De potentie van OpenStreetMap data in transport modellen: Een casus in Zoetermeer

Shaya Q. J. Joemmanbaks – TU Delft – shayajoemmanbaks@gmail.com
Ivo Hindriks – Panteia B.V. – i.hindriks@panteia.nl
Jan Kiel – Panteia B.V. – j.kiel@panteia.nl

Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk
13 en 14 oktober 2022, Utrecht

Samenvatting

Publiek beschikbare data is tegenwoordig in ruime hoeveelheden aanwezig. Vanwege privacy-, vertrouwelijkheids- en gegevensverzamelingskwesties zijn deze data vaak alleen op geaggregeerd niveau beschikbaar. Gedetailleerde data (microdata) zijn meestal niet direct beschikbaar. De laatste 10 jaar is er tevens een toename in de kwaliteit en gebruik van geodata zoals OpenStreetMap (OSM). OSM heeft bewezen een waardevolle dataset te zijn. Het levert nuttige informatie voor transportmodellen en transportnetwerken vanwege de gedetailleerde locaties en gegevens van huizen, bedrijven, scholen, infrastructuur en dergelijke. Deze set aan data wordt continu uitgebreid en geactualiseerd en is daarmee een kansrijke bron om in te zetten voor transportmodellen.

Gedesaggregeerde transportmodellen vragen gedetailleerde sociaaleconomische gegevens die niet altijd beschikbaar of gemakkelijk te verkrijgen zijn. Een zogenaamde populatie synthese kan deze beperking deels opvangen. Met een populatie synthese is het mogelijk om een gedetailleerde sociaaleconomische dataset samen te stellen. De statistische procedure die hiervoor het meest wordt gebruikt is 'Iterative Proportional Fitting (IPF)'. Deze methode maakt het mogelijk om met targets op een hoger aggregatieniveau uiteenlopende sociaaleconomische details in te vullen. IPF maakt de weg vrij om te zien of OSM data op gedetailleerd niveau kan worden toegepast en zodoende een methodologie te ontwikkelen voor populatie synthese met een ruimtelijke distributie op het niveau van woningen.

Hoewel het gebruik van OSM data in transportmodellen nog in de kinderschoenen staat, is het gebruik van OSM data onderzocht om het ruimtelijke detail te vergroten door de locaties van woningen nauwkeurig te modelleren en de huishoudens die gegenereerd zijn in de populatie synthese toe te wijzen aan de woningen. De centrale vraag hierbij is:

Hoe kan populatie synthese worden uitgevoerd voor woningen en in hoeverre kan OpenStreetMap data gebruikt worden voor het koppelen van huishoudens aan de woningen in een woonwijk?

Om deze vraag te beantwoorden is een methodologie ontwikkeld. Componenten hiervoor zijn geïdentificeerd in de bestudeerde literatuur. De methodologie is vervolgens toegepast in een casestudy in Zoetermeer en verder verfijnd door toevoeging van andere componenten. Uit de case is gebleken dat OSM data tekortkomingen heeft betreffende thematische gegevens van objecten. Verder heeft de methodologie praktische details inzichtelijk gemaakt zoals het voorbereiden van data, het programmeren van de procedures, strategieën om het tekort aan data op te vangen en toewijzing van de huishoudens. Dit heeft uiteindelijk geleid tot een populatie, waarbij de thuislocaties bekend zijn in OSM en zijn gebaseerd op de relaties tussen huis- en huishoud karakteristieken.

1. Introduction

Recent years have seen an increase in urbanization, spatial restructuring, and population growth. With the rise in computational power of computers, microdata and open-source data, new opportunities arise for microsimulation models. Most of these microsimulation models require a realistic population. Although microdata exists, these cause privacy and confidentiality concerns and are often not available. Because of this, a process called ‘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.

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 that are often not mapped accurately. Therefore, this synthetic 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 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 (a process denoted as household allocation).

For the allocation of houses, detailed data is needed and this can be found potentially in OpenStreetMap data (OSM data). Crowd-sourced OSM data has shown to be a viable data source in literature and was explored in this research to provide the spatial units for the synthesized population. This implies that the houses and residential units were retrieved from OSM data to function as the spatial units by which the synthesized population is distributed.

This paper concentrates on the application of OSM data in microsimulation models. The main question is:

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

This paper contributes to the current body of literature by providing information on:

- Implementation details, transparency and modelling a synthetic population for small areas in detail.
- A methodology that outlines steps for generating a population and adding spatial units through OSM data.
- Empirical evidence on the application through a case study.
- Analysis of the quality of OSM data and its suitability.
- A method for household allocation that combines population synthesis, OSM data and a statistical technique.

