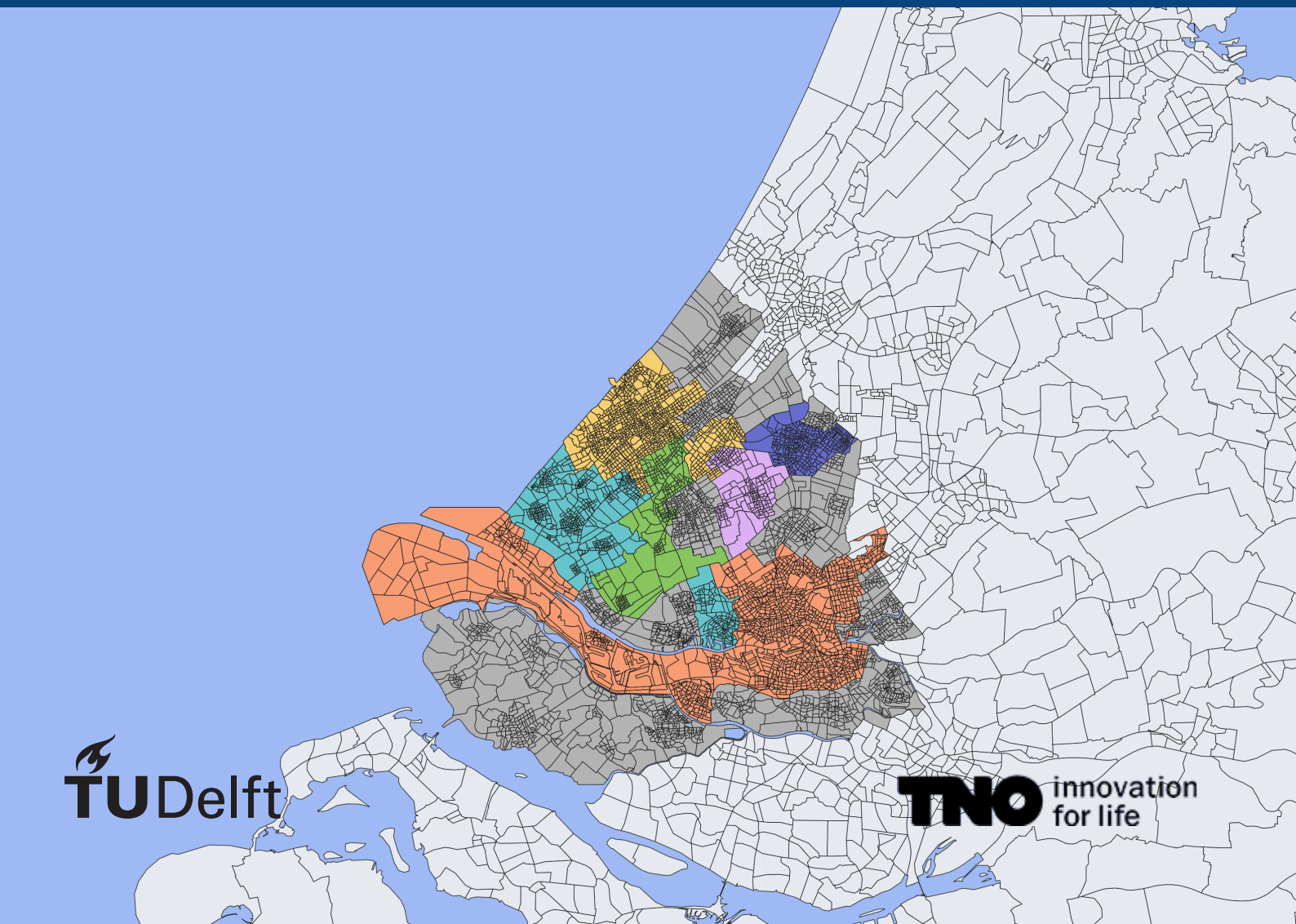
Estimation of an Activity-Based Model to Evaluate Sustainable Mobility Policies in the Netherlands

