimated regression equation (5). The residuals are equal to the observed data minus the predicted regression model values for the data. The red line is the Locally Weighted Scatterplot Smoothing (LOWESS) line and this creates a smooth line through a plot to illustrate trends. Ideally, this red line should be horizontal and close to the x-axis since the residuals should be randomly distributed around zero and there should not be an apparent pattern. This is not the case here as the red line has a downward slope, meaning that as the living area becomes bigger, the standardized residuals also increase indicating that it is not random. It can be concluded that there are heteroscedasticity issues. Linear regression assumes homoskedasticity and this assumption is not met in this case. Since the red line is not curved, there is no indication to opt for a nonlinear transformation. The ascending red line does indicate that there might be features of the model that are not currently captured.

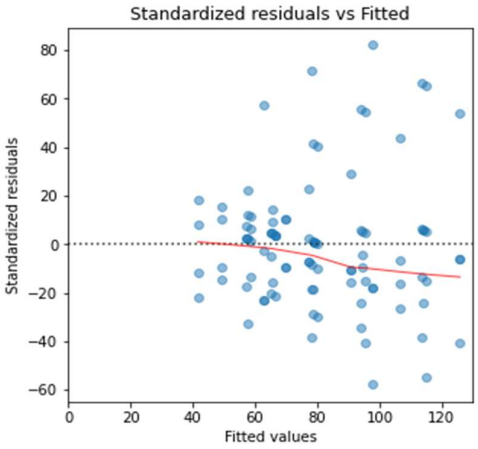

Figure 25 Standardized residuals vs the fitted values for expert judgement data set

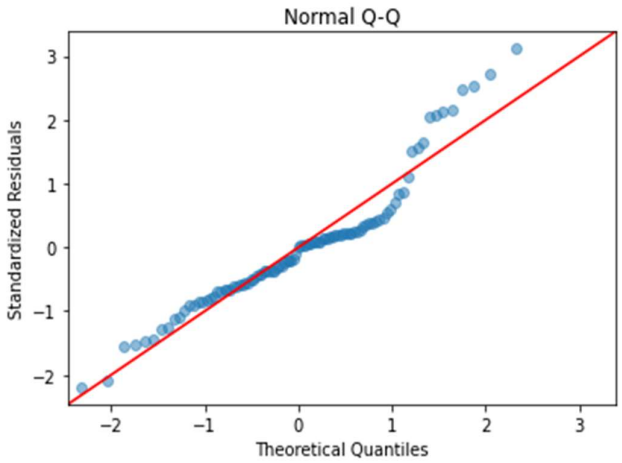

Figure 26 Q-Q plot for expert judgement

The quantiles of the standardized residuals are also plotted against the theoretical quantiles (standard normal variate with a mean of 0 and standard deviation of 1). This is illustrated in the normal quantile-quantile (Q-Q) plot in Figure 26. If the residuals are on the red line and do not deviate from it, the residuals are normally distributed. If it does deviate from the red line, it can indicate that the distribution has a heavy tail or is skewed. From the figure, it can be noticed that there are several points that fall far away from the red line. The distribution also appears to be right skewed. Thus, it is indicated that the errors are not being normally distributed throughout the data set. Thus, the assumption for normality of the residuals that is implicitly made with linear regression analysis does not hold.

From the graphs, it can be concluded that there is variability in the data and also in the behaviour under research, there are heteroscedasticity issues, and the residuals are not normally distributed at all values of the living area. Since data and time is limited in this research, there is no option to find new data with which the regression analyses can be done to gain better results or to involve more experts. Hence, this regression model will be used to calculate the desired living area for synthesized households.

4.9.5 SETUP OF HOUSEHOLD ALLOCATION

For the study area, a distinction is made between two categories of housing units. The first categories are single dwellings, rowhouses, townhouses and duplexes. And the second categories are for apartments and flats. OSM data allows to separate these two types of housing units accurately as well for the study area.

When looking at the property valuations of the first category and comparing them to the property valuations of the second category, it was found that the valuations of the first category are much higher. Therefore, the flats and apartments are less expensive. Based on this finding, an assumption was made that most households with a low income should be allocated to apartment buildings. Despite the income being part of the regression analysis and thus being captured in the desired living area, the income is also explicitly used to ensure that households with a low income are allocated first to the flats.

In the Netherlands, it is also common for apartments to have either no parking or parking for one car. For houses of the first category, there are most likely more than one parking spots available as these houses usually have their own garage or driveway. The second assumption is therefore that households with 0 or 1 cars should be allocated to apartment buildings first.

The third assumption is that households will only be allocated to a house when their desired living area is at least the living area of the house. Out of the list of candidate houses, the house with the smallest living area is chosen. If there are no candidate houses that meet this requirement, then there will be a compromise and the household will be placed in a house that has a living area that is closest to the desired living area but still smaller than the desired living area.

The fourth assumption is that one household can only be assigned to one house, so this does not allow for multiple households to be placed in the same house or residential unit. And it means that when a house is allocated, it should be removed from the set of available houses. The number of households is also equal to the number of houses in the study area. So, each household should be placed in a house after the allocation procedure is finished.

These four assumptions can be translated to rules, which leads to the household allocation becoming a rule-based model. These rules were chosen based on observations of the study area and simplifications. However, these rules can be changed, and other rules can be added depending on the area of interest. Now that the intended working of the model is explained, the code can be written. The next section explains how this is done.

4.9.6 WORKING OF HOUSEHOLD ALLOCATION

First, the houses and apartments are placed in separate DataFrames because the attributes from the houses are different from the attributes of the apartments and the flats in de apartment buildings do not have their own unique OSM ID like the houses do. By separating them, the allocation can be done more efficiently, and this makes implementing the assumptions easier. The apartments data set is also altered so every row corresponds to a flat.

The houses and apartments are sorted in ascending living area. Then a nested loop is used to go through all income groups and households that have 0 and 1 car available. The loop starts with households with the lowest income (income <€10,000) and 0 cars and starts allocating them. Afterwards, it goes through households with the lowest income and 1 car and allocates them too. The loop subsequently goes to the next income group and proceeds in the same way.

Before allocating a household, a selection of housing units is made that satisfy the desired area of the households and have not been allocated yet. The loop tries to place the household in an apartment first if possible. If there are no apartments left, the loop will allocate the household to a house. From the selection of houses and/or flats that satisfy the desired area, the house or flat with the smallest living area is chosen. In the case that there are no candidate houses for a household because the concerned household has a high desired area that cannot be satisfied by the living area of the houses in the study area, the model makes a selection of houses of which the living area is closest to the desired area and picks the largest house for the household as a compromise.

After allocating all households with 0 and 1 car, the nested loop is finished. Then a second nested loop starts that does the same as the first nested loops but now only loops through households of all income groups and with a car availability of 2 and 3+ cars. A counter is also added to both nested loops to keep track of the number of households allocated. After the second nested loop is finished, the counter should be equal to the total number of households in the study area. For the case study, this is 1122 households.

Since there was a lack of data for the case study, it was expected that this lack of data can occur for other areas as well and it can be helpful to still have a method for synthesizing disaggregate data for a study area. Therefore, it was decided to also formulate a random model for instances where nothing is known. The random model does not use the desired area of households or any other rules. It uses a random seed and shuffles a DataFrame in which all the apartments and houses are placed. This DataFrame is then

concatenated to the DataFrame containing the population of households for the study area. Using the random model and running it multiple times samples can be created. Then by using a Monte Carlo Simulator, samples can be drawn to make up the household allocation. The resulting code for the rule-based model and random model can be found in the Appendix V and the Jupyter Notebook file.

4.9.7 RESULTS OF HOUSEHOLD ALLOCATION

The obtained results for the household allocation of the rule-based model and random model were translated to colour coded maps that give an overview of the placement of each of the homogeneous household types. The results are presented on Pages 79, 80 and 81 for the housing units of the first category. The distribution of the second category housing units cannot be colour coded on the map as there are multiple household types in the same apartment building. To still give an overview of what type of households are placed in the apartment buildings, histograms are created. These are depicted in is presented through histograms in Figure 33 on Page 82. The names of the buildings were previously denoted in Figure 23.

From the colour coded maps, the following can be remarked:

- The outer edge of houses on the left of the study area are allocated to mostly couples with and without children in the rule-based model, whereas for the random model these houses have been allocated to a mixture of households including one-person households. Since the outer edge consists of houses that have high property valuations and living areas, it seems unlikely that there would be one-person households residing in these houses. It is thus regarded that the rule-based model led to more realistic results than the random model for the household composition.
- This same outer edge of houses was also predominantly assigned to high income households in the rule-based model. The random model allocated these houses even to low-income households, which is not in line with expectations. There could be instances where this occurs but then it is more likely through inheritance or other circumstances, and these would still be exceptions to the rule and not what often occurs.
- Most of the households seen in Figure 29 have incomes higher than €20,000 in the rule-based model. There are only a few low-income households placed here and this could be because all the flats in the apartment buildings were unavailable and had been allocated already. In the random model in Figure 30, the households are mixed, and all types of income groups can be found in the centre.
- For the car availability, the households with 0 and 1 car are found in the centre of the study area and the outer edge has mostly households with 2+ cars for the rule-based model in Figure 32. The random model in Figure 31, as expected, shows a mixture of household types in all categories of the car availability.

From the histograms made for the apartment buildings, the following can be stated:

- All household composition types can be found in the flats in the rule-based model, whereas in the random model only one person households and couple without kids have been allocated. Seeing as this is random, it coincidental that there are only two types of household compositions in the flats. It could also be because one person households and couple without kids are the biggest groups in the synthesized population with respectfully 364 and 315 households belonging to this group. Couple with kids' households occur 309 times, one parent households occur 113 times and other multiple person households are the smallest group with just 21 households.
- For the household income in the rule-based model, there are only households allocated with an income of less than €20,000. Meanwhile in the random model, all household income categories

can be found in the apartment buildings. This is not necessarily unrealistic because there can be households that prefer apartments over housing units of the first category.

- The car availability of the households allocated to flats in the rule-based model is mainly 0 and 1 car. And in the random model all households with 2 cars and less were allocated. There were no households with 3+ cars but this could be because this is the smallest group existing of only 54 households, so the chances are smaller for picking these households.

From these remarks, it can be concluded that there is value in the rule-based model and that this leads to results that are more intuitive and realistic than the random model. However, the random model is also useful for areas where no data is available. Both models can provide a transport model with a disaggregated synthesized population having a fine spatial distribution and are able to be used for adding spatial detail.