This paper will first describe a brief overview of the literature along with the methodology that was derived partially from literature for generating a synthetic population with a spatial distribution. Then the a case study in Zoetermeer is described which outlines practical details on the application of OSM data. Finally, a discussion, future work and recommendations are given.

2. Literature Review and Developed Methodology

A review of existing literature on the topics population synthesis, OSM data and house allocation techniques has led to identifying components that need to be part of the methodology that will be developed. This methodology is needed as there is currently a lack of literature that describes population synthesis and all its steps. Furthermore, the OSM data and household allocation will be added to this methodology as this has also been identified as a research gap. These components included the types of population synthesis, input data, control variables, validation, OSM data quality assessment and choice of method for household allocation.

2.1 Population Synthesis

There are different population synthesis techniques found in the 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. In our study, 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). An overview and comparison of the population synthesis methods can be found in Choupani and Mamdoohi (2016), Pritchard and Miller (2012) and Joemmanbaks (2022).

The 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. The 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. The more control variables chosen, the more the computational complexity increases (Lim, 2020).

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 (Lim, 2020).

Validation of the IPF procedure remains a difficult task. The population synthesis output is often at individual-level and detailed. So, to validate this, 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. For internal validation the aggregate constraint variables are compared with the aggregated results of the population synthesis using the same variables. For the comparison, the Pearson's correlation coefficient can be used and should indicate a perfect correlation between the aggregate constraint variables and the aggregated results of the population synthesis.

For external validation of the IPF procedure, the following options were identified by Lovelace, Ballas & Watson (2014):

- Use real spatial microdata as a comparison to the synthesized population data.
- Use surveys to collect primary data for the study area and then test the synthesized population with this.
- Make comparisons on aggregate levels with the synthesized data and an external data set.
- Sum and accumulate the small area synthesized population to a larger area population and then compare the results with real data from higher geographies.

2.2 OpenStreetMap Data

OpenStreetMap data is collaboratively collected by users and shared on an online community platform (Goetz & Zipf, 2012). This data consists of basic data structures such as nodes, ways and relations. Tags are attached to these structures to represent physical features. Each tag consists of a key and a value (OpenStreetMap Wiki, 2021). The tags are useful for identifying houses/residential units and their characteristics (such as addresses, building type, number of units).

Since OSM data is user generated, there are concerns about the quality and accuracy. 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). There are tools to help assess the quality of OSM such as OSMantic, TagInfo, Osmose, iOSMAnalyzer and OSM Inspector (Almendros-Jiménez & Becerra-Terón, 2018).

It should be noted that 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 quality of The Netherlands was assessed and it was concluded that the buildings were mapped almost completely.

2.3 Household Allocation

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. Iterative Proportional Fitting, choice modelling, regression analysis and statistical matching could all be candidate methods for the household allocation.

2.4 Developed Methodology

After all components were identified from literature and through application in a case study in the neighborhood 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.

The methodology starts with the specification of IPF type, control variables and model choices. Afterwards, the input data is assessed and harmonized. Then the IPF procedure is carried out. Meanwhile, the OSM data is retrieved and assessed. The houses are filtered from this set and their attributes along with household attributes are used in the household allocation. Finally the validation of the entire procedure is done. This framework will be explained through the case study and a more elaborate description can be found in Joemmanbaks (2022).