David Matheus



TU Delft

TNO innovation for life

Estimation of an Activity-Based Model to Evaluate Sustainable Mobility Policies in the Netherlands

by

David Matheus

to obtain the degree of Master of Science in Engineering and Policy Analysis
at the Delft University of Technology,

Student number:	5242223	
Project duration:	February 7, 2022 – August 31, 2022	
Thesis committee:	Dr. E. J. E. Molin,	TU Delft, chair and supervisor
	Dr. Y. Huang,	TU Delft, supervisor
	Ir. H. Zhou,	TNO, external supervisor

TU Delft

TNO innovation for life

Contents

Acknowledgements

It was probably the crisis in my native Venezuela what sparked in me a desire to contribute more to society, and what eventually pushed me to pursue a master's degree in Engineering and Policy Analysis. For me this has been both a personal and academic endeavor. It has been two years of hard work and dedication, facing during much of it the uncertainties and adversities of a global pandemic, but I can say that I am glad I took upon this journey.

This thesis marks the end of my studies at TU Delft, and I want to extend my gratitude to the people that helped me along the way. First, to my supervisors, Eric, Yilin, and Han, for providing me with their guidance to see this project through. To Haiko, for being there to give support and the occasional reassurance. To everyone at the SUMS department at TNO, including my fellow interns, for always lending their assistance when I needed it, the interesting conversations, and for making my internship a pleasurable experience. To my friends and peers, for the many times we took a break together and the crisis counseling when everything seemed to be a mess. To my family, for their encouragement and support, even from an ocean away. To Valeria, because I simply cannot imagine doing this without you; I cherish every minute together, even when I was too stressed to show it.

Now an even longer journey starts, and while contributing to Venezuela seems to be an impossible task, I am excited to see what I might achieve wherever I am.

Abstract

The transition towards greener mobility will play a significant role in decarbonizing the economy. Hence, policy makers need tools that allow them to test alternatives towards that goal. This drive has led to the development of increasingly accurate transport models, with the latest generation being activity-based transport demand models. These models, however, are hard to build, as they are very data-intensive and complex, therefore in this research project a methodology is conceived which attempts to use readily available data from the Dutch travel demand survey (ODiN) and open source software such as ActivitySim, originally developed as a package in Python to make activity-based models in the United States.

While a very time-intensive task, the data was able to be processed for use in ActivitySim. The data was mostly complete, but needed to be complemented with data from the Centraal Bureau voor de Statistiek (2019), and information about members of the household other than the survey respondent, and joint tours, is missing. The data, however, can be processed in a way that can be reapplied in the future, which lowers the barrier to develop a model with Dutch survey data.

Choice is modeled as logit discrete choice models, and the estimation of the parameters required is facilitated by ActivitySim, which has built-in functionality to support it, and with an integrated workflow the model parameters can be estimated with little effort. With this procedure, and choosing workplace and school locations in advance, a good degree of accuracy was achieved, but it was shown that the sampling method used to deal with the very large choice set introduced significant bias to the model output, as observed in the travel distances, which were shorter in the simulated output than in the observations in the survey data.

While ActivitySim has a sampling methodology to deal with large choice sets, an alternative method, Stratified Importance Sampling with activity spaces, is implemented based on the survey data, from where the sample is determined using the travel distances observed and which produces more accurate outputs when compared to the default sampling.

The result of this research is a framework to easily develop and estimate an activity-based transport demand model that is able to provide insights on the travel demand, and especially on how to influence individual choice behavior, which can facilitate the procurement of quality analysis for decision support in the arena of sustainable mobility, hopefully helping accelerate the mobility transition.