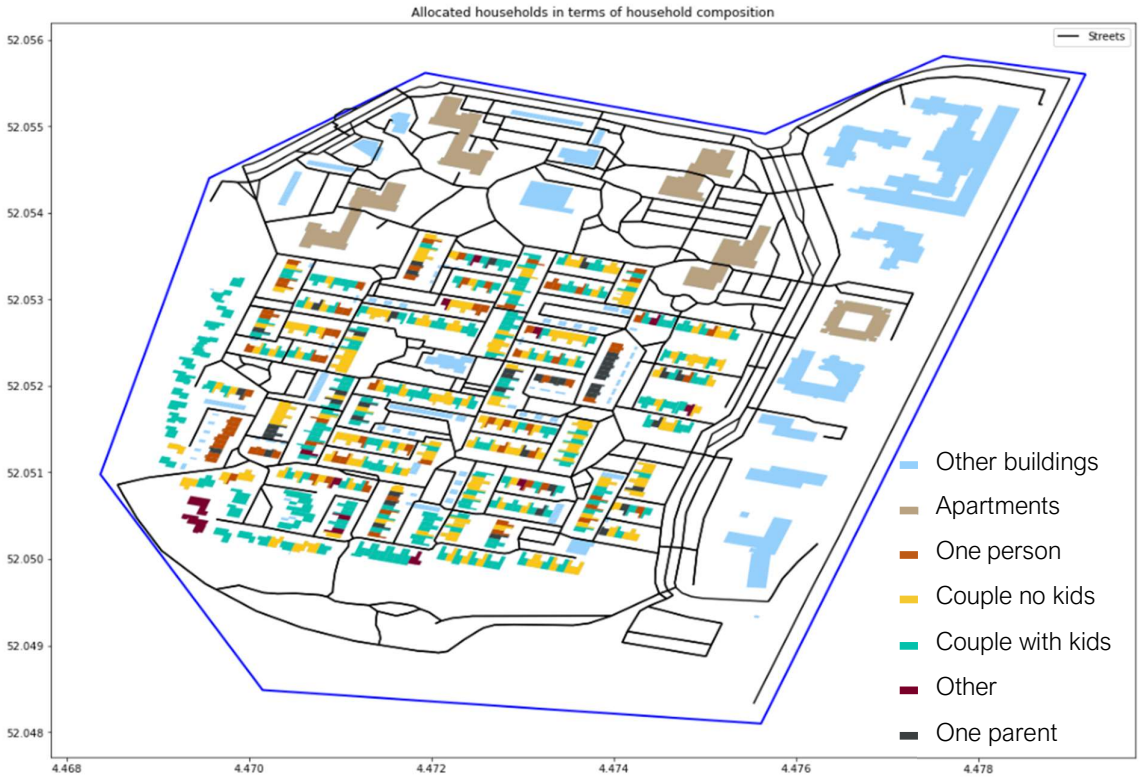

Figure 27 Results for the household composition for the rule-based model

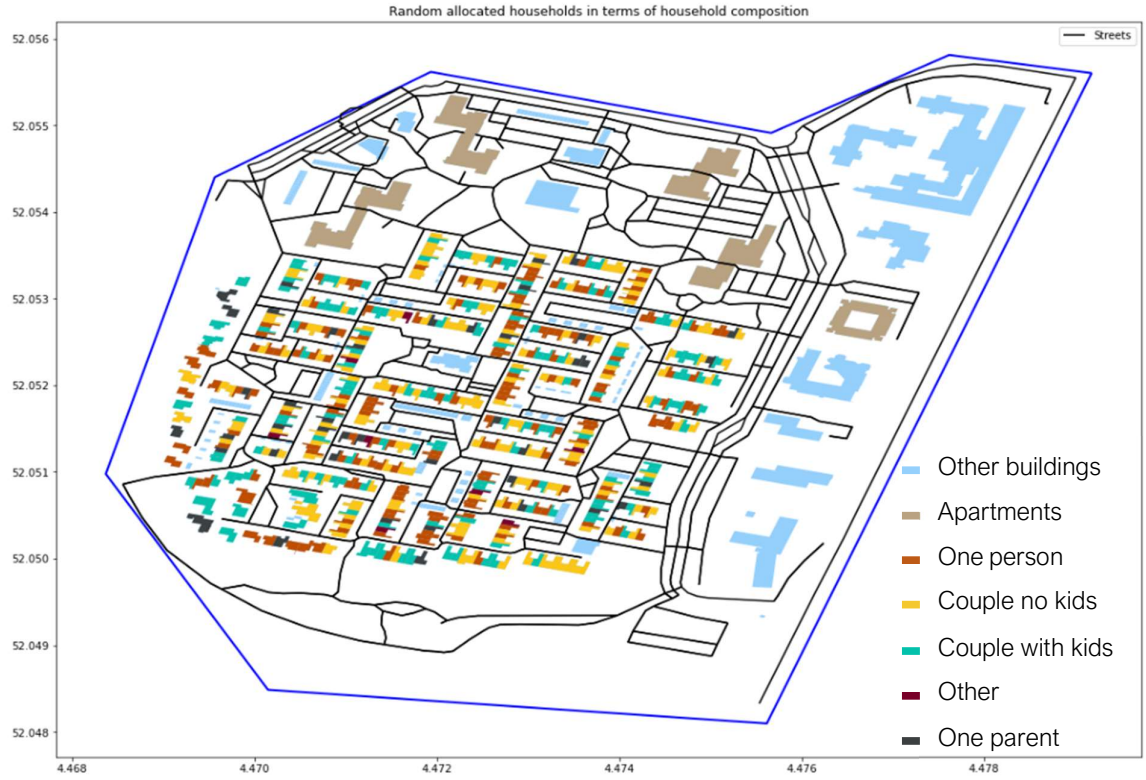

Figure 28 Results for the household composition for the random model

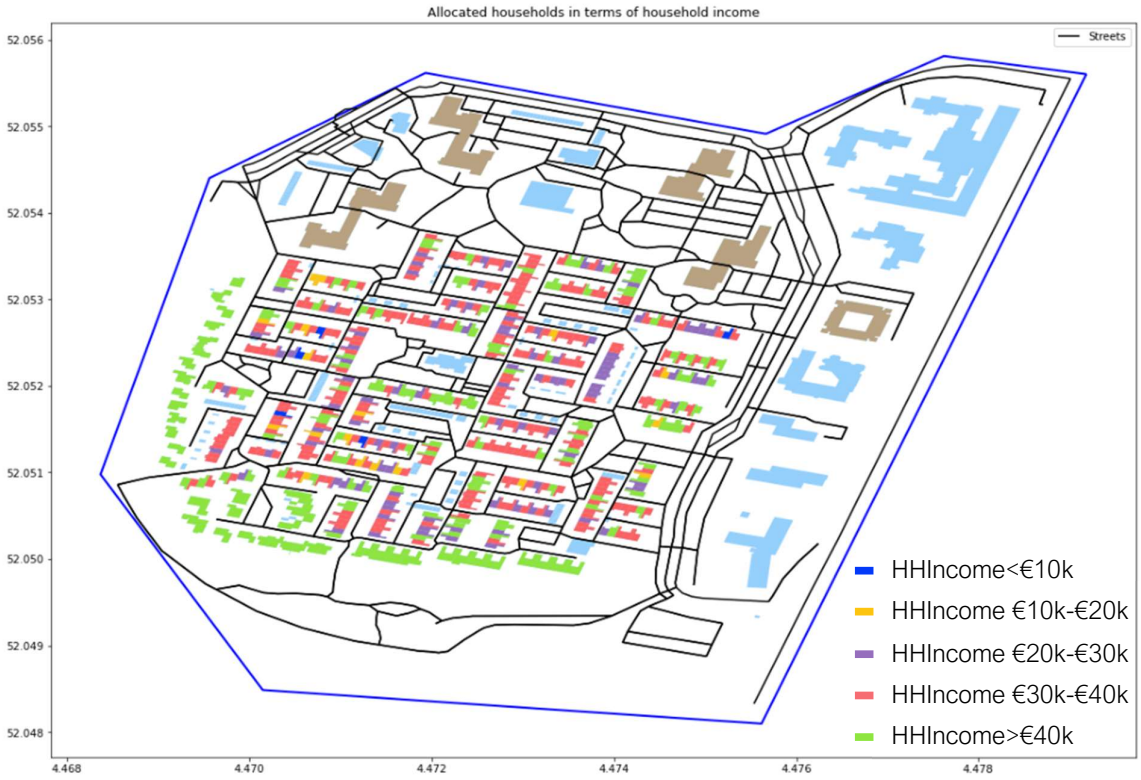

Figure 29 Results for household income for rule-based model

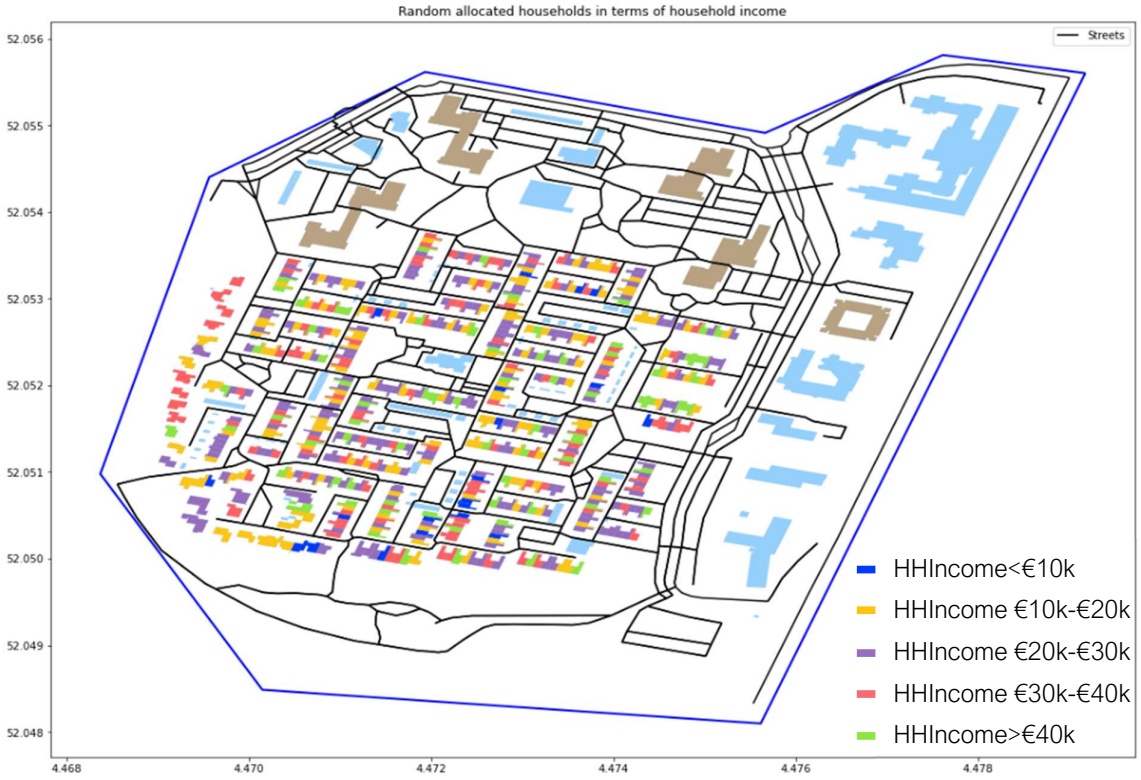

Figure 30 Results for household income for random model

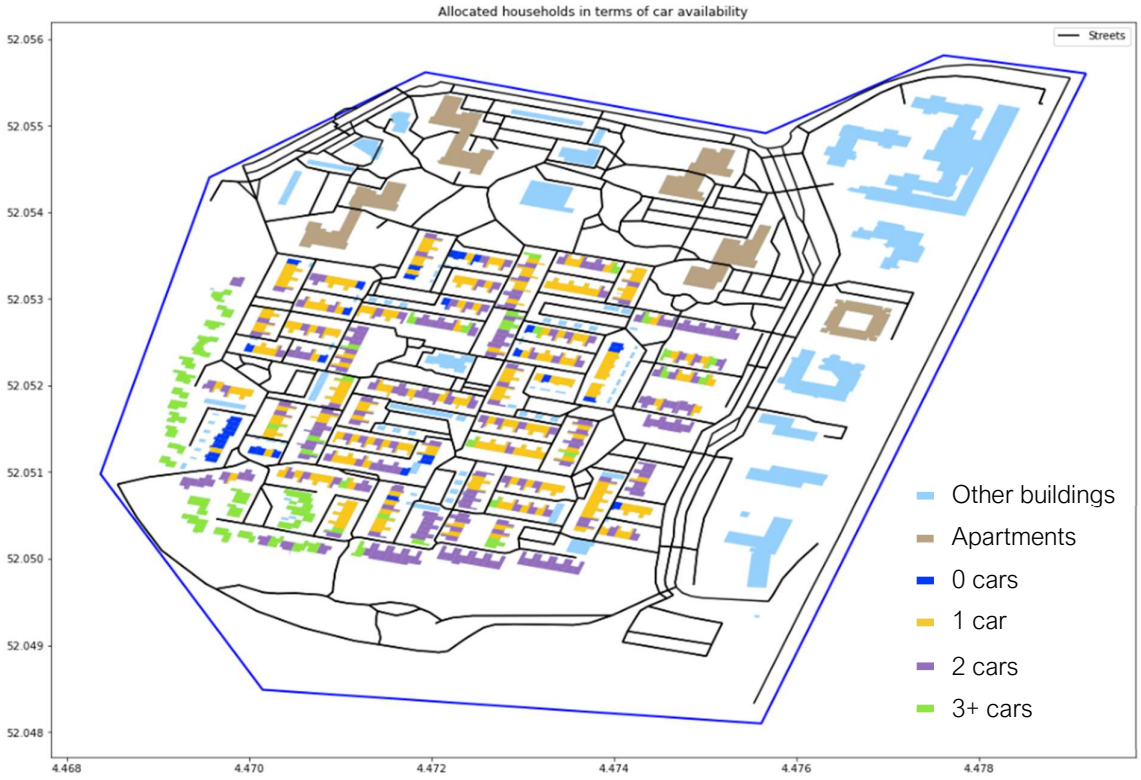

Figure 31 Results for car availability for the rule-based model

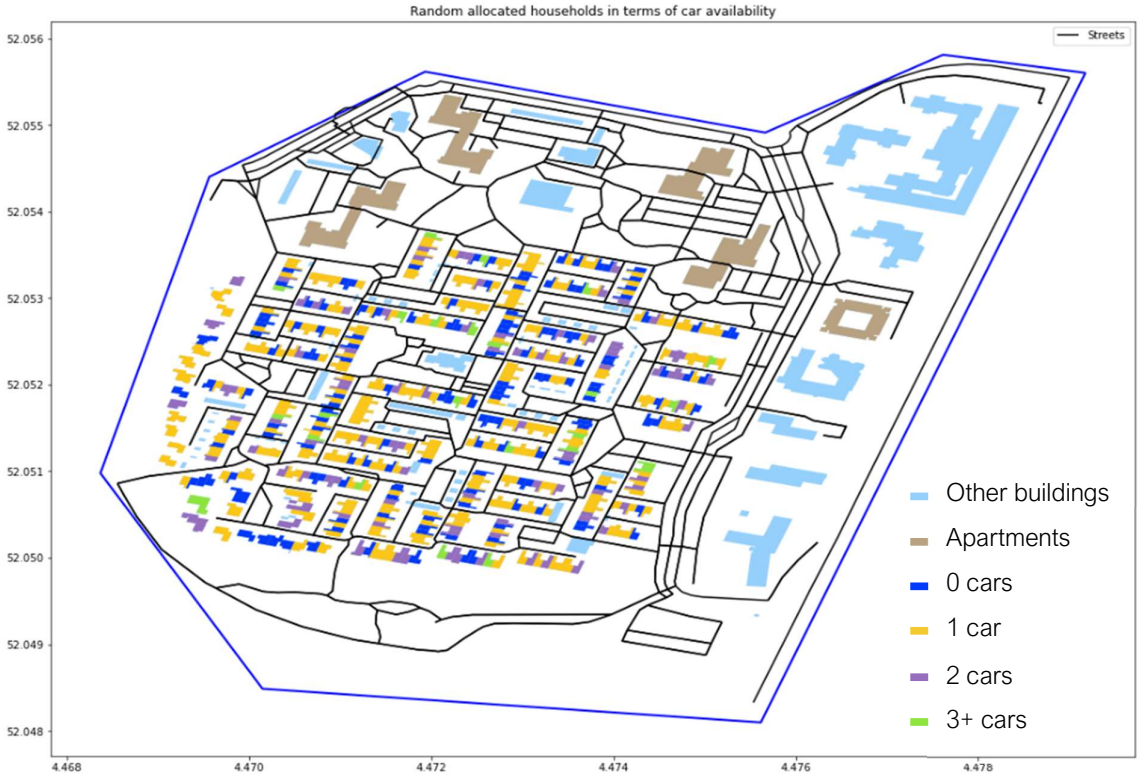

Figure 32 Results for car availability for the random model

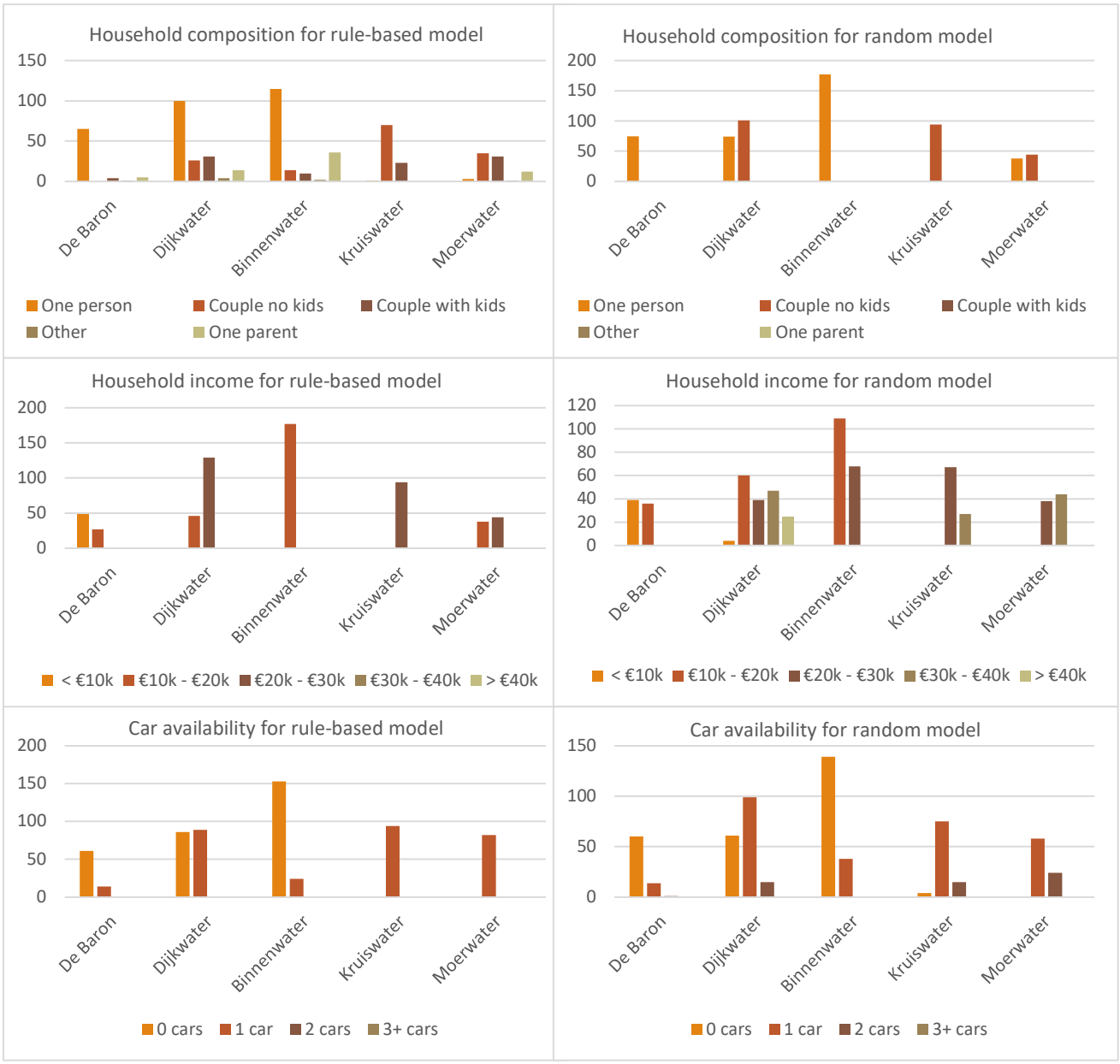

Figure 33 Results of apartment buildings for rule-based model (left) and random model (right)

4.10 VALIDATION

This section will shed light on the validation used for the case study. This includes the population synthesis as well as the household allocation.

4.10.1 VALIDATION OF IPF PROCEDURE

Since data was already scarce, the full external validation cannot be carried out. The internal validation is however carried out. For internal validation, the Pearson correlation coefficient is calculated. If this is equal to 1, then the methodology is internally validated. The formula for the correlation coefficient is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3)$$

Where,

r : Correlation coefficient

x_i : Values of x-variable in sample

$\bar{x}$: Mean of x-variable in sample

y_i : Values of y-variable in sample

$\bar{y}$: Mean of y-variable in sample

The generated synthetic population for the study area given in Appendix K were aggregated to get tables that are in the same format as Table 8, Table 9 and Table 11. For each of the control variables the sum of the rows/columns were calculated and compared to the marginal totals that were used as input. These were the univariate distributions. The x-variable is in this case the control variable in the synthesized population and the y-variable is the marginal total for the control variable that was derived from population census data. The formula is then filled in for all three control variables. The results can be seen in Table 13. For all the variables the correlation is 1, which is in line with expectations. It is hereby concluded that the control variables are internally validated for this IPF procedure.

Table 13 Correlation coefficients

Household composition				Household income				Car availability			
x_1	549.22	y_1	17,400	x_1	79.88	y_1	2,530.72	x_1	501.02	y_1	15,873
x_2	476.62	y_2	15,100	x_2	459.70	y_2	14,563.92	x_2	786.83	y_2	24,928
x_3	467.15	y_3	14,800	x_3	550.66	y_3	17,445.65	x_3	324.45	y_3	10,279
x_4	32.56	y_4	1,000	x_4	352.59	y_4	11,170.64	x_4	82.70	y_4	2,620
x_5	170.45	y_5	5,400	x_5	252.17	y_5	7,989.07	$\bar{x}$	423.75	$\bar{y}$	13,425
$\bar{x}$	339.0	$\bar{y}$	10,740	$\bar{x}$	339.00	$\bar{y}$	10,740	r	1		
r	1			r	1						

The external validation is a more complex process and as explained before would either require a survey or an external data set. Finding data in the suitable format for the IPF procedures in the first place, was already a challenge with just one control variable being directly available for Zoetermeer (household composition). The car availability was only available for the Netherlands in percentages and the household income was only available in quintile groups and it concerned the standardized income.

When looking for data, there was no separate external data set found to validate the population synthesis for Zoetermeer or for the study area. An effort was still made to do a less strict validation, which would be in the form of comparing averages available in the V-MRDH data set and the data set from the Databank of Zoetermeer to averages calculated from the synthetic population.

When looking at car availability, the V-MRDH had an average of 0.875 and if it is assumed that the fourth category (3+ cars) exists only out of 3 cars, the average from the population synthesis becomes 0.993. This differs from each other and a possible reason for this can be that the percentages of the Netherlands are not entirely representative of the case study. This could be because large areas in the Netherlands depend more on the car whereas in the study area, this is less the case.

For the income, because the category with >€40,000 does not have an upper bound, an assumption must be made. This category consists of the categories €40,000 - €50,000 and >€50,000 of OViN. There were 33 observations for the income group of €40,000 - €50,000 and 16 observations for the income group >€50,000. An exact average cannot be calculated when using categorical variables. The income consists of intervals of €10,000 and the last category is every income that is more than €40,000. The midpoints of each interval will be used along with the frequencies of the categories to fill in the formula:

$$mean = \frac{\Sigma f*m}{\Sigma f} \quad (4)$$

With f being the frequencies and m being the midpoints. For the last interval €50,000 is chosen as a midpoint based on the observations of OViN. Filling in the formula then gives a mean of €26,000. According to CBS, the mean standardized disposable household income for Zoetermeer is €29,000 for 2016 (CBS, 2020).

For the household composition, the databank of Zoetermeer had frequencies for the year 2015 for Meerzicht Oost. This data was then scaled down to the study area by using a multiplier. The comparison of the frequencies are shown in Figure 34. Differences can be seen and this is most likely because the targets and sample data were not for the specific study area. If this had been the case, a closer match would have been expected. The biggest issue is that there is no ground truth to which the synthesized population can be compared. Therefore, there is currently no method to properly validate the results of the population synthesis externally.