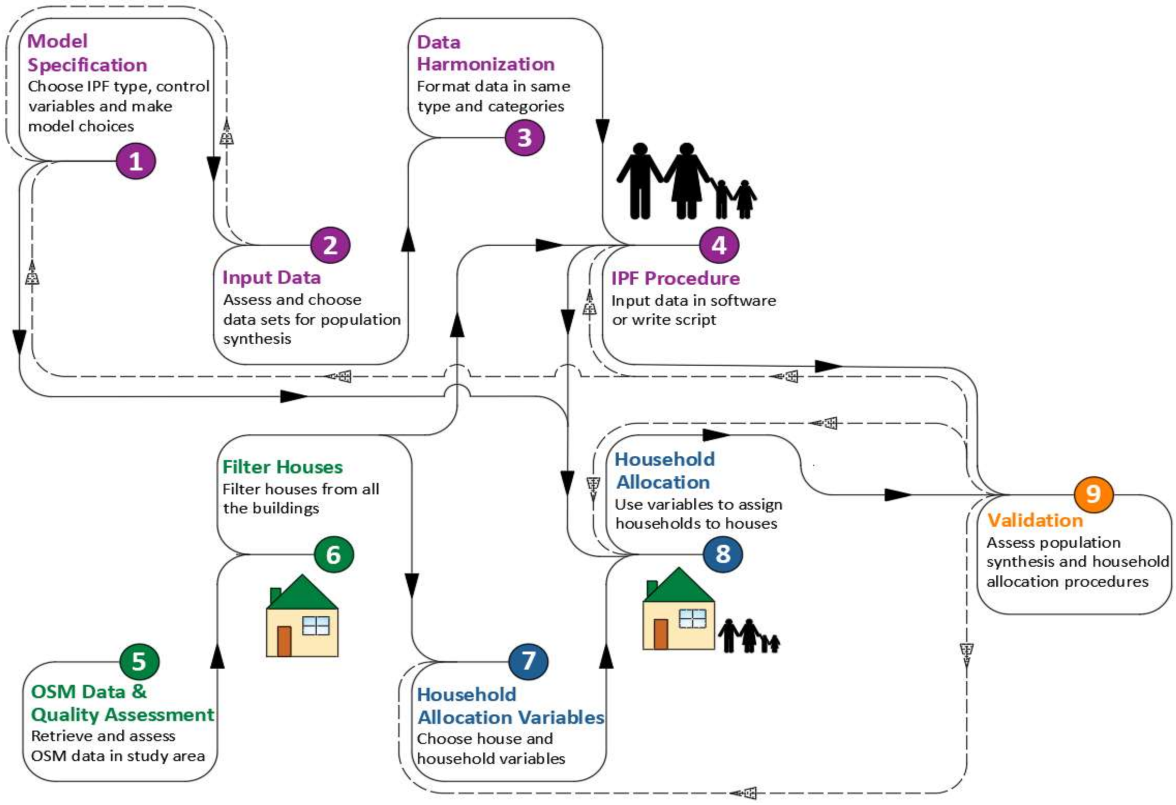

Figure 1 Developed methodology

3. Case Study Zoetermeer

The study area chosen is the size of a neighborhood and is part of Meerzicht Oost in the municipality Zoetermeer in The Netherlands. The area is mainly residential but also has several office buildings, a school and a clinic. Due to limited data availability and the restricted time budget, it was opted for the classic IPF procedure that uses single-level fitting and is zone-by-zone.

3.1 Model Specification

Simplifications and assumptions are introduced in this case study. This is needed to make results attainable within the timeline of this study. The goal is 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).
- Only opensource data sets can be used.

The IPF procedure was done at the level of households. The chosen control variables are the household composition, the standardized disposable household income and the car availability. The data for these control variables were not available at the level of the neighborhood. Therefore, several IPF procedures were used at the level of the municipality Zoetermeer and then scaled down to the size of the study area. This requires the assumption that the distributions for these variables are the same at the level of Zoetermeer and the neighborhood Meerzicht Oost.

3.2 Input Data and Data Harmonization

The next two steps are input data and data harmonization. For the seed data the OViN (Onderzoek Verplaatsingen in Nederland; the research in mobility patterns in the Netherlands) is used. Data from CBS (Central Bureau for Statistics) and the Databank of Zoetermeer were used for the marginals in the IPF procedures.

Since there are three control variables, this concerns a three-dimensional IPF with three-dimensional seed data and two-dimensional marginals. The three dimensions can be seen as rows, columns and slices. The seed data is therefore of the format household composition by household income by car availability. The three two-dimensional marginals needed here are of the format:

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

In the data harmonizing step, the categories for each of the variables is matched to each other. This leads to a reduction in the original categories from the OViN and CBS data sets. This was done for all control variables and led to the following categories:

- Household composition:
    - Type 1: One person household
    - Type 2: Couple without kids
    - Type 3: Couple with kids
    - Type 4: Other multiple person households
    - Type 5: One parent household
- Household income
    - Less than €10,000
    - €10,000 - €20,000
    - €20,000 - €30,000
    - €30,000 - €40,000
    - More than €40,000
- Car availability
    - 0 cars
    - 1 car
    - 2 cars
    - 3+ cars

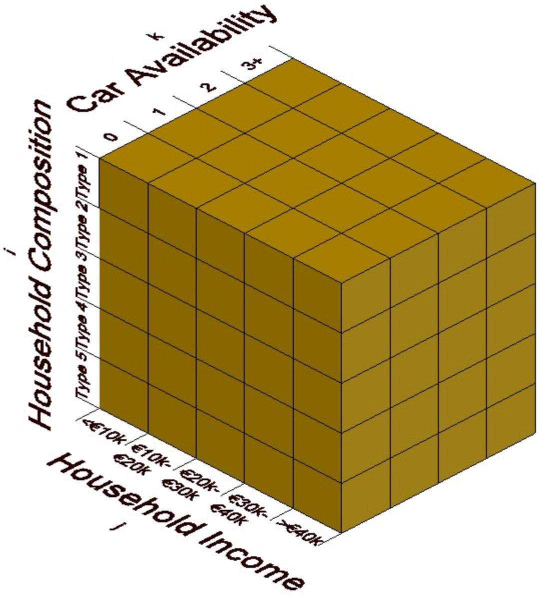

Figure 1 Visualization of seed data adapted from Deming & Stephan (1940)

Visualization of the three-dimensional seed data used in the three-dimensional IPF procedure is illustrated in Figure 2. After the categories were grouped together and new cross tabulations were determined, the cells with the value 0 were replaced with an arbitrarily small value (0.0001) in order to improve convergence. Afterwards, the IPF procedure is carried out.