It was concluded that using ODiN data and ActivitySim presented as advantages an easy and replicable formulation, and the availability of data that can be used for sustainable mobility policy analysis; yet, this formulation fails to account for household interactions, something that activity-based models often promise to do, and the documentation provided by ActivitySim while extensive is still inadequate in some regards to understand how to process the data.

The resulting model is, however, highly accurate, despite needing some considerations and improvements. The model needs to sample destination choice alternatives, because otherwise its big size would bring the model to a halt, and it does so using a sampling method that is programmed into ActivitySim. This method was shown to introduce bias to the simulation output even if the choice model was properly estimated, and hence an alternative sampling method based on Stratified Importance Sampling was implemented, and the model output greatly improved as a result. Hence, we conclude that it is possible to obtain highly accurate and efficient activity-based models using available data such as ODiN and open source software such as ActivitySim.

It is argued that a formulation like this can be highly beneficial to the evaluation of sustainable mobility policies, as it lowers the barrier to obtain the accurate and detailed outputs that other models cannot produce, and it provides accurate destination choices that will then inform other submodels that are necessary to evaluate sustainability impact, such as mode choice, travel distance and travel time to evaluate emissions.

We continue by discussing the limitations stemming from the available data and the lack of information for other household members and joint travel, the trend-breaking nature of the COVID-19 pandemic and is impact on mobility in further years, and the possible untested bias of the newly proposed sampling method; and giving recommendations on how to effectively use of a model developed with this framework.

Finally, further research is proposed regarding remote work, the improvement of the choice models, and on the sampling method used.

1

Introduction

Sustainability and climate change remain significant societal challenges for Europe and the world. The negative effects of climate change have prompted countries to attempt to tackle it, and global declarations and regional level initiatives such as the Sustainable Development Goals (European Commission, 2022), the Paris Agreement (United Nations, 2015), and the European Green Deal (European Commission, 2019) highlight that decision-makers all over the world consider it a priority.

According to the European Commission (2016), transport represents almost a quarter of Europe's greenhouse gas emissions and is the main cause of air pollution in cities, with road transport to blame for most of it. This implies that transport and mobility play a big role in climate change, one of the most important challenges in the near future to achieve the goals of the European Green Deal (European Commission, 2020). The transition to sustainable mobility will be key to decarbonize the economy and tackle climate change as socio-demographic transitions, increasing urbanization, and a push for sustainability have made it a priority for policy makers (Sharmeen & Meurs, 2019; Stam et al., 2021).

The European Commission (2016) has outlined a number of strategies to limit greenhouse gas emissions originated from mobility, which include integrating digital solution to optimize the transport system, pricing strategies to promote more energy efficient operations, promoting multi-modality, phasing in alternative fuels, and increasing the share of zero-emission vehicles.

To plan these interventions, policy-makers need information about travel demand and transportation system performance such as travel times, modal shares, and congestion (among others) under different policy alternatives to evaluate and compare them (Castiglione et al., 2015). Policy-makers then rely on transport models to obtain these insights (Lah, 2019).

The societal goal of this research is to help in the mobility transition towards sustainability by means of evaluating, improving, and facilitating the development of the models that are used to support relevant decision-making, in hopes that better, accessible models provide more and better quality information to decision-makers in the area and thus they make better decisions. Primarily, high accuracy and

good explainatory value are needed to evaluate these policies. In this sense, this research is scoped towards real applications and implementations, seeking to bridge the gap between existing knowledge and current practice.

To place this research in context it is useful to understand mainstream and emerging modeling techniques. Traditionally, the so-called four-step models are used to inform transport planning, however four-step models are not well equipped to provide the input needed to evaluate most modern sustainable mobility policies as they are heavily aggregated, do not account for interdependence in choices that travelers make, and do not provide much in terms of explaining traveler behavior, so a new generation of models, activity-based models, is presented as an alternative. Activity-based models represent the activities that individual travelers need to perform and derive travel from these, while promising to account for interdependence between chained trips and household members, allowing to model travelers much more realistically and accurately (Castiglione et al., 2015; Ortúzar & Willumsen, 2011). These models offer more explainatory power and detail than its predecessors, as it allows to capture the behavior of travelers and not just the result of their behavior, and they can be integrated with other models to produce outputs beyond the domain of transport, making them more useful for informing policy, including those targeted at achieving sustainable mobility (Castiglione et al., 2015; Tajaddini et al., 2020).