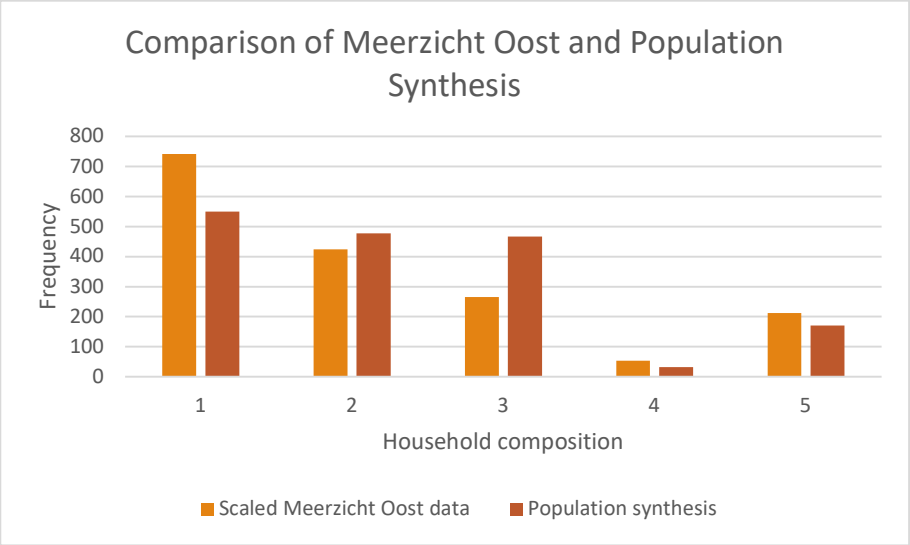

Figure 34 Comparison of Meerzicht Oost and population synthesis

4.10.2 VALIDATION OF HOUSEHOLD ALLOCATION

Due to a lack of data the validation for the household allocation can only be partially done. The data set that will be used for this is the housing survey of the year 2015 (WoON 2015). It should be noted that in this data set, there are no one person households, some combinations of attributes of households are not represented and the income is not standardized. Hence, the one person households cannot be validated. Before using this data in the rule-based model, it needs to be processed.

First, the data set is filtered to only include observations from Zoetermeer. Then all the observations for which the household composition was unknown were removed. Afterwards, the seven household composition categories were converted to household composition types used in the IPF and household allocation procedure:

- Pair without kids remains the same
- Pair with kids remains the same
- Pair with children and others becomes other multiple person household
- Pair with others becomes other multiple person household
- One parent household with children remains the same
- One parent household with children and others becomes other multiple person household
- Other composition is also placed under other multiple person household

The disposable income also has to be converted to the standardized disposable income. This is done by using the equivalence factors defined by CBS (2019). For the most common household compositions a crosstabulation is given by CBS that specifies the factor by which the disposable income has to be divided to calculate the standardized disposable income. For all other household compositions, the following formula should be used:

$$G = \frac{B}{(V+0.8*K)^{0.5}} \quad (5)$$

Where,

G : Standardized disposable household income

B : Disposable household income

V : Amount of adults in the household

K : Amount of children (age<18 years) in the household

Households for which the household size and number of kids were not specified, had to be removed from the data set as well because the standardized disposable income could not be calculated if these variables are unknown. After the conversion to standardized disposable income, the income was placed in the same five categories defined in the household allocation and population synthesis.

In the WoON 2015 data set, the living area is given for each of the household types. For the household allocation, it was decided to group all homogeneous households together and take the median of their living area and assign this value to households in the synthesized population that have the same household composition and income as the homogeneous household types. The median was chosen as this is a robust measure of central tendency and gives a good indication of the spread of data.

In the WoON 2015 data set, there are combinations of certain household compositions and household incomes that have no or little observations. To check if there are enough observations, the sample size is calculated using the following formula:

$$n = \left(\frac{Z\alpha_{/2} \times \sigma}{MOE}\right)^2 \quad (6)$$

The chosen confidence level is 90%, so the corresponding Z-score is 1.645. The chosen Margin of Error (MOE) is 12, meaning that the sample mean must be within 12 units of the true mean. For the standard deviation, the best estimate is chosen and this is the standard deviation of the WoON 2015 sample. This is equal to 36.12. Filling in the formula gives a sample of 24.5 observations. Therefore, when the observations of each homogenous household type (household composition and household income) in the WoON 2015 data set is 25 or higher, the median for those observations is calculated.

For homogeneous household types that have less than 25 observations, a regression analysis will be used that is estimated using the WoON 2015 data set. The regression analysis was carried out and the result was the following equation:

$$\begin{aligned} Desired\ Area = {} & 105.1 + \left(-9.91 * HHComp_{type2}\right) + \left(-23.40 * HHComp_{type}\ \right) \\ & + \left(-9.33 * HHComp_{type5}\right) + \left(6.43 * HHIncome_2\right) + \left(14.88 * HHIncome_3\right) \\ & + \left(29.82 * HHIncome_4\right) + \left(53.14 * HHIncome_5\right) \end{aligned} \quad (7)$$

In the intercept, the households consisting of couple with kids and an income of less than €10,000 are captured. According to the input data, the households that do not consist of couple with kids, all desire a smaller living area when only taking the household composition into account. When the income is also considered, as the income increases, the desired living area also increases. This is in line with the expectations.

The adjusted R-squared value is 0.201 and log-likelihood is -2506.3. For human behaviour, this is a normal model fit. Three of the variables (Inc10-20k, Inc20-30k and Inc30-40k) have p-values that are insignificant. The variables will still be kept as this is the closest estimate available. Diagnostic plots have also been made to explore the residuals of the data. In the standardized residual plot in Figure 35, it can be noticed that when the living area increases, the standardized errors also seem to increase. This indicates heteroskedasticity and that the homoskedasticity assumption does not hold for this regression analysis.

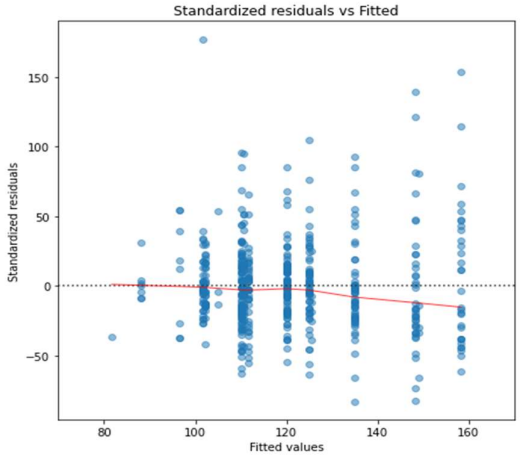

Figure 35 Standardized residuals vs fitted values for validation data set

In the normal Q-Q plot in Figure 36, it is also evident that after a certain value, the residuals are no longer normally distributed. This occurs around the second theoretical quantile. The distribution also appears to be right-skewed and thus has a longer tail to its right. For many of the data points, it does follow a normal distribution according to the normal Q-Q plot.

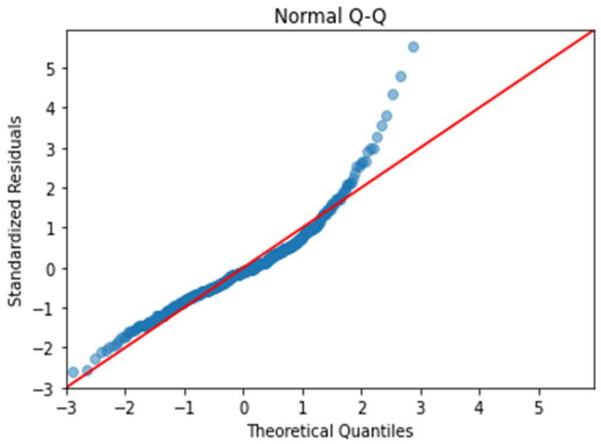

Figure 36 Q-Q plot (standardized residuals vs. theoretical quantiles) for validation data set

Since one person households are not included in the data set, an assumption was needed to still allocate one person households in the synthesized population using the validation data set. The chosen method was to include a dummy value for households consisting of one person and with an income of less than €10,000. The dummy value is chosen as 60 m^2 and for the rest of the income groups, the coefficients are used from the regression analysis with the validation data set. So, this resulted in the following equation for one person households:

$$Desired\ Area = 60 + (6.43 * HHIncome_2) + (14.88 * HHIncome_3) + (29.82 * HHIncome_4) + (53.14 * HHIncome_5) \quad (8)$$

Formula 8 is then used to calculate the desired living area. For all other household compositions, formula 7 is used to calculate the desired living area. The desired living area based on expert judgement and the WoON 2015 (validation data set) are then compared to each other for every household. The data sets were sorted by household composition. This comparison is presented in Figure 37. Households with ID 10000 up to and including 10363 are all one person households and are not present in the validation data set. These households have a dummy value so comparing them to the expert judgement data set is nonsensical as the dummy value chosen is also based on expert judgement. For the rest of the households, the experts tended to underestimate the living area. The living area seemed to be much higher than the experts anticipated. Since expert judgement elicitation almost always results in biased estimates, this is not a strange trend to see.

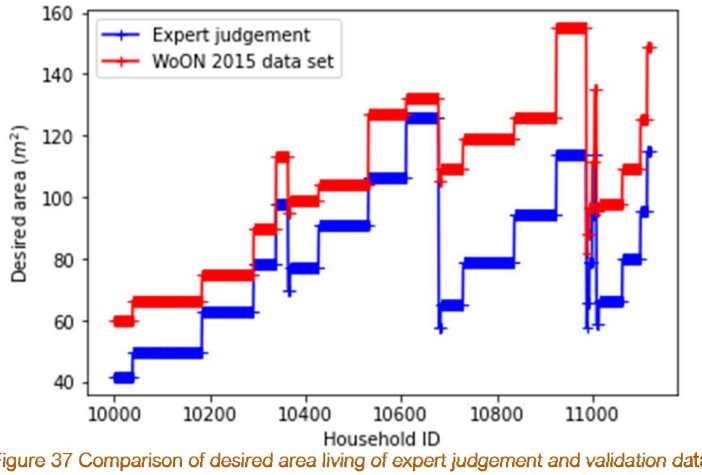

Figure 37 Comparison of desired area living of expert judgement and validation data

The results of the household allocation with the validation data set are given in Figure 38 on Page 89. For the household composition, it is seen that more couple without kids' households were placed in the outer edge with the validation data compared to the expert judgement data. The centre looks the same in both models, with just minor differences in the placement of certain household composition types.

For the household income, it appears that in the model with the validation data set more households with an income of €30,000 - €40,000 were placed in the outer edge. Whereas for the expert judgement model results, the households were majorly of the income group of €40,000 and up. And in the centre, there are just minor differences in the allocated households with the validation data set allocating households of income group €30,000 - €40,000 to houses that were previously (in the expert judgement model) allocated to the highest income group.

The same trend of the household income can be seen for the car availability. The outer edge is again mainly allocated to households having 2 cars in the model with the validation data set. While for the expert judgement data set, these houses were assigned to households with three or more cars. Also, the houses assigned with households with two cars in the expert judgement data, are now allocated to households with one car (for example the bottom right).

As with the expert judgement data set, histograms are made again for the apartment buildings for the validation data set in Figure 39 on Page 90. Looking at the household composition, there are only minor differences. All the household composition types can be found in De Baron in the expert judgement model. While only one person households were allocated to this building in the model with validation data set.

For the household income, a similar distribution can be seen for both the models with the expert judgement data set and the validation data set for all apartment buildings except the Binnenwater flats. In the validation model results, households with an income of less than €10,000 were also found here while none were allocated in the expert judgement model results.

The distributions seen for the car availability is comparable to the household income. There seems to be a lot of similarity between the two outputs. An obvious difference can be noted for the Moerwater flats. In the expert judgement model, there are no households allocated with 0 cars. On the other hand, there are households allocated with 0 cars to the Moerwater flats in the model with the validation data set.

From these observations, it can be concluded that for the houses (category 1), there are small differences especially in the outer edge where the allocation differs between the expert judgement and the WoON 2015 data sets. For the apartment buildings, the allocations were similar.

To check the household allocation even more thoroughly, the living area was analysed for the desired and allocated households. The desired living area is the living area as calculated by either the regression analysis or the median of the homogeneous household groups in the WoON 2015 data set. And the allocated living area is the living area of the house the household has been placed in after executing the household allocation. the households were sorted in ascending desired living area. The graphs are illustrated in Figure 40 on Page 91.

For the expert judgement model results (top graph), an upward trend is observed in both the desired and the allocated living area. For this data set, there were no instances where a compromise was needed because houses were always able to satisfy the desired living area. The same cannot be said for the validation data set. In this graph (bottom), there are points where the allocated living area is below the desired living area. The peaks in the allocated living area line are houses that have a big living area. Apart from these compromises, the graph also shows an upward trend.

Figure 38 Results of household allocation for validation data set

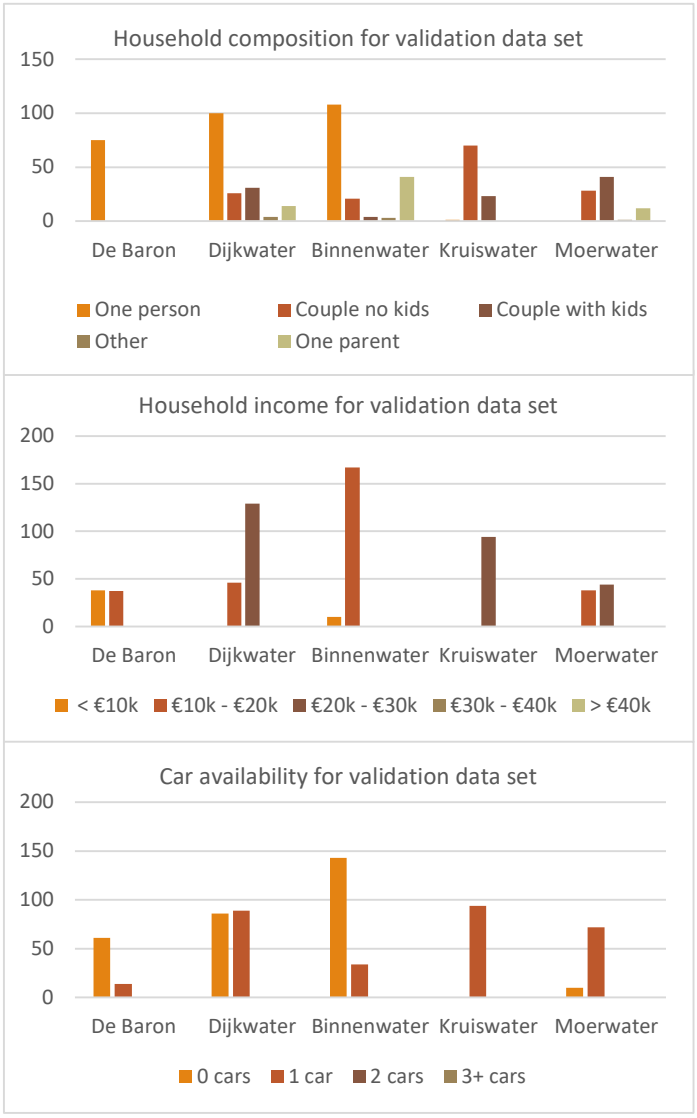

Figure 39 Apartment buildings for the validation data set

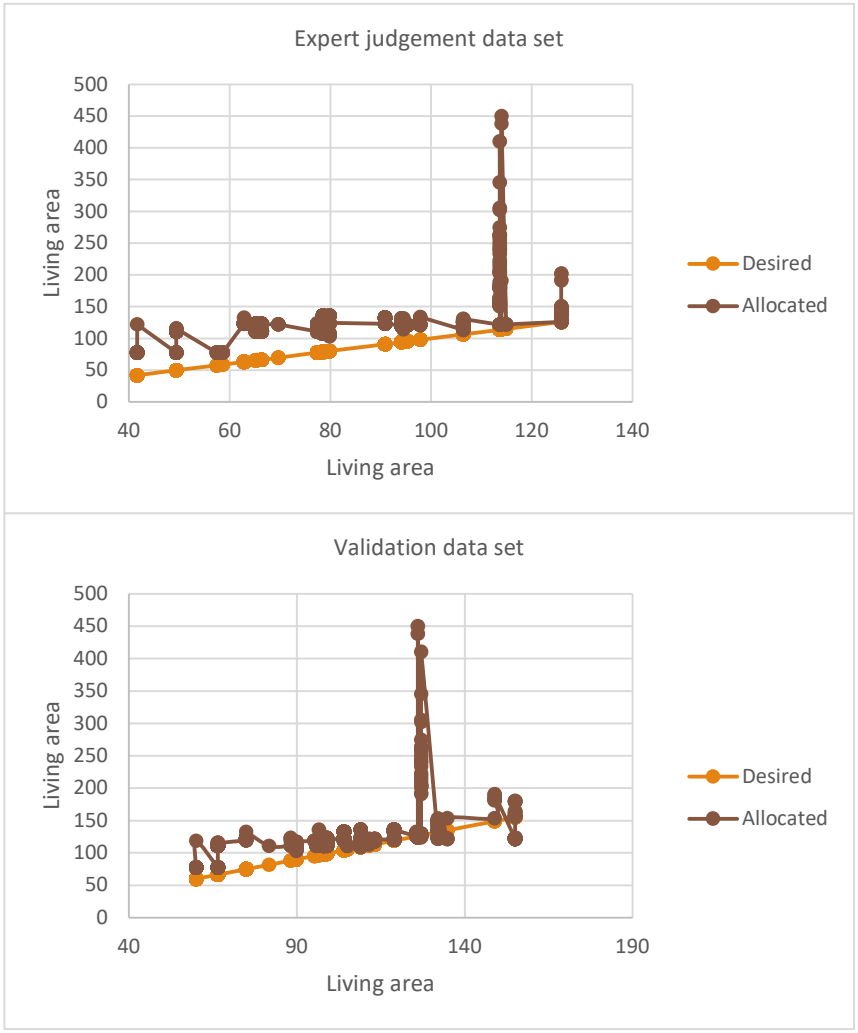

Figure 40 Comparison of living areas in expert judgement and validation data sets

5. DISCUSSION

In this chapter, the results of this research are discussed with interpretations and implications. This will be done in terms of the three components of this research which are population synthesis, OSM data and household allocation. Lastly, the uncertainties and limitations associated with the methodology are described.