3.3 OSM Data, Quality Assessment and Filter of Houses

A bounding polygon was used to retrieve the OSM data for the study area. A webtool developed by Almendros-Jiménez & Becerra-Terón (2018) to check the tagging quality was used. From the results, it was concluded that the tagging quality based on the versions did not have many, but this is due to the data imports in OSM. Based on the number of contributors and the sources used, the tagging quality was considered good. Furthermore, Osmose was used to find potential errors that have been marked for the study area. From the errors, it became clear that these would not have an influence on the results for the case study. Field research was also done to further specify buildings

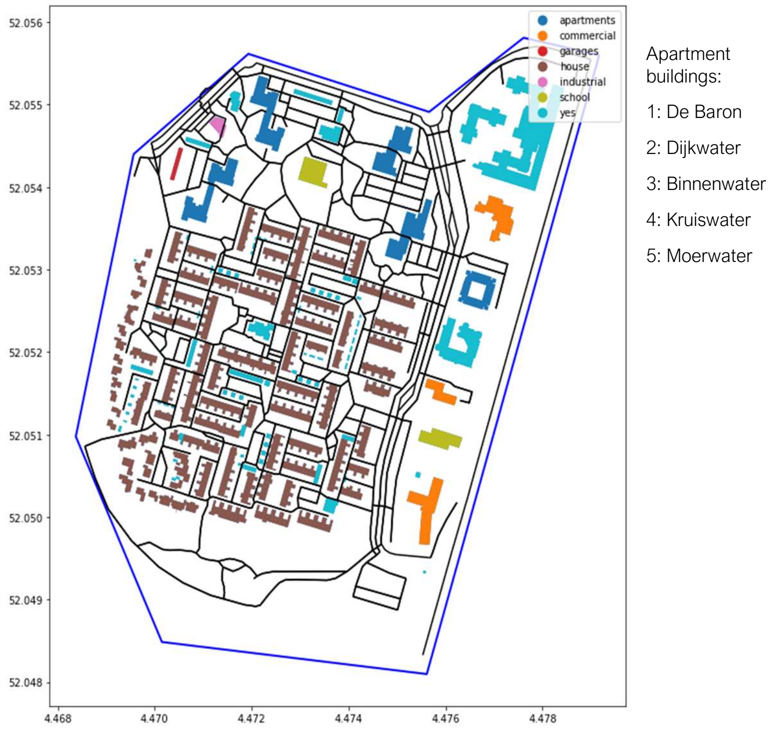

Figure 2 Adjusted graph with buildings

marked with the value 'yes'. Through observations, the value of these buildings was changed to 'apartments'. During the field research the number of apartments in these buildings were also counted and added to OSM along with the number of floors. The results of OSM for this area is shown in Figure 3. After these corrections, the following step was to filter all the houses by writing a script that uses the tags building: house and building: apartments.

3.4 Household Allocation Variables and Method

The available house variables from OSM were surface area and number of flats. The available household variables are the control variables from the population synthesis. However, car availability will not be included as this variable is correlated with the household income.

Due to a lack of open-source data on the relationship between household characteristics and house characteristics, it was opted for the elicitation of expert judgement. Experts were asked to give their estimate of the living area a household would desire based on their household composition and household income. Due to time constraints and COVID-19 only four experts were considered and no consecutive rounds were used for the elicitation. Hanea et al. (2017) outlines structured elicitation of expert judgement approaches. Based on the answers of the experts, a regression model was estimated. The deviating opinions of the experts can be seen in Figure 4. And the regression formula estimated is presented in the following formula:

$$Desired\ Area = 41.63 + \left(28.00 * HHComp_{typ}\ \ \right) + \left(15.75 * HHComp_{type3}\right) + \left(16.12 * HHComp_{typ}\ \ \right) + \left(17.00 * HHComp_{typ}\ \ \right) + (7.75 * HHIncome_2) + (8.4483 * HHIncome_3) + (23.9747 * HHIncome_4) + (43.4483 * HHIncome_5)$$

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 regression analysis was then used to estimate the desired living area for each household in the synthetic population. Then based on several rules, the houses are matched to the households. For the study area, a distinction is made between two categories of housing units. The first category are single dwellings, rowhouses, townhouses and duplexes. And the second category are for apartments and flats. OSM data allows to separate these two types of housing units accurately as well for the study area.