The choice models embedded in activity-based transport models need to be accurate and provide good explainatory value to be useful in informing policy. Additionally, they benefit from being easily specified, its parameters being easily estimated, and expertise in the field already being present. This research aims at providing an accurate, efficient and straightforward way to estimate activity-based models that can be easily be used to analyze a range of sustainable mobility policies. These factors were taken into account to define the research focus of this project, and to further define the research question, a literature review was performed.

Three main aspects are covered by the literature review: an overview of state-of-the-art choice modeling techniques that have emerged in recent times for transport models, the advantages and limitations of these methods, and practical barriers that hinder the deployment of activity-based models. The literature covered spanned scientific papers, where a close match with the search query, a recent publishing date, and number of citations were part of the criteria used to select the relevant literature, to varying degrees depending on the specific subtopic.

From the review it was identified that discrete choice (logit) models are widely used to model choices in activity-based models (Castiglione et al., 2015; Ortúzar & Willumsen, 2011), but alternatives like rule-based models have also been used, with some notable examples like ALBATROSS in the Netherlands (Arentze et al., 2000; Dianat et al., 2020; Hafezi et al., 2019). In addition, an extension to logit models, a multistate supernetwork, has been conceived to conceptualize stronger interdependence between choices (Arentze & Timmermans, 2004a; Fu & Lam, 2014; Liao et al., 2010; K. D. Vo et al., 2020). These models, however, have issues in accuracy and performance when dealing with very large choice sets, as is the case of possible destinations, an issue that has only been addressed in part for the case of unextended logit models by means of sampling methods (Ben-Akiva & Lerman, 1985; Berjisian & Habibian, 2019; Lemp & Kockelman, 2012; McFadden, 1977), possibly indicating that logit models have a better use case. The sampling methods, however, still need to be tested for suitability in activity-based models.

There are also a number of practical challenges that limit the application of activity-based models, large quality data requirements, needed computational power, and issues with the transferability of models to a different context (Tajaddini et al., 2020), even when the cost and time required to develop these models in their most basic form has decreased in recent times (Castiglione et al., 2015).

Nonetheless, activity-based modeling is still is not widespread, in part because of the challenges it poses, like large quality data requirements, needed computational power, and issues with the transferability of models to a different context (Tajaddini et al., 2020), even when the cost and time required to develop these models in their most basic form has decreased in recent times (Castiglione et al., 2015). Computational power is in part addressed in the case of logit models with the use of sampling methods, and Hörl and Balac (2021) suggest that a focus on replicability and adaptability, with the use of available data and open source tools helps address the other practical issues.

In the Netherlands, a couple of activity based models have been developed, of which the most salient is the ALBATROSS model developed by Arentze and Timmermans (2004b) for the Dutch government. However, this model needs extensions for it to be usable in policy analyses, and its verification, forecasting power, and transferability have not been fully established since its development as is the case with rule-based choice modeling formulations such as this (Dianat et al., 2020). Additionally, Knapen et al. (2021) developed a newer model with a focus on multimodality and using logit models, however, it leaves unresolved the issues of dealing with large choice sets and overcoming practical hurdles.

From this review, a gap could be identified in the formulation of activity-based models in the Netherlands. It appears that it would be useful to use a formulation that can prove easy to calibrate, use, extend, and replicate, and to refine the choice model, namely the destination choice models to improve results and run times. However, current formulations, while already benefiting from the use of readily available data like the ODiN travel survey, do not seem to follow the path highlighted by this search and have shortcomings because of it, like high cost of setting up due to model complexity, computation expense, unaccounted transferability and difficult calibration.