5.1 POPULATION SYNTHESIS

In the methodology that has been developed, the first step is the population synthesis. In literature, there seems to be limited comparisons of different population synthesis techniques, and this makes choosing a method difficult.

From the implementation of the single-level approach, it can be stated that synthetic reconstruction methods and specifically IPF has flexibility in terms of data requirements. The input data was not available at the geographical scale that was needed and yet it was still possible to downscale the population and get estimates for the synthetic population. Although, the assumption made that the distributions of household characteristics remain the same at the level of a city and the level of a neighbourhood does not have to be true and most likely will be different. This affects the accuracy of the synthesized population.

Most research on IPF do not discuss the data availability issues and therefore do not provide strategies to solve this issue when it occurs. So, it can only be assumed that the majority of studies profited from rich data sets. With the constricting requirement to only use open-source data in this research, it led to finding solutions to mitigate this issue. However, it is also important to note that public data often have filters, error terms and are rounded. Hence, open-source data itself may not be accurate and can introduce uncertainty by itself. The IPF has also proven to be a good tool when there are no multi-dimensional crosstabulations available as this allowed to still obtain constraints in the case study.

Farooq et al. (2013) reported that the IPF procedure has scalability issues. However, this was not specifically found in the case study. The study area could have easily been the size of a city with more control variables and the IPF algorithm would still be able to generate a population. It should be noted that in various studies, the IPF algorithm has been used for much bigger study areas than the one used in this case study. And that it concerns single-level fitting in the case study as well, which in general can synthesize with more control variables (no limit on this has been reported in literature).

As reported by Choupani and Mamdoohi (2016), the multilevel fitting synthesizers have problems converging when using more than four control variables, but this was not tested in this research. Upon reviewing the literature, it appears that the bigger problem is that these multilevel fitting synthesizers are comparable to black boxes, which makes comparison hard. There are still models that use these multilevel fitting synthesizers and use more than four control variables and are still able to synthesize populations. The pitfall is that there is little to no documentation on these synthesizers, it could be that these algorithms do not converge but instead reach their maximum tolerance or iterations and the resulting population is then not the best fit.

Another problem in terms of scalability with the amount of control variables are the data requirements. With each addition of a control variable, the dimension of the input data increases. Since the seed data is in the form of microdata containing multiple attributes, this did not pose any problems in the case study. However, finding aggregate data having more than three dimensions was difficult for the case study because aggregate data is not often collected in that manner but rather as unidimensional marginals.

The generated population with IPF consisted of fractions of households and these had to be rounded as well. This makes the household allocation process easier but as reported by Lovelace et al. (2014) can lead to breaking correlation structures. They also suggest more sophisticated methods for 'integerisation' than using a sum preserving rounding method used in the case study.

It is also important to note that a population can be generated for any study area if there is aggregate and disaggregate (when using sample-based methods) data available for the area. In that sense, the case study area was not suited for population synthesis as data spanning the specific geographical area was not available. Normally, population synthesis is also not carried out for small areas and this could most likely be because of a lack of data. Applications nowadays do need small area population synthesis, so it is recommended that the data is collected at this level. This may result in more accurate synthesized populations. Another reason this data should be collected at this level is for use in the external validation. In the case study, there was no ground truth available for the control variables in the study area, which hindered assessing the performance of the population synthesis.

5.2 OSM DATA

The quality of OSM data was tested in this research, however adjustments had to be made to improve the thematic accuracy. And there are still missing tags that are vital when an accurate household allocation is desired such as the living area, number of flats and number of floors. This could be realized by improving the link with the BAG register because currently not all information stored in the BAG is being utilized in OSM.

Linking POI's to individual flats would also make the household allocation easier. As of now, these points provide the full addresses (house numbers, postal code, street name) are not linked to the buildings and thus some information is lost and the flats in the building had to be manually separated in individual units. Some assumptions had to be made when these units were separated such as that the units are all equal in size and property valuation which is not the case in reality.

There are tools and methods to assess the quality of OSM. However, this might not be feasible to check if the study area is bigger or if it must be done for multiple study areas. The field research gave pivotal insight for specification of tags in OSM and allowed for the correction of features of OSM elements. With time, it is expected that the quality of OSM will continue to increase because of the densification process mentioned by Arsanjani et al. (2015). However, as of now it is not of sufficient quality to directly implement in transport models without limitations and compromises on the accuracy.

The case study used is mainly residential and had many uniform houses (single dwellings and row houses) and five apartment buildings. For average residential neighbourhoods, the case study is representable, and it can be assumed that the methodology is applicable to other residential areas. However, when the study area is commercial (city centres) or industrial or does not have uniform houses, the case study will not be representable. It will be important to have a proper calculation of living area of the houses in the study area and preferably to not use a calculation for this but rather extract the living area from the BAG in OSM. It is also essential that the tag (amenity and/or building) specifying the use of the buildings is accurate to distinguish residential buildings from commercial buildings. It is also known from literature that urban areas are mapped in greater detail than non-urban areas (Neis, Zielstra, & Zipf, 2012). The case study was in an urban area. In rural areas, the quality of OSM might not be sufficient for use in the household allocation.

5.3 HOUSEHOLD ALLOCATION

As input for the household allocation, housing data is needed. This was not available as open-source data so to overcome this issue, expert judgement was used. The expert judgement used in the case study considered too little experts and did not elicit the opinions over iterative rounds. If this had been done, the results might have been closer to reality or the housing data from the WoON 2015 data set.

For the household allocation, regression analysis was used in the case study, but it became clear that the assumptions for homoskedasticity and normality of the residuals did not hold. The variability seen in the behaviour of households in relation to the house they reside in, makes it difficult for implementing statistical methods such as regression analysis, IPF and choice modelling. There might be opportunities for more sophisticated techniques such as machine learning algorithms to allocate households as these may better deal with the high variability.

The predictor variables used in the regression analysis are also not the only deciding factors when households are picking a house. More variables could be added that can either be retrieved from OSM (such as proximity to schools, offices, grocery stores, etc.) or in the population synthesis itself (like household size, labour force association, number of children, etc.). Placing them in the population synthesis would mean that aggregate and disaggregate data need to be available for these variables and this may increase the complexity of the IPF procedure. In the household allocation in the case study there was only distinguished between houses and flats, but this can be expanded to include social houses, privately owned or rented to make the household allocation more accurate.

The results generated with the expert judgement data set and the housing survey data set do not widely differ from each other. The high variability in the behaviour of households and other social aspects that play a role, make it difficult to decide what makes a specific spatial distribution of the synthesized households plausible. Both distributions are realistic; intuitively both distributions could have been found for neighbourhoods and can be used as input for transport models. In contrast to the existing transport models, this spatially distributed synthesized population has value as this is concerns disaggregate data and is at a fine spatial resolution that enables analysis in detail as well in transport models.

In the case study, there were several assumptions made. In residential areas, it is often the case that flats have a lower property valuation than single dwellings, rowhouses, townhouses and duplexes. It is also common for apartment buildings to have one parking spot available per unit. So, it is expected that these assumptions are realistic, and they can be used in other residential neighbourhoods as well. However, the assumption that one household resides in each house or residential unit does not always hold in reality, so the application of the household allocation in the case study does not account for instances where this assumption does not hold.

Different assumptions and household allocation techniques will lead to different spatial distributions. The differences in the households to be allocated will also lead to a different spatial distribution. To assess the robustness of this method, it is recommended to also perform a sensitivity analysis.

5.4 REFLECTION ON METHODOLOGY

The methodology developed has proven to be a good guideline for synthesizing a population with spatial units through the implementation in the case study. Since there is no framework as of now in literature that specifies all the steps for population synthesis and using OSM, this is a valuable methodology. There are aspects such as the filtering of houses and the household allocation that were problematic because of the quality of OSM and the absence of an open-source housing data set. For these issues, the case study has

suggested ways to still implement the methodology proposed, proofing that the developed framework is flexible. This makes the methodology good for countries where data availability may be a constraint.

There could be an addition to this framework by including a feedback loop from the household allocation to the population synthesis. The households that would have to compromise on the living area in the study area would then be fed back to the population synthesis. And then these households could be taken out of the population and placed into the synthesized population of neighbouring areas where the households their requirement for the living area can be satisfied. This was not done in the current case study as this is a simple proof of concept. It is also not included in the methodology because this comes down to the researcher's preference whether to have compromises being made or not.

5.5 UNCERTAINTY

Given the lack of data, there is uncertainty in the model inputs and outputs. In an ideal situation, there would be data available, and the uncertainty would be reduced in the input and output of the model. In the whole procedure, most uncertainty lies in the household allocation because the IPF procedure itself is deterministic. The uncertainty of the IPF is mostly in the input data as this was not available for the geographic area and had to be scaled down with the assumption that the distributions of household characteristics will remain the same at the different geographic levels (at the level of the Netherlands, Zoetermeer and the Meerzicht Oost neighbourhood).

In the household allocation, there is uncertainty in the living area retrieved from OSM. This is indirectly calculated by comparing the area of the polygon to the living area reported in BAG. But there are some differences and thus the factor can be adjusted and will result in a slightly different living area. Ideally, the link to the BAG would be better and this could then allow extraction of the living area directly from OpenStreetMap Data. There is also uncertainty in the expert judgement because this leads by definition to subjective distributions of the households. In the survey, the confidence levels of the experts were also included, this could help to aggregate the results and make upper and lower boundaries for the living area. The regression analysis used introduces uncertainty as well because some variables were statistically insignificant and there could be more variables included.

5.6 LIMITATIONS

Due to resources in terms of time and data, the following limitations were found to be applicable in this research:

- For the house allocation, an assumption is made that there resides one household in one house/residential unit. In actuality, there are cases where this assumption does not hold and then the methodology will not be able to assign multiple households to a single house. This must then be adjusted when there are known cases of multiple households in one residential unit.
- The data from OSM and the population census data introduces uncertainty when these are not accurate or for the specific geographical location which limits how well the methodology will perform. The methodology heavily relies on OSM to provide the spatial units and this data has proven to not be up to par.
- The age of the data is also a limitation. The population census data is from 2015 and 2016 whereas the OSM data is from 2021. This leads to making current assumptions based on old data which is a limitation.

- The data collection method currently does not collect microdata at the level of neighbourhoods, which is also a limitation and requires assumptions to be made when using data from bigger geographies. In some instances, the microdata is collected but not accessible to the public which is also a limitation for this research.
- The current methodology does not take social or environmental factors that do influence the housing situation of households into account.
- The public data that was used is also a limitation because this data is often changed for privacy reasons.
- The different institutions that collect data are also inconsistent. As the collected data sets differ regarding the spatial aggregation, time, and definitions. This limits the usability and reality of the data sets.

6. CONCLUSIONS AND RECOMMENDATIONS

This chapter gives a summary of the conclusions that can be made in this research, the research questions are also answered and lastly recommendations are proposed.

6.1 CONCLUSIONS

As the computational power of computers increase and open-source data becomes more accessible, new opportunities arise for microsimulation models. All these models require a realistic synthetic population. This population can be further enriched and ready to be implemented in transport models if they include a spatial distribution of the households as well. This results in a population for which the home end of trips/tours and in activity schedules is known. The distribution can be made realistic and accurate by taking attributes of households and houses into account in the household allocation. Crowd-sourced OpenStreetMap data has shown potential to be a viable data source in literature and was brought into this research to provide the spatial units for the synthesized population.

Research questions were formulated to help establish a methodology and apply this methodology in a case study. The set of sub questions will be answered first using the literature review and the analysis of the results. Afterwards, the main research question will be addressed.

Sub question 1a: What population synthesis technique can be selected for this research?

Since there was more research done on synthetic reconstruction approaches and these methods had a wide range of benefits, the IPF-based synthetic reconstruction methods were chosen. However, the proposed framework can still be used for other population synthesis techniques to some extent.

Sub question 1b: What steps need to be outlined in the methodology?

From the literature review these steps were identified to be input data, choice for population synthesis technique, choice of control variables, validation, OSM data quality assessment and choice for household allocation. From implementation in the case study, steps such as data harmonization, filtering of houses, choice for household allocation variables and household allocation validation were also added to the methodology.

Sub question 1c: Which statistical technique can be used to allocate households to houses?

Regression analysis, choice modelling, IPF procedure and statistical matching with hot deck procedures were identified as candidate methods. In the case study, it was opted for the regression analysis approach as this gave flexibility, had less data requirements and was intuitive. Regression analysis could also be easily combined with the expert judgement approach used in the case study.

Sub question 1d: How can the generated synthetic population be validated?

In literature, it was found that internal validation can be done by calculating the correlation coefficient between the synthesized population and the aggregated data (marginals). A few methods were also proposed for external validation, such as aggregating the synthesized population and comparing this to an external data set at higher geographical levels or collect and compare real spatial microdata to the synthesized population data.

Sub question 2a: What are the data requirements for the chosen population synthesis technique?

From literature and implementation details from blogs and forums, it was found that most population synthesis techniques require aggregate data for the marginals and disaggregate data (sample data) for the seed. For single-level fitting IPF approaches, the format can be specified as n-control variables, (n-1)-dimensional marginals and n-dimensional seed data.

Sub question 2b: What data in OpenStreetMap can be used for the allocation of households to houses?

The data that can be used were identified as tags that specify the type of building, the number of floors, the number of flats, the address, the height and the living area. For the case study, these were not reported in OSM, so field observations were used to add the values for these tags in OSM. The living area was calculated using the geometry in OSM along with a multiplier for the houses. For the apartment buildings, the number of floors and units were used too. The units were evenly distributed over all floors and the living area was calculated as the surface area of one floor divided by the number of units on one floor.

Sub question 2c: How can the quality of OpenStreetMap data be assessed?

The methods proposed were OSM tools that check for inconsistencies and errors, field research and comparison to other geographical data such as Google Maps. The tagging quality can be checked using TagInfo or the webtool developed by Almendros-Jiménez & Becerra-Terón (2018) that is based on TagInfo. Indicators for completeness and positional accuracy were also suggested but were not used in the case study as the OSM data for the Netherlands is imported data that is already validated and accurate in terms of positional accuracy.

Sub question 2d: How can input data still be derived when confronted with a lack of (micro)data for small areas?

The solution proposed was to use data from higher geographies and scale this down to the size of the study area. This requires an assumption that the distributions of the control variables at higher geographies is the same for small areas. This does introduce uncertainty but there is no other data available so this is the best option to still generate a synthetic population. When this population for the small area is aggregated up to higher geographies, it will still be representative as the data from higher geographies were used as constraints and samples.

Sub question 3a: Which control variables should be used in population synthesis to get a representative population?

After reviewing the literature and the available data, the decision was made to include the household composition, standardised disposable household income and car availability as control variables in the population synthesis in the case study.

Sub question 3b: Which variables from the available OpenStreetMap data for the study area can be used to allocate houses to the generated households?

The OpenStreetMap data was corrected for errors and includes the surface area, number of floors and number of residential units in apartment buildings. The living area was directly used for the household allocation along with the household composition and household income from the synthesized population. The number of floors and number of residential units were indirectly used to calculate the surface area of the residential units and to calculate the total amount of houses and residential units that can be allocated.

For the research the following main question was formulated:

How can population synthesis be carried out for neighbourhoods and to what extent can OpenStreetMap data be used to add a spatial distribution to the synthesized population?

This research proposed a methodology that can be used to synthesize a population, (partially) validate the synthesized population and test the quality of OSM data. Then filter houses, choose a household allocation technique, and validate the spatial distribution of the synthesized population. In doing so, this research has attempted to bridge several existing literature gaps. The main contributions are the establishment of a framework that describes all steps of population synthesis and provides methods for household allocation, implementation details and transparency in the IPF procedure, the implementation of population synthesis for neighbourhoods and exploration of OSM data as a source.

This proposed framework was implemented in a case study, and it was found that the existing available data formed a big constraint. The different institutions from where data could be used also do not harmonize data sets (even data sets from the same institution were not harmonized) to ensure the same definitions, spatial aggregation and time. The data collection is not carried out at the level of neighbourhoods either. This reduces the opportunities for disaggregate modelling at the fine geographical scale proposed in this research. The OSM quality was also not sufficient for household allocation and required corrections and enrichment through data from the BAG. There was no open-source housing data available either for the allocation of households. If applications require more detailed modelling of populations, then it becomes vital for institutions to also start collecting data at this fine scale.

The proof of concept demonstrated in this research, shows that there are opportunities for population synthesis in small areas with OpenStreetMap data. However, the data as of now needs to be corrected and enriched using other data sets. The methodology, even with all uncertainties introduced through a lack of data, is still able to produce a plausible population synthesis with a spatial distribution. From the validation, it can be concluded that there were just minor differences present, and that this technique can be used for detailed population synthesis with spatial distributions in transport models and that this is still a better estimate of reality than randomly allocating households.

6.2 RECOMMENDATIONS

This research led to recommendations for future research, authorities/institutions and the OSM community. The recommendations for future research will be discussed first followed by the recommendations for authorities/institutions and lastly the recommendations for the OSM community. All the recommendations are based on the results and limitations of this research.

6.2.1 RECOMMENDATIONS FOR FUTURE RESEARCH

The recommendations that can improve the methodology and opportunities for applications that stem directly from this research are:

- Perform proper external validation of the population synthesis and household allocation by collecting microdata and marginals for the case study area. This would include collecting data about household attributes and housing situations. Through collection of this data, a ground truth is obtained for the study area.
- Considering lacking data and the inability to collect the aforementioned microdata at times, it is also recommended to implement the entire protocol for the elicitation of expert judgement. The

method recommended for this was IDEA. The confidence intervals reported should then also be used as weights when aggregating the opinions of the experts.

- Conduct a sensitivity analysis for the household allocation by altering attributes of the synthesized population and analysing the effect this has on the household allocation.
- Include more variables in the population synthesis (such as household size, labour market association, number of children in the household, etc.) and household allocation (the type of house, proximity to grocery stores and schools, etc.) and assess how the distribution of the households' changes in the study area.
- To analyse the transferability of the method, it is also recommended to implement it in areas other than residential neighbourhoods (commercial areas, industrial areas, and rural areas) and assess if the methodology is able to cope with these types of areas.
- Utilize the model output (spatially distributed synthesised population) by implementing this in a microsimulation model for transport to assess the value of having such a disaggregated population.
- Further research is also required in refining the household allocation by being able to allocate more than one household to a house for instances where this occurs. Moreover, the usage of more sophisticated allocation rules and techniques that also include stochasticity is recommended.