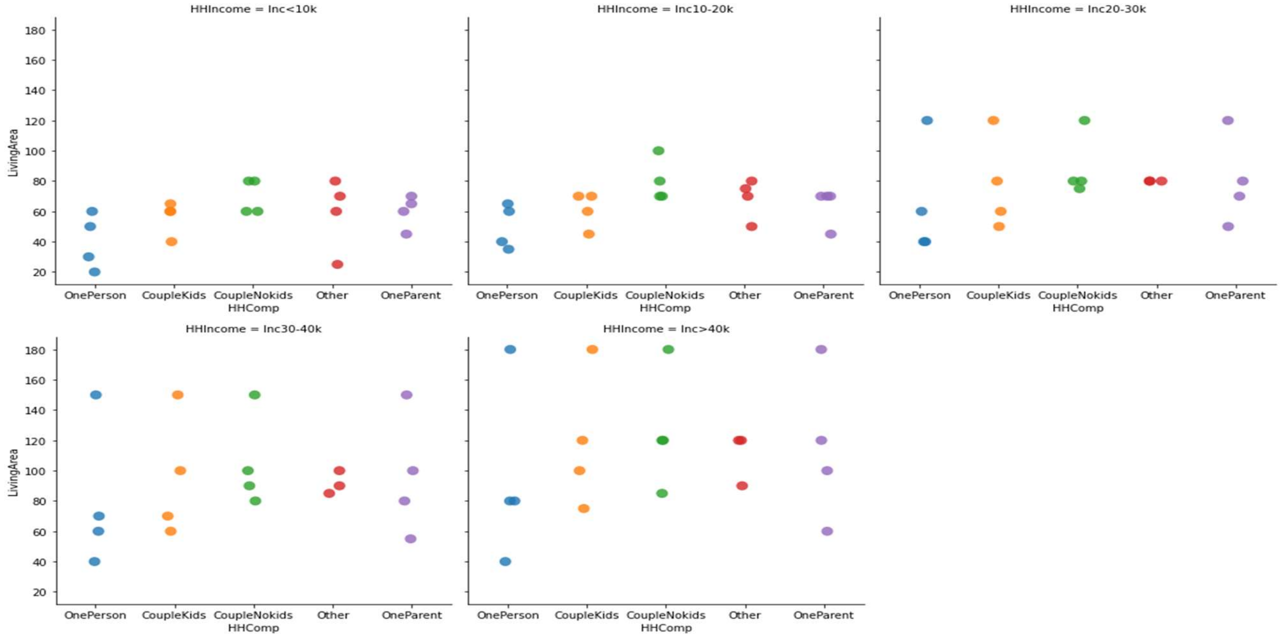

Figure 3 Input data for regression analysis

The following set of rules were used for the household allocation:

1. Most households with a low income should be allocated to the second category of housing units (flats/apartments) because the property valuation of these is lower.
2. Households with 0 or 1 car should be allocated to the second category of housing units first because these housing units usually have no parking or parking for one car in the Netherlands.
3. Households will only be allocated to a house when their desired living area is at least the living area of the house. If there is no candidate house, then there will be a compromise and the household will be placed in the house with the biggest living area out of the available houses.
4. One household can only be assigned to one house, so this allocation does not allow that multiple houses are placed in the same housing unit.

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. The algorithm can be found in Joemmanbaks (2022) along with pseudocode. The results for this model for the household composition is given in Figure 5 and for the income in Figure 6.