The use of open source software like ActivitySim and readily available travel survey data can offer a way to achieve models that are easy to calibrate, use, extend, and replicate in the Netherlands, however, since ActivitySim was not developed for use in the Netherlands some difficulties need to be overcome; the survey data that ActivitySim was designed to use can have a different design from Dutch travel surveys, and sampling methods used for destination choice sets can be unsuitable. An ambition to address these issues produces the main research question: ***What is the accuracy and efficiency of an activity-based model developed using travel survey data and open source software?***

Additionally, this is complemented by the following set of subquestions:

- What are the advantages and limitations of using ODiN data and the ActivitySim software to estimate activity-based model parameters??
- What is the accuracy of destination choice models developed using ODiN data and ActivitySim?
- What is the impact of Stratified Importance Sampling on model accuracy and performance?

In this project, a modeling procedure to estimate an activity-based model for sustainable transport policy

analysis in the Metropolitan Region Rotterdam The Hague in the Netherlands using Dutch travel survey data will be defined using the open source ActivitySim software, and its limitations and performance will be benchmarked. For this, Dutch travel survey data and zonal data will be used to calibrate the relevant choice models. Then, said choice models will be used to perform simulations in the ActivitySim framework for activity-based modeling. The model will then be adapted to used Stratified Importance Sampling to sample destination choice alternatives and the benchmarking repeated. Finally, the model will be evaluated for its usefulness in policy analysis.

This research is part of the graduation requirements for the MSc Engineering and Policy Analysis program at TU Delft, and it was performed in cooperation with the Sustainable Urban Mobility and Safety department at TNO, the Netherlands organization for applied scientific research, who provided support, data, and resources to achieve these results.

Following this introduction, chapter 2 contains an overview on the topic of sustainable mobility policies and transport modeling, chapter 3 has the literature review performed, chapter 4 describes the research approach followed, chapter 5 shows the results, and chapter 6 presents the contributions made to the existing literature, its implications for society, its limitations, and suggestions for further research.

2

Modeling sustainable mobility policies

The mobility transition has been given recent attention, as socio-demographic transitions, increasing urbanization, and a push for sustainability have made it a priority for policy makers (European Commission, 2016; Sharmeen & Meurs, 2019; Stam et al., 2021).

When using models to evaluate the policy alternatives, the model needs to be sensitive to a set of policies of interest, therefore, in the trade-offs between realism and practicality, the modeler needs to prioritize the aspects of the model that affect and reflect the desired set of policies (Arentze & Timmermans, 2004b; Shiftan & Ben-Akiva, 2010). Understanding these policies, their characteristics, their sensitivities, and their performance indicators is key to defining the requirements and suitability for any transport model that aims to incorporate them. Modern policies for sustainable mobility impose that models need to be sensitive to individual choices, are able to provide disaggregated outputs, and are able to provide explanations for the behavior of travelers, conditions that are not met by more traditional trip-based models and that activity-based models are better suited to tackle instead.

In this chapter, a quick overview of modern sustainable mobility policies is given along a description of the requirements they impose, and then traditional and activity-based models and important concepts are briefly explained.

2.1. Modern policies for sustainable mobility

In Europe, and in the Netherlands, a couple of avenues have been conceived to tackle the sustainability issue of transport, namely to more efficiently manage and use the existing infrastructure, and to implement technological advancements that would allow to diminish the environmental impact of mobility (Council of the EU, 2021).

One such policy is the implementation of mobility hubs, or locations where travelers can change their

travel mode within a trip (Knapen et al., 2021). The idea is that by making many transportation modes available on a given location, multimodal travel is facilitated, and greater access is given to modes with less impact on the environment, thus decreasing the environmental footprint of the trip (Aydin et al., 2022).