There are also aspects found in literature that require further research but do not directly stem from implementation of the methodology. These recommendations are:

- More research into the fitting and allocation stage of the IPF procedure that considers spatial units, integer conversion and selection stages. More research is also needed to get a well-established framework for the validation of IPF.
- Thorough comparison of all the population synthesis techniques requires attention too. This should provide insights into robustness, computational effort, transferability, ease of convergence, data requirements, memory requirements and performance of all methods. This makes choosing a population synthesis method easier because most of the advantages and disadvantages will be known.
- An open-source code for multilevel fitting would also be helpful and give more accessibility, implementation details and transparency to population synthesis.
- The usage of IPF, choice models, hot deck procedures and even machine learning algorithms for household allocation. This can showcase the suitability of these methods and makes comparing the methods in terms of performance possible.

6.2.2 RECOMMENDATIONS FOR AUTHORITIES/INSTITUTIONS

The first recommendation is to collect microdata at the level of neighbourhoods. This would enable transport modelling at a fine scale and give more accurate results. Research can thus develop better tools for advice on policies for the authorities. The second recommendation is to harmonize different data sets by defining the variables in the same manner, using the same categories for general variables (such as household composition) and cooperate to increase the usability of these data sets. This creates more consistency and better quality for the data sets.

6.2.3 RECOMMENDATIONS FOR THE OSM COMMUNITY

The first recommendation is to improve the link with the BAG register. The BAG is very detailed and has a lot of information that can be stored in the tags of OSM elements that would immediately make OSM data a more valuable resource. Another recommendation is to include the different residential units as their own entities in residential buildings so the number of flats and their size can easily be retrieved. Another recommendation would be to link the POIs to the apartment buildings that they are placed in.

REFERENCES

Almendros-Jiménez, J. M., & Becerra-Terón, A. (2018). Analyzing the Tagging Quality of the Spanish OpenStreetMap. *ISPRS International Journal of Geo-Information, 7*(8).

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating Synthetic Household Populations: Problem and Approach. *Transportation Research Record: Journal of the Transportation Research Board, 2014*, 85-91.

Arsanjani, J. J., Helbich, M., Bakillah, M., & Loos, L. (2015). The Emergence and Evolution of OpenStreetMap: a Cellular Automata Approach. *International Journal of Digital Earth, 8*(1), 76-90.

Auld, J., & Mohammadian, A. (2010). Efficient methodology for generating synthetic populations with multiple control levels. *Transportation Research Record: Journal of the Transportation Research Board, vol. 2175*(1), 138-147.

Auld, J., Mohammadian, A., & Wies, K. (2009). Population Synthesis with Subregion-level Control Variable Aggregation. *Journal of Transportation Engineering, 135*, 632-639.

Balac, M., & Hörl, S. (2021). Synthetic Population for The State of California Based on Open-data: Examples of San Francisco Bay Area and San Diego County. *The 100th Annual Meeting of the Transportation Research Board*, (pp. 5-29). Washington D.C.

Ballas, D., & Clarke, G. P. (2009). Spatial Microsimulation. (A. S. Fotheringham, & P. A. Rogerson, Eds.) *The Sage Handbook of Spatial Analysis*. doi:https://dx.doi.org/10.4135/9780857020130.n15

Bar-Gera, H., Konduri, K., Sana, B., Ye, X., & Pendyala, R. (2009). Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. *88th Annual Meeting of the Transportation Research Board.* Washington, D.C.

Barthelemy, J., & Toint, P. L. (2013). Synthetic Population Generation without a Sample. *Transportation Science, 47*(2), 266-276.

Beckman, R., Baggerly, K., & McKay, M. (1996). Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice, 30*, 415-429.

Bedeian, A. G., & Mossholder, K. W. (2000). On the Use of the Coefficient of Variation as a Measure of Diversity. *Organizational Research Methods, 3*(3), 285-297.

Bhandari, P. (2020, September 7). *Measures of Variability*. Retrieved from Scribbr: https://www.scribbr.com/statistics/variability/

Boeing, G. (2017). OSMNX: New Methods for Acquiring, Constructing, Analyzing and Visualizing Complex Street Networks. *Computers Environment and Urban Systems, 65*, 126-139. doi:10.1016/j.compenvurbsys.2017.05.004

Bowman, J. L. (2009). Population Synthesizers. *Traffic Engineering & Control, 49*(9), 342-342.

Briem, L., Heilig, M., Klinkhardt, C., & Vortisch, P. (2019). Analyzing OpenStreetMap as Data Source for Travel Demand Models: A Case Study in Karlsruhe. *Transportation Research Procedia, 41*, 104-112. doi://doi.org/10.1016/j.trpro.2019.09.021.

CBS. (2017, March 9). *Huishoudens in bezit van auto of motor; huishoudkenmerken, 2010-2015*. Retrieved from CBS Statline:

https://opendata.cbs.nl/#/CBS/nl/dataset/81845NED/table?searchKeywords=verhuisde%20personen

CBS. (2019, June). *Welvaart in Nederland 2019*. Retrieved from CBS: https://longreads.cbs.nl/welvaartinnederland-2019/bijlagen/

CBS. (2020, December). *Inkomen van huishoudens; huishoudenskenmerken, regio (indeling 2020)*. Retrieved from CBS: https://opendata.cbs.nl/#/CBS/nl/dataset/84866NED/table?dl=56C35

CBS. (2021, October). *Inkomen van huishoudens; inkomensklassen, huishoudenskenmerken*. Retrieved from CBS Statline: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83932NED/table?dl=1CA36

Centraal Bureau voor de Statistiek. (2017). *Huishoudens in bezit van auto of motor; huishoudkenmerken, 2010-2015*. Retrieved from CBS StatLine: https://opendata.cbs.nl/#/CBS/nl/dataset/81845NED/table?searchKeywords=verhuisde%20personen

Centraal Bureau voor de Statistiek. (2019). *Welvaart in Nederland 2019.* Den Haag: Centraal Bureau voor de Statistiek. Retrieved from https://www.cbs.nl/nl-nl/publicatie/2019/27/welvaart-in-nederland-2019

Centraal Bureau voor de Statistiek. (2021). *Gemiddelde huishoudgrootte 2016 - Buurten*. Retrieved from Gemeente Zoetermeer: https://zoetermeer.incijfers.nl/jive/

Centraal Bureau voor de Statistiek. (2021, August 18). *Regionale Kerncijfers Nederland*. Retrieved from Opendata CBS: https://opendata.cbs.nl/#/CBS/nl/dataset/70072ned/table?searchKeywords=inkomen%20en%20vermogen

Chen, N. K., & Cheng, H. L. (2017). House Price to Income Ratio and Fundamentals: Evidence on Long-horizon Forecastability. *Pacific Economic Review, 22*(3), 293-311.

Choupani, A. A., & Mamdoohi, A. R. (2016). Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia, 17*, pp. 223-233.

Deming, W. E., & Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics, 11*(4), 427-444.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., & Spaziani, M. (2014). Statistical Matching of Income and Consumption Expenditures. *International Journal of Economic Sciences, 3*(3), 50-65.

D'Orazio, M. (2017). *Statistical Matching and Imputation of Survey Data with StatMatch.* Rome: Italian National Institute of Statistics.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* John Wiley & Sons.

Ellebjerg, L. (2007). *Noise Control Through Traffic Flow Measures: Effects and Benefits.* Hovedstaden: Danish Road Institute. Retrieved from https://www.vejdirektoratet.dk/api/drupal/sites/default/files/publications/noise_control_through_traffic_flow_measures.pdf

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation Based Population Synthesis. *Transportation Research Part B: Methodological, 58*, 243-263.

Forthomme, D., & Ballis, H. (2021, February 14). *Dirguis/ipfn*. Retrieved from GitHub: https://github.com/Dirguis/ipfn

Fournier, N., Christofa, E., Akkinepally, A. P., & Azevedo, C. L. (2018). An Integration of Population Synthesis Methods for Agent-based Microsimulation. *The Annual Meeting of the Transportation Research Board*, (pp. 13-17). Washington DC.

Frenkel, A., Bendit, E., & Kaplan, S. (2013). Residential location choice of knowledge-workers: The role of amenities, workplace and lifestyle. *Cities, 35*, 33-41. doi:https://doi.org/10.1016/j.cities.2013.06.005

Garguilo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PloS One, 5*(1).

Girres, J. F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS, 14*(4), 435-459.

Goetz, M., & Zipf, A. (2012). Using Crowdsourced Geodata for Agent-Based Indoor Evacuation Simulations. *ISPRS International Journal of Geo-Information, 1*(2), 186-208. doi:10.3390/ijgi1020186

Google Maps. (2021). Retrieved from Google Maps: https://www.google.com/maps/@52.0542282,4.4734422,18.78z

Guo, J., & Bhat, C. (2007). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board, 2014*(1), 92-101.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design, 37*(4), 682-703.

Hanea, A. M., McBride, M. F., Burgman, M. A., Wintle, B. C., Fidler, F., Flander, L., . . . Mascaro, S. (2017). I nvestigate D iscuss E stimate A ggregate for Structured Expert Judgement. *International Journal oof Forecasting, 33*(1), 267-2799. doi:https://doi.org/10.1016/j.ijforecast.2016.02.008

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. H. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of Artificial Societies and Social Simulation, 15*(1).

Hobeika, A. (2005). *TRANSIMS Fundamentals: Chapter 3: Population Synthesizer.* US Department of Transportation.

Huang, Z., & Williamson, P. (2001). *A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata.* Liverpool: University of Liverpool, Department of Geography.

Hunsinger, E. (2008, May). *IPF Description.* Retrieved from GitHub: https://edyhsgr.github.io/IPFDescription/AKDOLWDIPFTHREED.pdf

Jovicic, G. (2001). *Activity Based Travel Demand Modelling - A Literature Study.* Danmarks TransportForskning. Retrieved from

https://resources.nctcog.org/trans/modeling/nextgeneration/ActivityBasedTravelDemandModelingLiteratureStudy.pdf

Kagho, G. O., Balac, M., & Axhausen, K. W. (2020). Agent-based Models in Transport Planning: Current State, Issues and Expectations. *The 9th International Workshop on Agent-based Mobility, Traffic and Transportation Models, Methodologies and Applications. 170*, pp. 726-732. Warsaw: Elsevier.

KC De Entree. (2021). *Welkom op de Entree*. Retrieved from KC De Entree: https://entree.unicoz.nl/welkom-op-de-entree/

Keßler, C., & De Groot, R. T. (2013). Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. *Geographic Information Science at the Heart of Europe*, 21-37.

Klimaatadaptatie Nederland. (n.d.). *Klimaatadaptatie en de omgevingswet*. Retrieved from Klimaatadaptatie Nederland: https://klimaatadaptatienederland.nl/kennisdossiers/omgevingswet/

Kounadi, O. (2009). *Assessing the Quality of OpenStreetMap Data.* London: University College of London: Department of Civil, Environmental and Geomatic Engineering.

Lim, P. P. (2020). *Population Synthesis for Travel Demand Modelling in Australian Capital Cities.* Institute for Social Science Research. Brisbane: University of Queensland.

Lim, P. P., & Gargett, D. (2013). Population Synthesis for Travel Demand Forecasting. *Proceedings of the 36th Australasian Transport Research Forum (ATRF)*, (pp. 2-4). Brisbane.

Liu, J., Ma, X., Zhu, Y., Li, J., He, Z., & Ye, S. (2021). Generating and Visualizing Spatially Disaggregated Synthetic Population Using a Web-Based Geospatial Service. *Sustainability, 13*(3).

Lomax, N., & Norman, P. (2016). Estimating Population Attribute Values in a Table:"Get Me Started in Iterative Proportional Fitting". *The Professional Geographer, 68*(3), 451-461.

Long, Y., & Shen, Z. (2015). Population Spatialization and Synthesis with Open Data. *Geospatial Analysis to Support Urban Planning in Beijing*, 115-131.

Lovelace, R., Ballas, D., & Watson, M. (2014). A spatial Microsimulation Approach for the Analysis of Commuter Patterns: from Individual to Regional Levels. *Journal of Transport Geography, 34*, 282-296. doi:https://doi.org/10.1016/j.jtrangeo.2013.07.008.

Lovelace, R., Birkin, M., Ballas, D., & van Leeuwen, E. (2015). Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique. *Journal of Artificial Societies and Social Simulation, 18*(2). doi:10.18564/jasss.2768

Ma, L., & Srinivasan, S. (2015). Synthetic Population Generation with Multilevel Controls: a Fitness-Based Synthesis Approach and Validations. *Computer-Aided Civil and Infrastructure Engineering, 30*, 135-150.

Marrel, A., Iooss, B., Jullien, M., Laurent, B., & Volkova, E. (2011). Global Sensitivity Analysis for Models with Spatially Dependent Outputs. *Environmentrics, 22*(3), 383-397.

Müller, K., & Axhausen, K. W. (2010). Population Synthesis for Microsimulation: State of the Art. *Arbeitsberichte Verkehrs-und Raumplanung, 638*.

Müller, K., & Axhausen, K. W. (2011). Hierarchical IPF: Generating a Synthetic Population for Switzerland. *51st Congress of the European Regional Science Association: "New Challenges for European Regions and Urban Areas in a Globalised World".* Barcelona: European Regional Science Association (ERSA).

Municipality of Zoetermeer. (2021). *Huishoudens - Zoetermeer*. Retrieved from Gemeente Zoetermeer: https://zoetermeer.buurtmonitor.nl/jive/

Navrud, S., & Bråten, K. G. (2007). Consumers' Preferences for Green and Brown Electricity : a Choice Modelling Approach. *Revue D'économie Politique, 117*(5), 795-811.

Neis, P., Zielstra, D., & Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet, 4*(1), 1-21.

O'Donoghue, C., Morrissey, K., & Lennon, J. (2014). Spatial Microsimulation Modelling: A Review of Applications and Methodological Choices. *International Journal of Microsimulation*.

OpenStreetMap. (2021). Retrieved from OpenStreetMap: https://www.openstreetmap.org/#map=18/52.05437/4.47382

OpenStreetMap Wiki. (2021, June 2). *Key:building*. Retrieved from OpenStreetMap Wiki: https://wiki.openstreetmap.org/wiki/Key:building#Accommodation

OpenStreetMap Wiki. (2021, February 15). *Map Features*. Retrieved from OpenStreetMap Wiki: https://wiki.openstreetmap.org/wiki/Map_features

Ortúzar, J. d., & Willumsen, L. G. (2011). *Modelling Transport* (4th ed.). United Kingdom: John Wiley & Sons, Ltd.

Pritchard, D. R., & Miller, E. J. (2012). Advances in Population Synthesis: Fitting Many Attributes Per Agent and Fitting to Household and Person Margins Simultaneously. *Transportation, 39*, 685-704.

Pukelsheim, F., & Simeone, B. (2009). *On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis.* Augsburg: Universität Augsburg. Retrieved from https://opus.bibliothek.uni-augsburg.de/opus4/1229

Rich, J. (2018). Large-scale Spatial Population Synthesis for Denmark. *European Transport Research Review, 10*(2).

Roick, O., Hagenauer, J., & Zipf, A. (2011). OSMatrix–grid-based analysis and visualization of OpenStreetMap. *The 1st European State of the Map Conference (SOTM-EU).* Vienna.

Rolfe, J., & Bennett, J. W. (1996). *Valuing International Rainforests: a Choice Modelling Approach.*

Rose, A. N., & Nagle, N. N. (2017). Validation of Spatiodemographic Estimates Produced Through Data Fusion of Small Area Census Records and Household Microdat. *Computers, Environment and Urban Systems, 63*, 38-49.

Spetter, R. (2019, November 5). *Is The New Dutch Environment and Planning Act Good For The Commons?* Retrieved from Commons Network: https://www.commonsnetwork.org/news/is-the-new-dutch-environment-and-planning-act-good-for-the-commons/

Strathman, J. G., Dueker, K. J., & Davis, J. S. (1994). Effects of household structure and selected travel characteristics on trip chaining. *Transportation, 21*(1), 23-45.

Su, X., Yan, X., & Tsai, C. (2012). Linear Regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(3), 75-294.

Teekens, J. (2017, February 28). *Environment and Planning Act - Explanatory Memorandum.* Retrieved from Government.nl: https://www.government.nl/topics/spatial-planning-and-infrastructure/documents/reports/2017/02/28/environment-and-planning-act-%E2%80%93-explanatory-memorandum

van Boeijen, A., Daalhuizen, J., & Zijlstra, J. (2020). *Delft Design Guide: Perspectives, Models, Approaches, Methods* (2 ed.). BIS Publishers. Retrieved from https://www.bispublishers.com/delft-design-guide-revised.html

van de Werken, A. (2018, October 29). *Verkeersmodel.* Retrieved from MRDH: https://mrdh.nl/project/verkeersmodel

Van Der Honing, R., & Henckel, J. (2021, March 25). *Verkeersmodel oplossing voor gemeenten voor rapportage geluidemissies Omgevingswet*. Retrieved from Goudappel: https://www.goudappel.nl/projecten/verkeersmodel-oplossing-voor-gemeenten-voor-rapportage-geluidemissies-omgevingswet/

van Duijn, M., & Rouwendal, J. (2012). Analysis of Household Location Behaviour, Local Amenities and House Prices in a Sorting Framework. *Journal of Property Research, 29*(4), 280-297. doi:https://doi.org/10.1080/09599916.2012.717100

van Eck, G., Kouwenhoven, M., & Hofman, F. (2021). 1.1.2 Lessen geleerd uit een backcast met het LMS. *PLATOS 2021 – Modellen in de actualiteit.* PLATOS Colloquium.