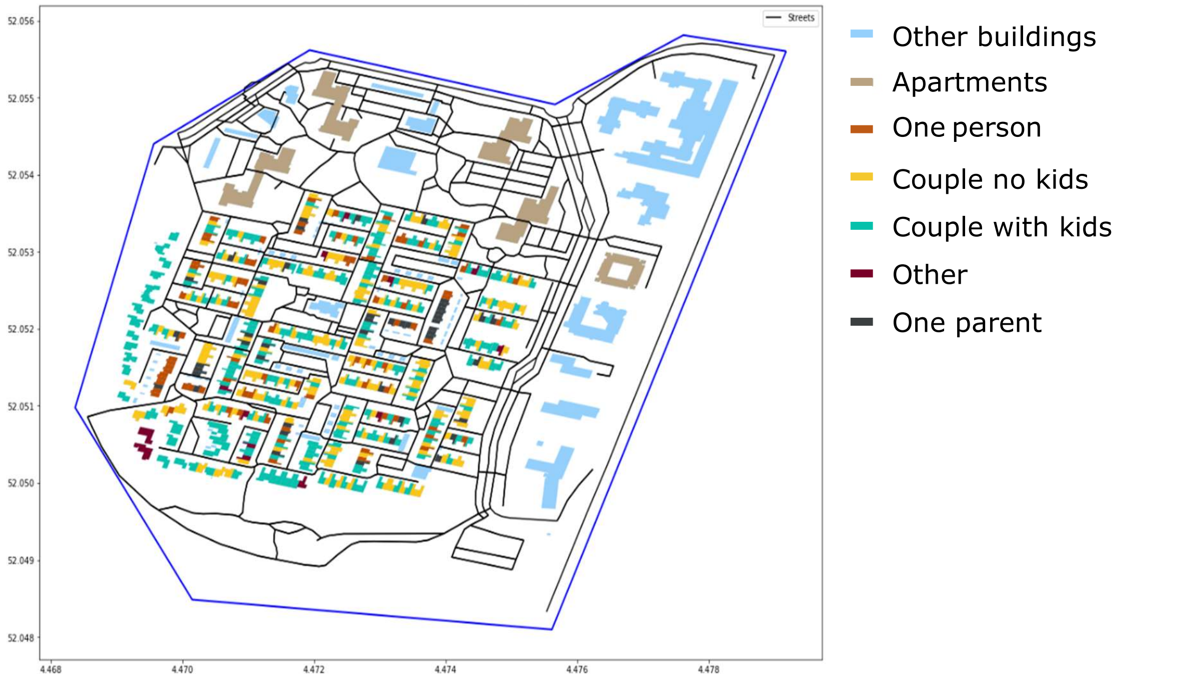

Figure 4 Household allocation based on household composition

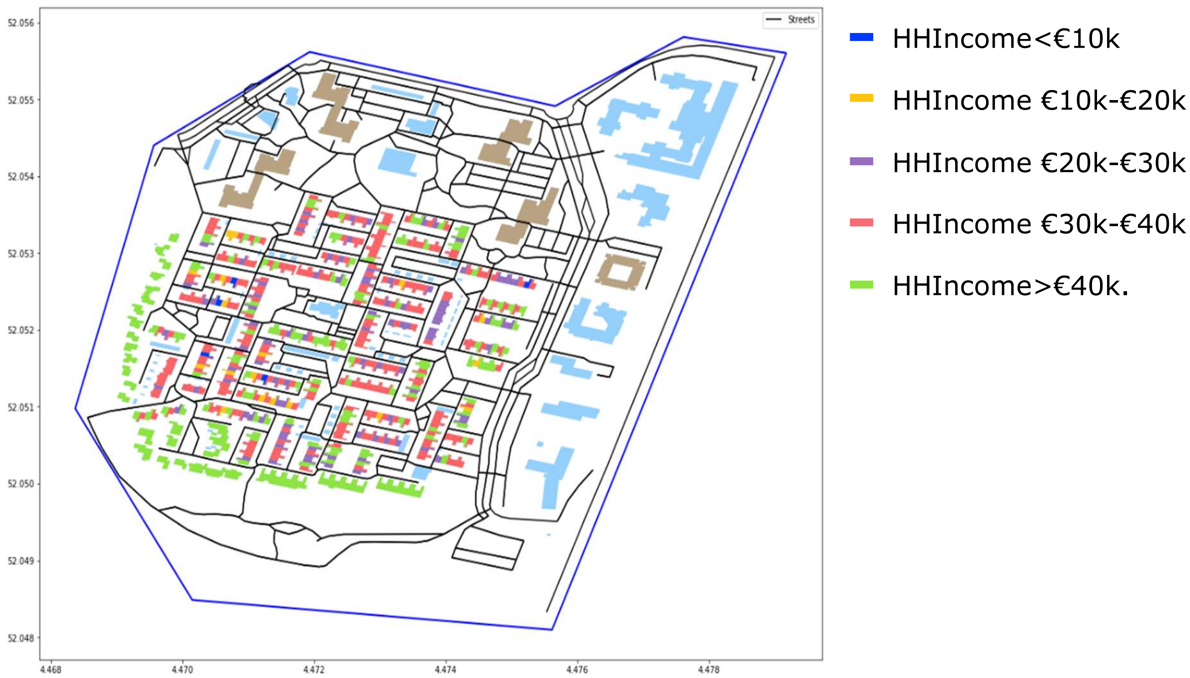

Figure 5 household allocation based on household income

3.5 Validation

Internal validation of the IPF procedure was done through the use of the Pearson correlation coefficient. It was calculated for all three control variables. The result was a correlation of 1, which is in line with expectations. It is hereby concluded that the IPF procedure is internally validated. External validation for the IPF procedure was not possible due to a lack of data.

The external validation of the household allocation was done partially through the housing survey of the Netherlands (WoON). This was not an opensource data set and it also required pre-processing. There were no records included of one person households in this data set and the disposable household income had to be converted to the standardized disposable household income.