Another proposed alternative to induce a modal split in favor of more environmentally friendly alternatives is to internalize the cost of using private vehicles, increasing the perceived costs of using them compared to other modes. This cost can take the form of road pricing (usually in the form of tolls), parking pricing, or congestion charges (Gallo & Marinelli, 2020).

The use of emerging modes of transportation have also been considered. In this regard, according to Gallo and Marinelli (2020), aided by mobile technology, vehicle sharing is a concept that appears as an interesting alternative. Car sharing can decrease the need for private vehicle ownership, and its users have shown to be more prone to use sustainable transport modes. On a similar fashion, shared micromobility options offer similar advantages with lower fees and the convenience of a free-floating scheme.

Vehicle electrification is a technological implementation that has also been hailed as an alternative to achieve zero emissions transportation. One of the issues that hampers its adoption is the range anxiety effect, where limited battery range and charging infrastructure prevents people from buying electric vehicles. However, the implementation of such new technology is still clouded in uncertainty, as decision-makers in charge of the infrastructure do not build it fast enough to avoid a situation where they are left with stranded investment, and energy grid load management could become a bigger issue when the number of electric vehicles increase energy demand (Council of the EU, 2021; Gallo & Marinelli, 2020).

Daisy et al. (2020) and Katoshevski-Cavari et al. (2011) also suggest that the efficiency with which the infrastructure is used is heavily dependent in land use. That means that changes in urban planning seemingly independent of transportation, like the adoption of a compact city or a multi-nuclear city as a design alternative, can have an effect on transport and its environmental effects. The implication is that both sides, transport and land use planning, need to be properly incorporated to produce decision-making outcomes that reach sustainability goals.

However, designing policies to provide sustainable mobility is a challenging endeavor; shifting car users towards more environmentally friendly public transport comes at the expense of flexibility and convenience for travelers (Alonso-González et al., 2018; Bruzzone et al., 2020; Lah, 2019), and the implementation of greener private transportation alternatives such as electric vehicles (Patyal et al., 2021) and micromobility options (Esztergár-Kiss & Lopez Lizarraga, 2021) pose significant challenges to achieve and to foresee its effects. For this reason, policy-makers need decision support tools that let them compare the alternatives. These tools allow policy-makers to understand the different impacts of the alternatives, to anticipate uncertainties and the reaction to them, and to plan ahead. Travel models fulfill this purpose by making possible the evaluation of parameters such as accessibility, travel distance, daily trips, congestion, emissions, energy demand, and traveler behavior (Castiglione et al., 2015; Gallo & Marinelli, 2020; Katoshevski-Cavari et al., 2011; Melkonyan et al., 2022; Philip et al., 2013).

2.2. Classic transport modeling

Ortúzar and Willumsen (2011) provide a comprehensive view of transport modeling practices, and they explain the ”classic” transport model, the so-called four step model (figure 2.1), called like that because of the four main stages that are needed to obtain results. First, the input data is used to estimate the total number of trips generated and attracted by each zone in the trip generation stage. Second, these trips are assigned to different origin-destination pairs in a trip matrix on the trip distribution stage. Third, the travel mode distribution is obtained in the modal split stage. Finally, the trips are assigned to the network to evaluate congestion in the assignment stage. Since the congestion obtained on the final step affects travel times, and these can then result different to the assumptions made in the distribution and modal split, these stages are usually run in many iterations to achieve consistency.

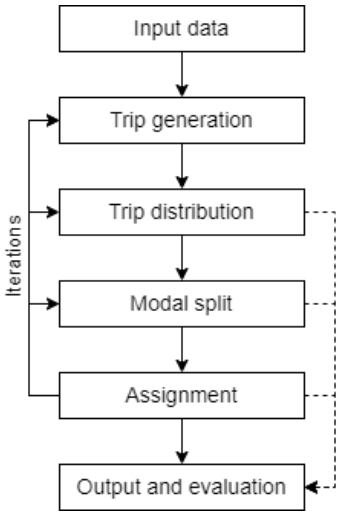


Figure 2.1: Representation of the classic four-stage transport model. Adapted from Ortúzar and Willumsen (2011)

The input for these models include a zoning system spanning the geographical zone to study, the population and land use data for these zones, and the transport network.