Vereniging van Nederlandse Gemeenten (VNG). (2020, July). *Webcollege - De Omgevingswet in vogelvlucht - juli 2020*. Retrieved from VNG: https://vng.nl/publicaties/webcollege-de-omgevingswet-in-vogelvlucht-juli-2020

Voas, D., & Williamson, P. (2000). An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata. *International Journal of Population Geography, 6*(5), 349-366.

Ye, P., Hu, X., Yuan, Y., & Wang, F. Y. (2017). Population synthesis based on joint distribution inference without disaggregate samples. *Journal of Artificial Societies and Social Simulation, 20*(4).

Ye, P., Wang, X., Chen, C., Lin, Y., & Wang, F. (2016). Hybrid Agent Modeling in Population Simulation: Current Approaches and Future Directions. *Journal of Artificial Societies and Social Simulation, 12*(1). doi:10.18564/jasss.2849

Ye, X., Konduri, K., Ram, P., Sana, B., & Waddell, P. (2009). A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. *The 88th Annual Meetings of the Transportation Research Board.*

APPENDIX A: HARRIS PROFILE FOR POPULATION SYNTHESIS METHODS

The criteria for the Harris profiles used for the comparison of population synthesis techniques are described here:

1. *Computation efficiency and memory requirements*: this entails how efficient (in terms of speed and time) the algorithm is and how much storage is required for the algorithm. If the speed is high, the time is low and the storage requirements are low as well, the method will be graded with a high value (+2).
2. *Data requirements*: this compares how much data each method requires and specifically whether there is a sample data set needed. If sample data is necessary along with aggregate data, the data requirements will be high, and the score given will be low (-2).
3. *Convergence*: this describes whether a technique has trouble converging or converges easily and always when conditions (such as no zero-cells) are met. The guarantee of convergence has only been proven for IPF, so this method gets maximum points and the rest have no proof of convergence so are awarded negative points.
4. *Flexibility*: this concerns the scalability and the ease of adding dimensions or constraints to the algorithm. The score awarded will be high if it is easy to scale up or down with the method.
5. *Transferability*: this pertains to the size and detail of the case study areas used in the methods to prove their performance. If it has been used in limited case studies and only specific areas (for example to a big geographical scale) then the score here will be low.
6. *Performance*: this entails the reported errors between the marginals and joint distribution of the real population and the synthesized population. If the deviations are small, the score will be high (+2) meaning that it performs well.

Note: A common criterium that is often used to evaluate models is robustness. However, the robustness is only thoroughly researched for synthetic reconstruction and not for the other methods. This makes the comparison for robustness between the methods uncorroborated (although it is seen as an advantage that synthetic reconstruction techniques are robust).

APPENDIX B: TAGGING QUALITY INDICATORS

The tagging quality indicators defined by Almendros-Jiménez & Becerra-Terón (2018) are:

1. Completeness: this checks whether attributes that are important for navigation are present in the area. These attributes are names of streets and buildings, maximum speeds, direction of traffic, house numbers, name of Points of Interest (POIs) and information such as opening hours and phone numbers.
2. Compliance: TagInfo is used to gain the tagging procedure commonly used for a certain entity in the study region and the adherence to this procedure is then measured and used as a indicator of quality in an area.
3. Consistency: here the contributors' agreement on the tagging procedure of an OSM element is examined and the standard deviation of the number of attributes used to explain an element is measured.
4. Granularity: the average and median of the number of attributes utilized for explaining an entity is computed. If an entity contains a large number of attributes, this implies that the quality is better and there is more detailed information.
5. Richness: here the accuracy with which a specific entity is classified into categories (i.e., values associated to classes) is evaluated. A larger number of categories equates to a more comprehensive classification.
6. Trust: this indicator is based on the "many eyes principle" and links this to the quality of the data. This means that the more versions of entities there are, the better the quality. The contributors' local and global experience are also assessed (Almendros-Jiménez & Becerra-Terón, 2018).

APPENDIX C: VALIDATION OF MEERZICHT OOST NEIGHBORHOOD

Panteia had a data set available of 2016 from the traffic model of the Rotterdam – The Hague metropolitan area (V-MRDH) in which data collection zones were outlined and some data was available for these zones. The data that would be useful for the case study has been filtered out and is presented in the table below

Table 14 Zonal data of study area from V-MRDH

Zone ID	Houses	Residents	Labour force	Student places 0-12 years	Jobs in Retail	Jobs in industry	Other jobs	Total amount of jobs	Average Cars per household
1	0	0	0	0	0	0	1280	1280	0
2	0	0	0	0	0	5	245	250	0
3	180	325	160	0	0	0	0	0	1.1
4	0	0	0	275	0	0	0	0	0
5	0	0	0	0	0	0	580	580	0
6	110	190	95	0	5	0	55	60	0.85
7	510	825	410	0	0	0	0	0	0.85
8	200	310	155	0	0	5	5	10	0.85
9	130	315	155	0	0	0	10	10	0.8
10	185	365	180	335	5	5	5	15	0.85
11	230	550	275	0	0	5	15	25	0.85
12	150	335	165	0	0	0	10	15	0.85

Apart from the data collection zones, the variability between each of the zones is important too as this shows that the study area is heterogenous and the methodology is suited for these types of areas that have a variety of agent/household characteristics and activities. The variability can be measured qualitatively and quantitively. Both will be tested for the study area.

When done qualitatively, a list should be made of typical zones that can be expected in a study area and compare these typical zones to the actual zones that are within this chosen study area. These typical zones that were desired were formulated for this research as follows:

Table 15 Zone types for qualitative assessment of case study

Zone type	Description	Presence in Meerzicht Oost neighborhood
Type 1	Zone with more dependence on a certain transport mode e.g. car.	Zone 3, this was concluded because the average amount of cars per household is the highest for this zone
Type 2	Zone with the primary function of residency.	Zones 3, 6, 7, 8, 9, 10, 11 and 12.
Type 3	Zone with primary function of occupations (offices, retail, industry or other).	Zones 1, 2 and 5 are more work (office, detail), retail and industry oriented. This is evidenced by the number of jobs in these zones and the fact that there are zero residents and zero houses here.
Type 4	Zone with education (school or university) or day care centre.	Based on the quantity of student places for students of 12 years and under, there are two zones, namely zone 4 and zone 10, that have a school or day care centre.

Type 5	Zone with supermarkets, retail shops, restaurants, fuel stations, gym, salons.	From OSM, it is also evident that zone 10 is of type 5 as there is a car wash and fuel station there.
Type 6	Zone with a train station or other major hub.	None
Type 7	Zone with medical centre, clinic, dentist or physiotherapy clinic.	Zone 11 is of type 7 as there is a dentist located there.

Having at least a few of these types of zones, would mean that in terms of quality, the study area possesses enough variability. From the above table, it can be concluded that the study area possesses enough variability for this research.

To quantitatively assess the variability, the following measures could also be implemented:

- Range: for the important parameters find the zone with the lowest and highest value and calculate the difference. This is the easiest measure of variability but gives no information about the distribution.
- Standard deviation: this would give the average amount of variability in the data set for a chosen parameter. The bigger the standard deviation, the more spread out the distribution of the data is.
- Variance: this gives the average of squared deviations from the mean, and it also gives information on the spread of data. The more spread out the data, the larger the variance (Bhandari, 2020).
- Coefficient of variance: in contrast to the other (absolute measures), this one is a relative measure and indicates what the size is of the standard deviation compared to the mean. The measure is given in a percentage. It is used as an indicator of diversity (Bedeian & Mossholder, 2000). The higher the percentage, the bigger the spread. 100% means that the data is relatively highly spread and lower than 100% can be considered low variance.

The data set for the chosen case study area contains the variables houses, residents, labour force, total amount of jobs and number of cars per household. These will be used to check the variability.

Table 16 Quantitative measures for variability

Measure	Houses	Residents	Labour force	Jobs	Cars per household
Mean	141.25	267.92	132.92	187.08	0.58
Range	510	825	410	1280	1.1
Standard deviation	145.14	125.78	384.02	384.02	0.44
Variance	21064.2	64024.81	15820.27	147470.3	0.19
Coefficient of variance	102.75%	94.44%	94.63%	205.27%	74.93%

From the table above, the range can be considered to widely vary when comparing it to the mean. The standard deviation and variance are also high and indicate a lot of diversity. The best measure would be the coefficient of variance, for which most of the indicators (apart from cars per household, albeit this is still on the higher side) indicate a high variance and therefore a good variability.

From the qualitative and quantitative measures, it can be concluded that the study area possesses enough variability for the research and that the data collection zones are not homogenous. This heterogeneity will

require that the houses and households need to be matched and cannot simply be randomly distributed. The study area can therefore be deemed as validated.

APPENDIX D: HOUSEHOLD COMPOSITION DATA FROM ZOETERMEER

Table 17 Household composition data (Municipality of Zoetermeer, 2021)

Year	2015	2016
Other Households	1000	.
One parent households	5400	.
Pair with children	14,800	.
Pair without children	15,100	.
One person households	17,400	18,460
Total households	53,700	.

APPENDIX E: HOUSEHOLD COMPOSITION BY STANDARDIZED DISPOSABLE HOUSEHOLD INCOME

Table 18 Household composition by standardized disposable household income (CBS, 2021)

(2015)	Standardized disposable income (rounded)				
Household composition	<€10,000	€10,000-€20,000	€20,000-€30,000	€30,000-€40,000	>€40,000
One person household	1847.17	6905.58	5208.96	2252.51	1185.79
Pair without children	206.99	2899.3	4965.73	3744.06	3283.92
Pair with children	191.59	2217.22	5138.45	4233.83	3018.91
Other multiple person household	51.16	197.62	311.84	255.08	184.30
One parent household	233.81	2344.2	1820.67	685.17	316.15

APPENDIX F: CAR AVAILABILITY BY HOUSEHOLD COMPOSITION

Table 19 Car Availability percentages for the Netherlands (Centraal Bureau voor de Statistiek, 2017)

% (2015)	Cars in household			
Household composition	0	1	2	3+
One person household	55	42.2	2.5	0.5
Multiple person household	13	52	28.6	6.5
One parent household	36	51.3	10.4	2.5
Pair, total	10	52.3	31	7
Pair with children	6	44	39.1	10.5
Pair without children	13	59.9	23.5	3.8
Other multiple person household	38	34.7	20.2	7.1

APPENDIX G: POTENTIAL SURVEY

Date:…./…./2021

Consent:	☐ I, the participant, understand that I am being asked to participate in a survey that forms part of Shaya Joemmanbaks' thesis work at the Delft University of Technology and consent that my provided answers may be used for this research.
Postal code (4 digits):	
Email (optional):	

1. How many members are in your household?
   ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 or more

2. What is the composition of the household?
   ☐ One person household
   ☐ Couple without children
   ☐ Couple with children
   ☐ One parent household
   ☐ Student household
   ☐ Other

3. How many cars does the household own?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 or more

4. What is the disposable annual household income?
   ☐ ≤ €9.999
   ☐ €10.000 – €19.999
   ☐ €20.000 – €29.999
   ☐ €30.000 – €39.999
   ☐ €40.000 – €49.999
   ☐ ≥ €50.000
   ☐ Income unknown

5. What is the property value of your house (if applicable)?
   ☐ ≤ €100.000
   ☐ €100.000 – €149.999
   ☐ €150.000 – €199.999
   ☐ €200.000 – €249.999
   ☐ €250.000 – €299.999
   ☐ ≥ €300.000
   ☐ unknown

6. What is the rent monthly if it concerns a rental home?
   ☐ ≤ €999
   ☐ €1000 – €1999
   ☐ ≥ €2000

7. What is the living area of the house?
   - ☐ $< 50m^2$
   - ☐ $50m^2 - 74m^2$
   - ☐ $75m^2 - 99m^2$
   - ☐ $\geq 100m^2$

APPENDIX H: THREE-DIMENSIONAL SEED DATA

Table 20 Three-dimensional seed data (unfitted) from OViN 2015 and OViN 2016

Rows	Columns	Slices	
Household Composition (i)	Household Income (j)	Car availability (k)	Amount
1	1	0	0.0001
1	1	1	1
1	1	2	0.0001
1	1	3+	0.0001
1	2	0	22
1	2	1	13
1	2	2	0.0001
1	2	3+	0.0001
1	3	0	9
1	3	1	14
1	3	2	3
1	3	3+	0.0001
1	4	0	5
1	4	1	8
1	4	2	0.0001
1	4	3+	0.0001
1	5	0	0.0001
1	5	1	3
1	5	2	1
1	5	3+	0.0001
2	1	0	0.0001
2	1	1	0.0001
2	1	2	2
2	1	3+	0.0001
2	2	0	2
2	2	1	16
2	2	2	2
2	2	3+	0.0001
2	3	0	5
2	3	1	38
2	3	2	11
2	3	3+	0.0001
2	4	0	1
2	4	1	28
2	4	2	22
2	4	3+	1
2	5	0	0.0001
2	5	1	16

2	5	2	17
2	5	3+	1
3	1	0	1
3	1	1	3
3	1	2	2
3	1	3+	0.0001
3	2	0	4
3	2	1	30
3	2	2	13
3	2	3+	0.0001
3	3	0	2
3	3	1	52
3	3	2	51
3	3	3+	6
3	4	0	0.0001
3	4	1	24
3	4	2	40
3	4	3+	5
3	5	0	0.0001
3	5	1	8
3	5	2	20
3	5	3+	7
4	1	0	1
4	1	1	0.0001
4	1	2	0.0001
4	1	3+	0.0001
4	2	0	1
4	2	1	3
4	2	2	2
4	2	3+	0.0001
4	3	0	0.0001
4	3	1	0.0001
4	3	2	0.0001
4	3	3+	0.0001
4	4	0	0.0001
4	4	1	0.0001
4	4	2	0.0001
4	4	3+	0.0001
4	5	0	0.0001
4	5	1	0.0001
4	5	2	0.0001
4	5	3+	0.0001
5	1	0	2
5	1	1	1
5	1	2	0.0001
5	1	3+	0.0001

5	2	0	8
5	2	1	6
5	2	2	1
5	2	3+	1
5	3	0	1
5	3	1	16
5	3	2	2
5	3	3+	0.0001
5	4	0	0.0001
5	4	1	4
5	4	2	0.0001
5	4	3+	0.0001
5	5	0	0.0001
5	5	1	1
5	5	2	0.0001
5	5	3+	0.0001

APPENDIX I: SCRIPT FOR TWO-DIMENSIONAL IPF PROCEDURES

```
# Build seed data matrix t1 (Household composition X Car availability)

t1 = np.loadtxt('t1_2.txt')

#print(t1)

# Set the marginal totals

xip_1 = np.array([17400, 15100, 14800, 1000, 5400])

xpj_1 = np.array([15873, 24928, 10279, 2620])

aggregates1 = [xip_1, xpj_1]

dimensions1 = [[0], [1]]

IPF = ipfn.ipfn(t1, aggregates1, dimensions1, verbose=0)

t_1 = IPF.iteration()

print(t_1)

# Build next seed data matrix (Household composition X Standardized disposable income) already fitted

t2 = np.loadtxt('t2.txt')

print(t2)

# Build next seed data matrix (Car availability X Standardized disposable income)

t3 = np.loadtxt('t3_2.txt')

# Set the marginal totals

xip_3 = np.array([15873, 24928, 10279, 2620])

xpj_3 = np.array([2530.71777018045, 14563.9244202699, 17445.6545718995, 11170.6380937792,
7989.06514387104])

aggregates3 = [xip_3, xpj_3]

dimensions3 = [[0], [1]]

IPF = ipfn.ipfn(t3, aggregates3, dimensions3, verbose=0)

t_3 = IPF.iteration()

print(t_3)
```

APPENDIX J: THREE-DIMENSIONAL IPF PROCEDURE

```
# make the seed matrix for whole of zoetermeer (derived OViN data from 2016)

#HHSam X HHGest X HHAuto (this is the order: rows, columns, slices)

#m1 = np.zeros((5, 5, 4))

m1r = np.loadtxt('m1_2D_2.txt')

# Note that this returned a 2D array!

#print(m1r.shape)

# However, going back to 3D is easy if we know the

# original shape of the array

m1 = m1r.reshape((5, 5, 4))

print(m1)

#Preserved dimensions along which we sum to get the corresponding aggregates

xipp = xip_1  # Household composition (1x5)

xpjp = xpj_3  # Household Income (1x5)

xppk = xip_3  # Car Availability (1x4)

xijp = t2    # Household composition x household income (5x5)

xpjk = t_3.T  # Household income x Car availability (5x4)

# Make sure the dimensions match

aggregates4 = [xipp, xpjp, xppk, xijp, xpjk]

dimensions4 = [[0], [1], [2], [0, 1], [1, 2]]

IPF = ipfn.ipfn(m1, aggregates4, dimensions4, convergence_rate=0.00001, max_iteration=5000)

m_1 = IPF.iteration()

print(m_1)

totalhouseholds = m_1.sum() # A check if the IPF was performed correctly and if the constraints were programmed right

if round(totalhouseholds) == 53700:

    print('Households generated in IPF are equal to the households in the population')

else:

    print('Incorrect, Check indices, dimensions and aggregates again')
```

```
#downscale the generated population to that of the study area: number of houses are from corrected OSM data set

pop_st = (1122/53700) * m_1

print(pop_st)

darray=pop_st.flatten()  #convert the 3D array to 1D to be able to use a sum-safe rounding method

# print(darray)

pop_round_list = iteround.saferound(darray, 0) #the saferound function returns a list so later this has to be turned into array again

pop_round_array = np.array(pop_round_list) # turn list into an array so the reshape function can be used

pop_round_st = pop_round_array.reshape((5,5,4))

print(pop_round_st) # get the original shape of the output of IPF but now with the rounded values
```

APPENDIX K: GENERATED POPULATION FOR ZOETERMEER AND STUDY AREA

Table 21 Generated population for Zoetermeer and study area (rounded)