In this data set the living area for households having a certain income and composition is given. 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 characteristics as the homogeneous household types. Furthermore, for the homogeneous households that had less than 25 observations (sample size calculated using a Z-score of 1.645 and Margin of Error of 12), a regression analysis was carried out using the WoON as the training data set. For the one person households a dummy value was used along with the coefficients of the regression analysis mentioned in Section 3.4.

When comparing the estimates for desired living area from the expert judgement data set with the housing survey data set for all household types except for one person households, it could be seen that the experts tended to underestimate the living area. The results of using the housing survey data set in the household allocation are presented based on the household composition in Figure 7 and based on the household income in Figure 8.

The outer edge of houses on the left of the study area are mostly allocated to couples with kids in the expert judgement model and in the housing survey model there seem to be more households of couples without kids. The center of the study area looks the same in both models, with just minor differences in the placement of certain household composition types.

For the household income, it appeared 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 center, 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.

Similar trends were also seen for the household allocation when comparing maps that were color-coded based on the car availability. The same analysis was carried out for the apartment/flats and these appeared to be similar as well.

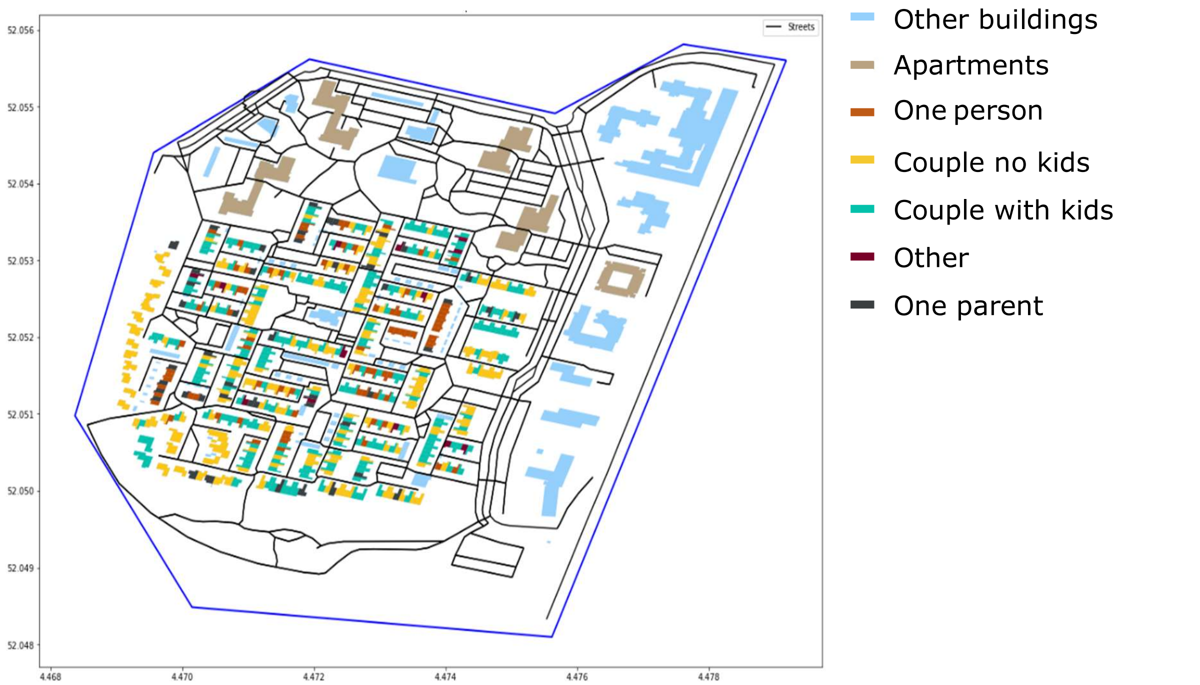

Figure 7 Household allocation for validation based on household composition

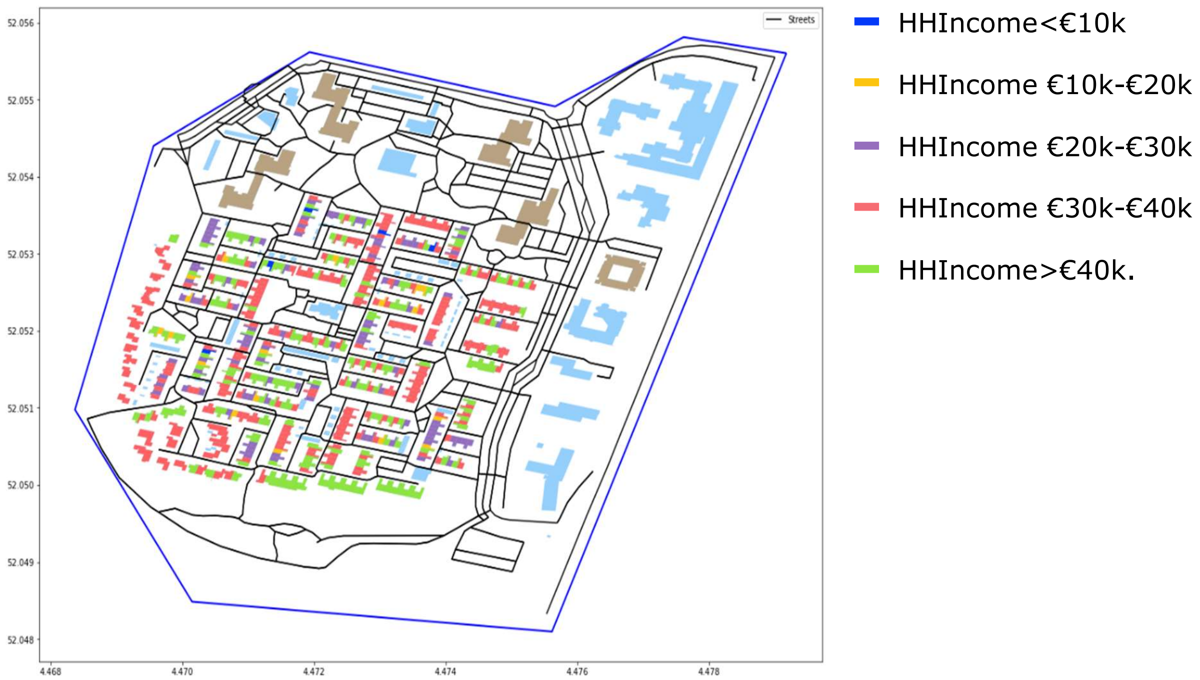

Figure 8 Household allocation for validation based on household income

3.6 Discussion

In the first step in the methodology developed, the population synthesis technique has to be chosen. Due to limited comparisons of different population synthesis techniques, this remains a difficult task. Through the implementation of the single-level IPF in the case study, it can be stated that this method has flexibility in terms of data requirements. Since data for the specific study area was not available, but downscaling data sets from higher geographies could still be done. However, this affects the accuracy of the synthesized population. This research also had the strong limitation of only using opensource data which led to data availability issues. These opensource data sets may not be accurate because they often contain filters, error terms and are rounded. This affects the accuracy as well.

The quality of OSM data was tested in this research, however adjustments had to be made to improve the thematic accuracy. There are still missing tags that are vital when an accurate household allocation is desired. These could be added in by properly combining the Basic Registry of Buildings and Addresses (BAG) with OSM data. Linking the Points of Interest (POI's) to individual buildings would also make the household allocation easier. As of now, these points provide the full addresses and are not linked to the buildings. Thus some information is lost and the flats in the apartment buildings had to be manually separated in individual units using assumptions that do not always hold. This mitigation could still be used in the case study as the study area was mainly residential and had many uniform houses.