Rows	Columns	Slices	Amount (households)	Amount (households)
Household Composition (i)	Household Income (j)	Car availability (k)	Zoetermeer	Study area
1	1	0	1125.29	23.51
1	1	1	651.12	13.60
1	1	2	70.74	1.48
1	1	3+	0.02	0.00
1	2	0	5635.42	117.75
1	2	1	1270.13	26.54
1	2	2	0.01	0.00
1	2	3+	0.01	0.00
1	3	0	2747.67	57.41
1	3	1	2127.34	44.45
1	3	2	333.91	6.98
1	3	3+	0.02	0.00
1	4	0	1328.06	27.75
1	4	1	924.41	19.31
1	4	2	0.01	0.00
1	4	3+	0.01	0.00
1	5	0	0.00	0.00
1	5	1	951.66	19.88
1	5	2	234.07	4.89
1	5	3+	0.05	0.00
2	1	0	0.16	0.00
2	1	1	0.00	0.00
2	1	2	206.83	4.32
2	1	3+	0.00	0.00
2	2	0	664.10	13.88
2	2	1	2026.40	42.34
2	2	2	208.79	4.36
2	2	3+	0.01	0.00
2	3	0	889.16	18.58
2	3	1	3363.40	70.27
2	3	2	713.16	14.90
2	3	3+	0.01	0.00
2	4	0	186.60	3.90
2	4	1	2272.96	47.49
2	4	2	1179.76	24.65
2	4	3+	104.74	2.19

2	5	0	0.00	0.00
2	5	1	1738.80	36.33
2	5	2	1363.21	28.48
2	5	3+	181.91	3.80
3	1	0	170.17	3.56
3	1	1	0.03	0.00
3	1	2	21.39	0.45
3	1	3+	0.00	0.00
3	2	0	454.12	9.49
3	2	1	1299.09	27.14
3	2	2	464.01	9.69
3	2	3+	0.00	0.00
3	3	0	200.15	4.18
3	3	1	2590.15	54.12
3	3	2	1860.77	38.88
3	3	3+	487.38	10.18
3	4	0	0.02	0.00
3	4	1	1786.58	37.33
3	4	2	1967.01	41.10
3	4	3+	480.23	10.03
3	5	0	0.00	0.00
3	5	1	700.55	14.64
3	5	2	1292.31	27.00
3	5	3+	1026.05	21.44
4	1	0	51.16	1.07
4	1	1	0.00	0.00
4	1	2	0.00	0.00
4	1	3+	0.00	0.00
4	2	0	71.26	1.49
4	2	1	81.54	1.70
4	2	2	44.81	0.94
4	2	3+	0.00	0.00
4	3	0	116.62	2.44
4	3	1	58.05	1.21
4	3	2	42.52	0.89
4	3	3+	94.66	1.98
4	4	0	111.70	2.33
4	4	1	48.59	1.02
4	4	2	32.10	0.67
4	4	3+	62.69	1.31
4	5	0	0.04	0.00
4	5	1	54.01	1.13
4	5	2	39.85	0.83
4	5	3+	90.40	1.89
5	1	0	233.80	4.89
5	1	1	0.01	0.00

5	1	2	0.00	0.00
5	1	3+	0.00	0.00
5	2	0	1699.52	35.51
5	2	1	486.17	10.16
5	2	2	66.79	1.40
5	2	3+	91.72	1.92
5	3	0	187.84	3.92
5	3	1	1495.86	31.25
5	3	2	136.96	2.86
5	3	3+	0.02	0.00
5	4	0	0.04	0.00
5	4	1	685.09	14.31
5	4	2	0.01	0.00
5	4	3+	0.02	0.00
5	5	0	0.00	0.00
5	5	1	316.08	6.60
5	5	2	0.02	0.00
5	5	3+	0.05	0.00

APPENDIX L: SCRIPT FOR STUDY AREA DEMARCATION

```
# Make polygon from the coordinates of the study area

coords=([(4.479169, 52.055601), (4.477609, 52.05581), (4.47566, 52.054911), (4.471933, 52.055615), (4.469567, 52.0544), (4.468374, 52.050977), (4.47015, 52.048487), (4.475612, 52.048098),    (4.479169, 52.055601)])

P = Polygon(coords)

G = ox.graph_from_polygon(P, network_type='all')

gdf1 = ox.geometries_from_polygon(P, tags={'building':True})  # Geo dataframe for buildings

# plot all the subcategories of buildings

nodes, edges = ox.graph_to_gdfs(G)  # Geo dataframe for network or else they cannot be plotted in same figure

# checken welke projectie

gdf1.crs

# Plotting both graphs

fig, ax = plt.subplots(figsize=(10,10))

edges.plot(ax=ax,label='Streets', edgecolor='k')

gdf1.plot(ax=ax, label='Buildings', legend=True)

gpd.GeoSeries(P).plot(ax=ax, linewidth=2, edgecolor='blue', facecolor='none')  #plotting the boundary of the study area

gdf1.plot(ax=ax, column='building', legend=True)

plt.tight_layout()  # plot graph
```

APPENDIX M: FINDINGS FROM OSMOSE

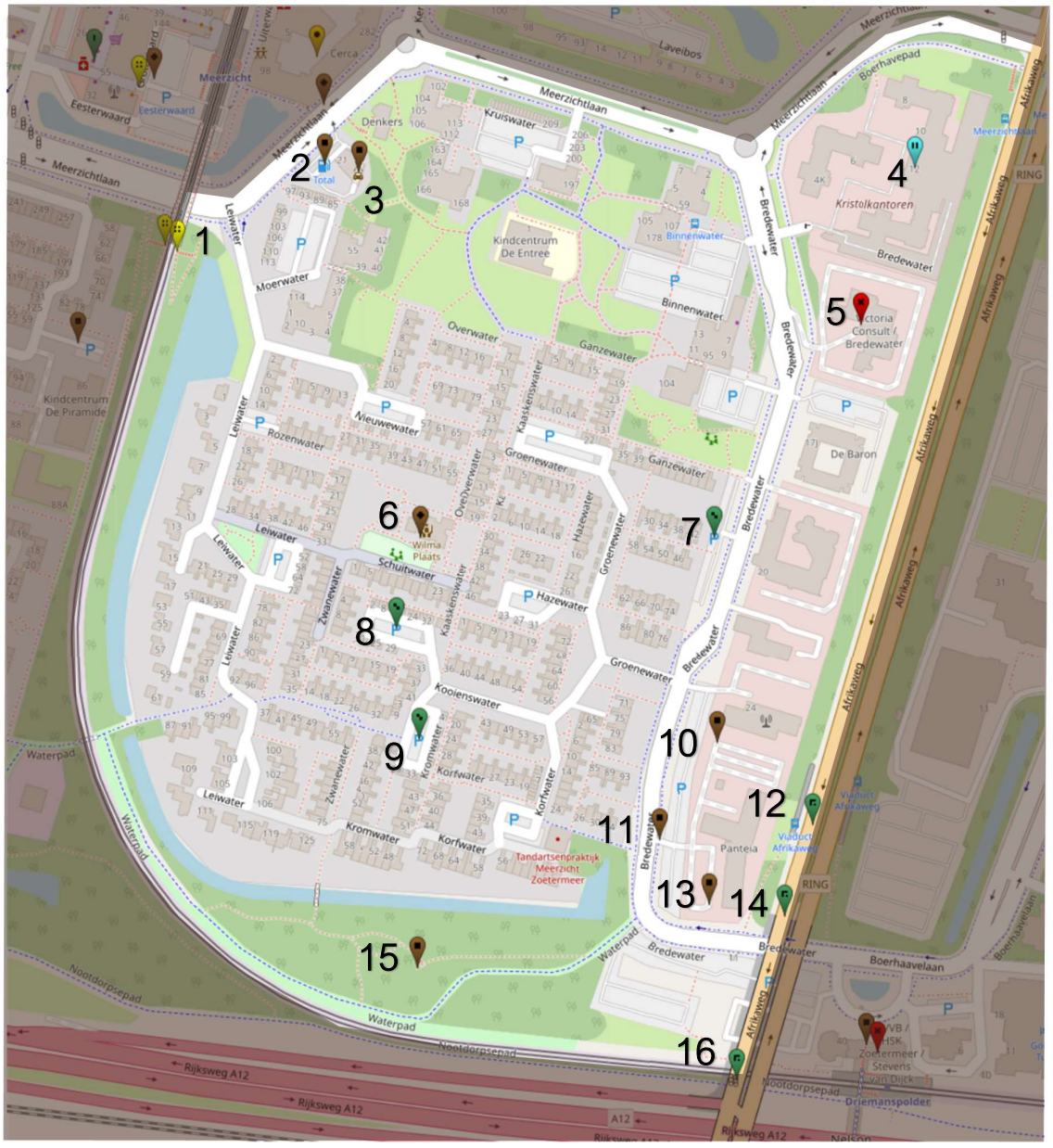

Figure 41 Map with errors

Table 22 Error types

Error no.	Error type	Description	Impact on research
1	Tag for bridge missing	There is a railway specified above the ground but there is no bridge here in a tag.	No
2	Uncommon key value	The fuel station that is mapped has a key named service and value named fuel_station and Osmose is stating that it is an uncommon key value. (Not an actual error)	No

3	Uncommon key value	The carwash that is mapped has a key named service and value named car_wash and Osmose is stating that it is an uncommon key value. (Not an actual error)	No
4	Untagged named object	There is a node that is near the SRK Rechtshulp building and it is not tagged.	No
5	Name tag contains two names	There is a commercial building that is names Victoria Consult and Bredewater. (None of these names are actually correct)	No
6	Missing tag	There seems to be a tag for Wilma Plaats in which the key is amenity and the value is social_facility and OSM wants to remove the social_facility for some reason. (Not an actual error).	No
7	Missing access way to parking	For one of the parking's on the Bredewater there seems to be a missing link to access it according to Osmose.	No
8	Missing access way to parking	For the parking on the Kooienswater there seems to be a missing link to access it according to Osmose.	
9	Missing access way to parking	For the parkings on the Kromwater there seems to be a missing link to access it according to Osmose.	No
10	Uncommon key value	One of the parking's that is mapped on the Bredewater has a key named service and value named parking and Osmose is stating that it is an uncommon key value. (Not an actual error)	No
11	Uncommon key value	One of the parking's that is mapped on the Bredewater has a key named service and value named parking and Osmose is stating that it is an uncommon key value. (Not an actual error)	No
12	Bad turn lanes order	The tag specifying the turns for the lanes on the Afrikaweg seem to have a wrong order.	No
13	Uncommon key value	One of the parking's that is mapped on the Bredewater has a key named service and value named parking and Osmose is stating that it is an uncommon key value. (Not an actual error)	No
14	Bad turn lanes order	The tag specifying the turns for the lanes on the Afrikaweg seem to have a wrong order.	No
15	Key is unspecific	There seems to be a key named barrier with value "yes" that is seen as unspecific and needs to be replaced with a specific value.	No
16	Bad turn lanes order	The tag specifying the turns for the lanes on the Afrikaweg seem to have a wrong order.	No

APPENDIX N: SCRIPT FOR RETRIEVING BUILDING INFO

```
display(gdf)

file_name = 'OSMdataMeerzichtOost1.xlsx'

gdf.to_excel(file_name)  # Export building data to Excel

print('DataFrame is written to Excel File successfully.')
```

APPENDIX O: STORED OSM VARIABLES AND EXPLORATION FOR STUDY AREA

The OSM data set had the variables:

- Element_type
- Osmid
- Nodes
- Addr:city
- Addr:country
- Addr:housenumber
- Addr: postcode
- Addr:street
- Amenity
- Building
- Building:levels
- Ref:bag
- Roof:levels
- Source
- Source:date
- Start_date
- Geometry
- Leisure
- Name
- Operator
- Ways
- Type

In the Excel file, these variables were not specified for all items. All the elements were provided with an element_type, osmid, the nodes that the element had, the geometry, the reference and source which is from BAG specified. From the 763 elements, only one had the city, country, postcode and street specified. 759 had no specified amenity and 4 OSM elements were specified to be schools, a dentist and a social facility. For the building tag, 224 items appear to have the value 'yes', 518 items are specified as 'house', 2 items are specified as schools, 3 items are specified as 'commercial' and 16 were specified as 'garages'. For leisure, two sport centres were specified and the rest of the 761 items were not specified.

The only names specified for the elements were:

- Panteia (Office)
- Denkers (Sport center)
- Wilma Plaats (Youth center)
- Tandartsenpraktijk Meerzicht Zoetermeer (Dentist)
- Victoria Consult/Bredewater (Office)
- Kindcentrum De Entrée (School)
- De Baron (Residential building)

Panteia B.V. mentioned that a lot of the buildings in the zones that are bounded by the Afrikaweg, frequently have other tenants. This concerns commercial buildings. So, in the current OSM data there are still buildings with company names and the companies have since then relocated or closed. This is the case for

Victoria Consult. This could also be the reason why most of the companies that are actually located here are not mapped.

APPENDIX P: COMPARISON OF OSM AND GOOGLE MAPS & FIELD RESEARCH

When comparing OpenStreetMap to Google Maps, there seem to be more companies located in the study area according to Google Maps. Field research was carried out to check whether Google Maps or OpenStreetMap was correct. The field research consisted of observing the buildings and looking for banners, signs and boards that mark the location of the companies and interviewing neighbourhood residents. The companies, offices and schools that were found in Google Maps but not in OSM have been listed in the table below along with an indication of whether its presence was also found during the field observations.

Table 23 Crosscheck of Google Maps findings and field research

Company From Google Maps	Presence found	Remarks
Royal Taxi Service	✓	-
Eim	✗	Eim is the economic institute for middle to small businesses and has now become Panteia B.V. This is thus outdated.
Administratiekantoor De Tichel	✓	Located in the same building as Panteia B.V.
PGH Holland Fysiek Goudhandel	✗	Building has been vacated.
Let's Play Incasso B.V.	✗	Building has been vacated.
Van Dongen Uitvaartzorg & Uitvaartcentrum	✓	-
GABE-IT	✗	Building has been vacated.
GOODZO	✓	GOODZO is the building name and the company located here is UNI3.
BlinktUit	✗	-
Kledingbank Zoetermeer	✓	-
BSA B.V.	✓	-
Bupa Global Travel MyCard	✗	-
Boeddha.org	✗	-
Discotek	✗	-
Your Online Magazine	✗	-
Kromkamp Rijopleiding	✗	-
Stichting Studie der Nadere Reformatie	✗	-
Prive Wellness Bellavita	✓	No boards or signs, but from the front door and the gate that was open to enter, it was seen that this is indeed a wellness resort. There was also confirmation from neighbours.
Admin & Eve	✗	-
Adam Car Repair	✗	-
AVIVO Audio & Video	✗	-
Stichting Ginkgo Biloba	✗	-
Eysbroek Administration	✗	-
Wobo Holding B.V.	✗	-
Prins Clausschool	✗	When searching, this school seems to be located in Rokkeveen and not in Meerzicht Oost. So, this is wrongly mapped in Google Maps. This is illustrated in figure 42.
Prinses Amaliaschool	✗	This was the old name of the school 'De Entrée'.

Stichting Unicoz Onderwijsgroep	✗	This is also wrongly mapped in Google Maps. The school 'De Entrée' is managed by this group and it is not a separate school (KC De Entree, 2021).

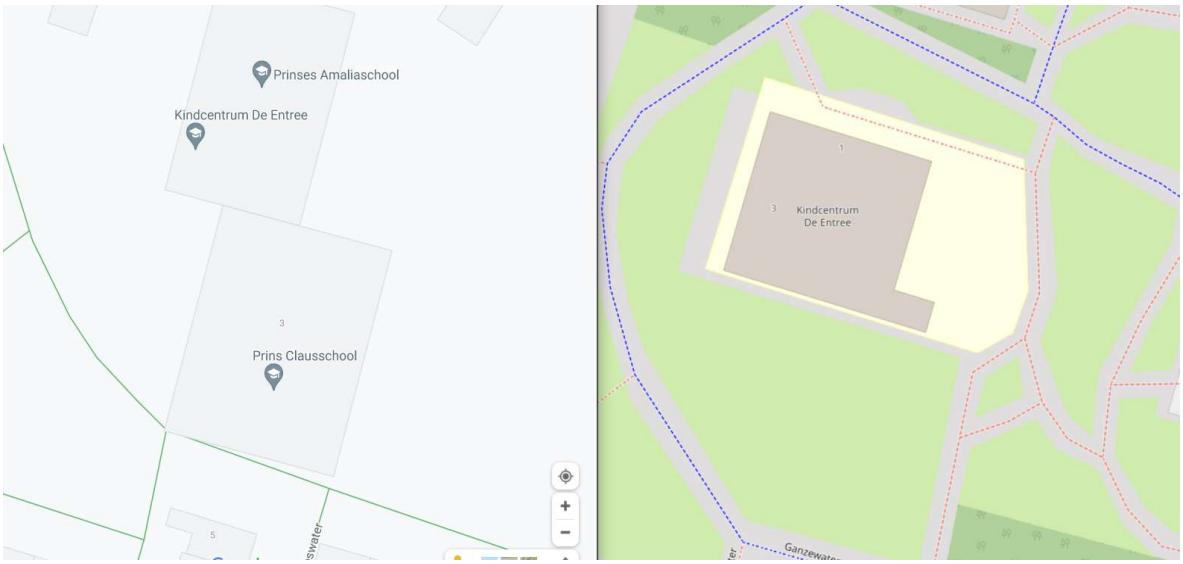

Figure 42 Google Maps (left) (Google Maps, 2021) and OpenStreetMap (right) (OpenStreetMap, 2021) for the schools

APPENDIX Q: EXPLORATION OF BUILDINGS WITH VALUE 'YES'

Apart from the companies, the buildings that were tagged as 'yes' still require more research. To give a proper overview of this, the map of the study area is presented in Figure 43. The buildings with value 'yes' are marked in green. All of the bigger buildings are numbered and one of the small ones to show what is classified as big and small.

Figure 43 Specification of green buildings (value: 'yes')

During the field observation, the following became clear:

- The smaller sized green buildings like number 1 and 2 are either energy houses, garages or storage.
- The elongated green buildings like number 3 are garages.
- Number 4 is a dentist clinic.
- Number 5 is a vacated building that will be redesigned as a residential building in the future.
- Number 6 is a residential building (apartment building) that has 75 house numbers.
- Number 7 is an office building.
- Numbers 8, 9, 11 and 13 are residential buildings (apartment building) that have respectfully 175, 177, 94 and 82 house numbers.

- Number 10 is ladder and equipment storage building.
- Number 12 is a sport centre named Denkers.
- Number 14 is the building of the neighbourhood house named ‘De Ankers’ and youth centre named ‘Wilma Plaats’.

APPENDIX R: CORRECTION OF OSM DATA

```
gdf1.reset_index(inplace=True)  # change index to column

gdf1.loc[gdf1.osmid == 254892100, 'building'] = 'house'  # random house that was classified as yes

gdf1.loc[gdf1.osmid == 3414475, 'building'] = 'apartments'  # De Baron

gdf1.loc[gdf1.osmid == 254885293, 'building'] = 'apartments'  # Dijkwater

gdf1.loc[gdf1.osmid == 254885288, 'building'] = 'apartments'  # Binnenwater

gdf1.loc[gdf1.osmid == 254888486, 'building'] = 'apartments'  # Kruiswater

gdf1.loc[gdf1.osmid == 254885309, 'building'] = 'apartments'  # Moerwater part 1

gdf1.loc[gdf1.osmid == 254885273, 'building'] = 'apartments'  # Moerwater part 2

gdf1['building:flats'] =  np.nan

gdf1.loc[gdf1.osmid == 3414475, 'building:flats'] = 75  # De Baron

gdf1.loc[gdf1.osmid == 254885293, 'building:flats'] = 175  # Dijkwater

gdf1.loc[gdf1.osmid == 254885288, 'building:flats'] = 177  # Binnenwater

gdf1.loc[gdf1.osmid == 254888486, 'building:flats'] = 94  # Kruiswater

gdf1.loc[gdf1.osmid == 254885309, 'building:flats'] = 44  # Moerwater

gdf1.loc[gdf1.osmid == 254885273, 'building:flats'] = 38  # Moerwater

gdf1.loc[gdf1.osmid == 3414475, 'building:levels'] = 4  # setting levels for De Baron

gdf1.loc[gdf1.osmid == 254885293, 'building:levels'] = 11  # setting levels for Dijkwater

gdf1.loc[gdf1.osmid == 254885288, 'building:levels'] = 11  # setting levels for Binnenwater

gdf1.loc[gdf1.osmid == 254888486, 'building:levels'] = 6  # Kruiswater (one part has 4 and one part has 8
so the average is taken because the building is not split in 2 parts like Moerwater )

gdf1.loc[gdf1.osmid == 254885309, 'building:levels'] = 6  # Moerwater part 1

gdf1.loc[gdf1.osmid == 254885273, 'building:levels'] = 4  # Moerwater part 2
```

APPENDIX S: LIVING AREA CALCULATION

```
gdf= gdf1.to_crs('EPSG:28992')  #RD coordinates projection for the Netherlands

gdf["area"] = gdf['geometry'].area

gdf["livingarea"] = gdf['area'] * 1.878     # apartments should be excluded

display(gdf)


df_apartments = gdf.loc[gdf['building'] == 'apartments']

df_houses = gdf.loc[gdf['building'] == 'house'].set_index('osmid')

# adjust living area for apartments based on area, number of flats and levels

df_apartments['livingarea'] = df_apartments['area'] / (df_apartments['building:flats'] / df_apartments['building:levels'])

# display(df_apartments)
```

APPENDIX T: SURVEY FOR HOUSEHOLD ALLOCATION

Questionnaire for expert judgement on relationship between house surface area (living area), household composition and household income

Date:	Click or tap to enter a date.
Name:	Click or tap here to enter text.
Company/department:	Click or tap here to enter text.
Consent:	☐I, the participant, understand that I am being asked to participate in a survey that forms part of Shaya Joemmanbaks' thesis work at the Delft University of Technology and consent that my provided answers may be used for this research.

As part of research for my thesis about generating synthetic households and assigning them to houses in OpenStreetMap, it is important to define the relationship that exists between household characteristics and house characteristics. The household characteristics at my disposal are household composition and household income. The house characteristic available is the surface area of houses (living area).

The house surface area is defined as the living space of a house in square meters. The household composition gives information on the structure of a household. In this case, there are 5 types:

- One person household
- Pair without children
- Pair with children
- Other multiple person household (for example student household)
- One parent household

The household income concerns the standardized disposable income and is defined as the net household income adjusted by factors that correct for differences in household size and composition. It consists of 5 categories, namely:

- < €10,000
- €10,000 - €20,000
- €20,000 - €30,000
- €30,000 - €40,000
- > €40,000

Upon searching for data that might help understanding and estimating the relationship between the household composition and household income on one side and the house surface area on the other side, I was confronted with a lack of data. To still gain insight and make a model for allocation of households, expert judgement can be used. This means that correlations will be decided by experts based on their knowledge and experience. Questions will be asked to infer about the correlation between the variables. Furthermore, it should be noted that this data will be used for the residential neighbourhood Meerzicht Oost in Zoetermeer.

Question 1:

a. Do you think that there is a positive or negative correlation between the living area of a house and the household composition?
   ☐ Positive ☐ Negative
b. Do you think that the correlation between the living area of a house and the household composition is strong, intermediate or weak?

☐ Strong ☐ Intermediate ☐ Weak

c. If you were to give this correlation a numerical value (between 0 and 1), what value would you give it?
   Value
d. How confident (between 0 and 100%) are you about the given answers at questions 1a, 1b and 1c?
   Percentage %

Question 2:

a. Do you think that there is a positive or negative correlation between the living area of a house and the household income?
   ☐ Positive ☐ Negative

b. Do you think that the correlation between the living area of a house and the household income is strong, intermediate, or weak?
   ☐ Strong ☐ Intermediate ☐ Weak
c. If you were to give this correlation a numerical value (between 0 and 1), what value would you give it?
   Value
d. How confident (between 0 and 100%) are you about the given answers at questions 2a, 2b and 2c?
   Percentage %

Question 3:

a. What is the minimum living area each of the following households requires according to you:
   I. A one-person household: area m^2
   II. A couple without kids: area m^2
   III. A couple with kids: area m^2
   IV. Other multiple person household: area m^2
   V. One parent household: area m^2
b. How confident are you about the given surface areas?
   Percentage %

Question 4:

a. What is the living area each of the following households can afford according to you:
   I. Household with income of less than €10,000: area m^2
   II. Household with income between €10,000 and €20,000: area m^2
   III. Household with income between €20,000 and €30,000: area m^2
   IV. Household with income between €30,000 and €40,000: area m^2
   V. Household with income of more than €40,000: area m^2
b. How confident are you about the given surface areas?

   Percentage%

Question 5:

In the table, a cross tabulation is made specifying households based on their household composition and household income. For each of the cells, can you give a living area that these households most likely will have.

	<€10,000	€10,000-€20,000	€20,000-€30,000	€30,000-€40,000	>€40,000
One person household	area m^2	area m^2	area m^2	area m^2	area m^2
Pair without kids	area m^2	area m^2	area m^2	area m^2	area m^2
Pair with kids	area m^2	area m^2	area m^2	area m^2	area m^2
Other multiple person household	area m^2	area m^2	area m^2	area m^2	area m^2
One parent household	area m^2	area m^2	area m^2	area m^2	area m^2

Remarks (if you have any):

Click or tap here to enter text.

APPENDIX U: REGRESSION ANALYSIS AND DIAGNOSTIC PLOTS

```
X_alt = pd.read_csv('X.csv', index_col=[0])

X_alt.drop('OnePerson', axis=1, inplace=True) # drop one column to avoid multicollinearity issues

X_alt.drop('Inc<10k', axis=1, inplace=True) # drop one column to avoid multicollinearity issues

X_alt = sm.add_constant(X_alt)  # Add intercept, standard model does not come with constant

# display(X_alt)

Y = pd.read_csv('Y.csv', index_col=[0])

# display(Y)

model_alt = sm.OLS(Y, X_alt).fit()

model_alt.summary()

# get the standard error of the regression model

SE = model_alt.scale**.5

print(SE)

# have both x and y in the same dataframe

df_altxy = pd.concat([X_alt, Y], axis=1).reset_index(drop=True)

# add predicted values for y (livingarea)

df_altxy['ypred'] = model_alt.params[0] + (model_alt.params[1]*df_altxy['CoupleKids']) +
(model_alt.params[2]*df_altxy['CoupleNokids']) + (model_alt.params[3]*df_altxy['Other']) +
(model_alt.params[4]*df_altxy['OneParent']) + (model_alt.params[5]*df_altxy['Inc10-20k'])+
(model_alt.params[6]*df_altxy['Inc20-30k']) + (model_alt.params[7]*df_altxy['Inc30-40k']) +
(model_alt.params[8]*df_altxy['Inc>40k'])

#calculate residuals (y-ypred)

df_altxy['Residuals'] = df_altxy['LivingArea'] - df_altxy['ypred']

# display(df_altxy)

df_altxy.plot(x='LivingArea', y='Residuals', style='o')

plt.ylabel('Residuals')

plt.title('LivingArea vs. Residuals')

# Can also use: fig = sm.graphics.plot_regress_exog(model_alt,'LivingArea', fig=fig) but this is not working

model_fitted_y = model_alt.fittedvalues

model_residuals = model_alt.resid # alternative way of getting residuals

model_norm_residuals = model_alt.get_influence().resid_studentized_internal
```

```
model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
model_abs_resid = np.abs(model_residuals)
model_leverage = model_alt.get_influence().hat_matrix_diag
model_cooks = model_alt.get_influence().cooks_distance[0]
plot_lm_1 = plt.figure(figsize=(5,5))
plot_lm_1.axes[0] = sns.residplot(model_fitted_y, df_altxy.columns[-3], data=df_altxy,
                          lowess=True,
                          scatter_kws={'alpha': 0.5},
                          line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8})

plot_lm_1.axes[0].set_xlim(0, 130)
plot_lm_1.axes[0].set_title('Standardized residuals vs Fitted')
plot_lm_1.axes[0].set_xlabel('Fitted values')
plot_lm_1.axes[0].set_ylabel('Standardized residuals');
QQ = ProbPlot(model_norm_residuals)
plot_lm_2 = QQ.qqplot(line='45', alpha=0.5, color='#4C72B0', lw=1)
plot_lm_2.axes[0].set_title('Normal Q-Q')
plot_lm_2.axes[0].set_xlabel('Theoretical Quantiles')
plot_lm_2.axes[0].set_ylabel('Standardized Residuals');
```

APPENDIX V: HOUSEHOLD ALLOCATION SCRIPT

```
#getting the regression variables from the population synthesis dataframe

D_OnePerson = df['OnePerson']

D_CoupleNoKids = df['CoupleNoKids']

D_CoupleKids = df['CoupleKids']

D_Other = df['Other']

D_OneParent = df['OneParent']

D_Incless10k = df['Inc<10k']

D_Inc10to20k = df['Inc10-20k']

D_Inc20to30k = df['Inc20-30k']

D_Inc30to40k = df['Inc30-40k']

D_Inc40kplus = df['Inc>40k']

DesiredArea = []

for i in range(len(D_OnePerson)):

    k = model_alt.params[0] + (model_alt.params[2] * D_CoupleNoKids[i]) + (model_alt.params[1] * D_CoupleKids[i]) + (model_alt.params[3] * D_Other[i]) + (model_alt.params[4] * D_OneParent[i]) + (model_alt.params[5] * D_Inc10to20k[i]) + (model_alt.params[6] * D_Inc20to30k[i]) + (model_alt.params[7] * D_Inc30to40k[i]) + (model_alt.params[8] * D_Inc40kplus[i])

    DesiredArea.append(k)

# print(DesiredArea)

#add desired area to the population synthesis dataframe

df_rand = df.copy()  # set of households to be allocated in the randomized allocation

df_val = df.copy()  # Set of households to be allocated for validation

df['DesiredArea'] = DesiredArea

# df  # set of households with desired area for the rule and regression based allocation

df_apartments = gdf.loc[gdf['building'] == 'apartments']

df_houses = gdf.loc[gdf['building'] == 'house'].set_index('osmid')

df_apartmentsdoubled = df_apartments.loc[df_apartments.index.repeat(df_apartments['building:flats'])].copy().reset_index()

df['osmidunit'] = 0   # column to show where the household has been assigned
```

```
df_apartmentsdoubled.sort_values('livingarea', inplace=True)  #sort dataframe based on livingarea from small to large

df_apartmentsdoubled['flatid'] = df_apartmentsdoubled['osmid'].astype(str)+"_"+df_apartmentsdoubled.index.astype(str)

df_apartmentsdoubled.set_index('flatid', inplace=True)

df_randapp = df_apartmentsdoubled.copy()  # set of apartments to be used for the randomized allocation

df_appval = df_apartmentsdoubled.copy()  # set of apartments to be used for the validation

df.index = range(10000, 10000 + len(df))

# df_apartmentsdoubled

df_houses.sort_values('livingarea', inplace=True)  #Sort df houses from small to large as well

# df_houses.set_index('osmid', inplace=True)

df_randhouses = df_houses.reset_index().copy()  # set of flats to be used for the randomized allocation

df_housesval = df_houses.copy()

# df_resunitsrand = pd.concat([df_randapp, df_randhouses], axis=0)

income_groups = ['Inc<10k', 'Inc10-20k', 'Inc20-30k', 'Inc30-40k', 'Inc>40k']

car_groups = ['0cars', '1car']

car_groups2 = ['2cars', '3+cars']

c=0

allocated_flatids = []

allocated_housesids = []

df['flatid'] = 0

df['osmidunit'] = 0

for i in income_groups:
    for j in car_groups:
        df_filt = df[(df[i]==1) & (df[j]==1)].copy()
        print('length filt df', len(df_filt))
        c += len(df_filt)
        for ih, rh in df_filt.iterrows():
            df_app_sel = df_apartmentsdoubled.loc[(rh['DesiredArea'] <= df_apartmentsdoubled['livingarea'])]
            df_app_sel_filt = df_app_sel[~df_app_sel.index.isin(allocated_flatids)]
```

```
            try:
                flatid_sel = df_app_sel_filt['livingarea'].astype(float).idxmin(skipna=True)
            except:
                df_houses_sel = df_houses.loc[(rh['DesiredArea'] <= df_houses['livingarea'])]
                df_houses_sel_filt = df_houses_sel[~df_houses_sel.index.isin(allocated_housesids)]
                try:
                    housesid_sel = df_houses_sel_filt['livingarea'].astype(float).idxmin(skipna=True)
                except:
                    df_houses_filt = df_houses[~df_houses.index.isin(allocated_houseids)]   #if there is no house that satisfies, compromise
                    housesid_sel = df_houses_filt['livingarea'].astype(float).idxmax(skipna=True)  #choose house that is closes to desired area
                allocated_housesids.append(housesid_sel)
                df.loc[df.index==ih,'osmidunit'] = housesid_sel
            else:
                allocated_flatids.append(flatid_sel)
                df.loc[df.index==ih,'flatid'] = flatid_sel
            #print(flatid_sel)
        print('number of allocated flats', len(allocated_flatids))
        print('length house id updated 0-1 cars',len(allocated_housesids))
for i in income_groups:
    for k in car_groups2:
        df_filt = df[(df[i]==1) & (df[k]==1)].copy()
        print('length filt df2', len(df_filt))
        c += len(df_filt)
        for ih, rh in df_filt.iterrows():

            df_app_sel = df_apartmentsdoubled.loc[(rh['DesiredArea'] <= df_apartmentsdoubled['livingarea'])]
            df_app_sel_filt = df_app_sel[~df_app_sel.index.isin(allocated_flatids)]
            try:
```

```
            flatid_sel = df_app_sel_filt['livingarea'].astype(float).idxmin(skipna=True)
        except:
            df_houses_sel = df_houses.loc[(rh['DesiredArea'] <= df_houses['livingarea'])]
            df_houses_sel_filt = df_houses_sel[~df_houses_sel.index.isin(allocated_housesids)]
            try:
                housesid_sel = df_houses_sel_filt['livingarea'].astype(float).idxmin(skipna=True)
            except:
                df_houses_filt = df_houses[~df_houses.index.isin(allocated_houseids)]  #if there is no house that satisfies, compromise
                housesid_sel = df_houses_filt['livingarea'].astype(float).idxmax(skipna=True)  #choose house that is closes to desired area
            allocated_housesids.append(housesid_sel)
            df.loc[df.index==ih,'osmidunit'] = housesid_sel
        else:
            allocated_flatids.append(flatid_sel)
            df.loc[df.index==ih,'flatid'] = flatid_sel
        #print(flatid_sel)
    print('length flat id updated 2-3 cars',len(allocated_flatids))
    print('length house id updated 2-3 cars',len(allocated_housesids))
print('counter allocation', c)
df['osmidunit2'] = np.where(df['flatid']!=0, df['flatid'].str.split('_', n=1, expand=True)[0], df['osmidunit'])

#for random model
# Adding houses and flats in one dataframe
df_resunitsrand = pd.concat([df_randapp, df_randhouses], axis=0).reset_index(drop=True)
# df_resunitsrand
df_rand = df_rand.sample(frac=1, random_state=1).reset_index(drop=True)  # randomized households
df_resunitsrand = df_resunitsrand.sample(frac=1, random_state=1).reset_index(drop=True)  # alle residential units in 1 df randomized
# df_resunitsrand.drop('livingarea', axis=1, inplace=True) # note: run once
```

```
# df_resunitsrand.drop('level_0', axis=1, inplace=True)  # note: run once
df_resunitsrand.drop('index', axis=1, inplace=True)
df_randall = pd.concat([df_resunitsrand, df_rand], axis=1)
```