Regression analysis was used in the household allocation but it should also be noted that assumptions for homoskedasticity and normality of the residuals did not hold due to the high variability seen in the behavior of households in relation to the house they reside in. There might be opportunities for more sophisticated techniques such as machine learning algorithms to allocate households as these may deal better with the high variability.

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 behavior 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 neighborhoods 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.

3.7 Future Work

The recommendations that can improve the methodology and opportunities for applications 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.
- Conduct a sensitivity analysis for the household allocation by altering attributes of the synthesized population and analyzing the effect this has on the household allocation.
- Include more variables in the population synthesis (such as household size, labor 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 analyze the transferability of the method, it is also recommended to implement it in areas other than residential neighborhoods (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 synthesized 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.

3.8 Conclusions

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 neighborhoods and exploration of OSM data as a source.

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.

4. Acknowledgements

This research has been carried out in the form of a graduate internship at the company Panteia B.V. and was supervised by the company and professors at the department of Transport and Planning of the Delft University of Technology. The findings of the research were then summarized in this paper. The authors of this paper would therefore like to thank Rob van Nes, Adam Pel, Winnie Daamen, Roeland Houkes and Yuko Kawabata for their guidance and support.

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).*

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.

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

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), 267-279. doi:https://doi.org/10.1016/j.ijforecast.2016.02.008

Joemmanbaks, S. Q. J. (2022). *The Potential of Using OpenStreetMap Data in Population Synthesis: A Case Study in Zoetermeer.* Delft University of Technology. Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3Aa8317337-aad2-4a92-9414-9d909195f262?collection=education

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.

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.

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

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

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.