The geographical area that the model needs to cover is usually split into smaller units known as Traffic Assignment Zones (TAZs), whose size need to balance the need for detail with the need to aggregate households and premises from the population and zonal data for manageability.

Transport network data is obtained from its representation as a directed graph, and it is assumed that travelers minimize their perceived and anticipated generalized cost function (that includes, for example, travel time and travel costs) to select the links they must traverse between origins and destinations, and then the level of service (cost) of the best route between origin and destination pairs are expressed in what is known as a skim matrix.

The output is a measure of production and attraction of trips per TAZ, a collection of trips per origin and destination TAZ pair on a trip matrix, an aggregate modal split of all the trips performed, and the loads that these trips inflict on the network along with the possible congestion caused.

The precise methods to perform the different stages of the four-step model can vary, but in general the units of analysis are trips, and the outputs are aggregated at zone level. Castiglione et al. (2015)

highlights some of the limitations imposed by this, for example, by stating that assuming that trips are done independently of one another or ignoring the interrelations between individuals in a household reduces sensitivity to policies that require accounting for these dependencies, as decisions made in one trip could affect a following one or someone else in the household. Likewise, the level of aggregation forces the modeler to assume that individuals in a same group (for example, households of the same type) share the same behavior; and the use of average values also distorts the sensitivity of the model and do not allow to provide detailed information on the impact of policies. Also related to the level of aggregation is the impossibility to trace back a trip and explain why it was generated, which limits the explainatory value of these models, especially when it comes to understanding traveler choices, and at the same time, the modeler is forced to make some assumptions about traveler behavior to obtain the output, which might be hard to do when little data about behavior is available.

These limitations contrast with the needs that can be identified from the trends in the policies described in section 2.1. First, many policy alternatives require the implementation and support of new technologies and schemes, for which currently little data is available, especially since they are not widely available or adopted. Second, many of these policies are oriented towards changing the behavior of the population. In line with Diana (2012) and Ortúzar and Willumsen (2011), these reasons imply a necessity to understand and accurately portray how individuals make (possibly interrelated) choices and the factors that drive them to be able to anticipate the impact of such policies on a wider scale, even when little precedent, as is the case, exists.

This calls for models that not only need to be very disaggregated, but that also need to be transparent and provide reasonable explanations for the behavior shown to be useful for policy analysis. The limitations of trip-based models in these regards highlight the need for a more advanced modeling method that allows for effective policy analysis. Activity-based models have been proposed as such an alternative to provide disaggregated and transparent outputs, and to overcome the lack of interaction between choices, by modeling individual travelers and using a tour-based formulation with interrelated choices across the day.

2.3. Activity-Based Models

In this regard, more traditional transport models have given way to more powerful activity-based transport models, whose main difference according to its proponents is that it assumes transport demand to derive from how individuals schedule their activities, who then subsequently need to travel to perform them (Arentze et al., 2005; Arentze & Timmermans, 2004b; Daisy et al., 2020; Hafezi et al., 2019; Ortúzar & Willumsen, 2011).

For each individual, activity-based models determine an activity schedule and then produce chained trips (called tours, see figure 2.2) to perform those activities. This means that activity-based transport models, when compared to their predecessors, are able to represent travel patterns in a disaggregate manner (per person) and thus can explain variations in travel behavior across the population in a much more detailed way, as well as making it easier to incorporate new explanatory variables, interdependence, and sensitivities (Castiglione et al., 2015), making the model more useful and interpretable.

Activity-based transport models are able to capture phenomena such as joint travel scheduling and tour

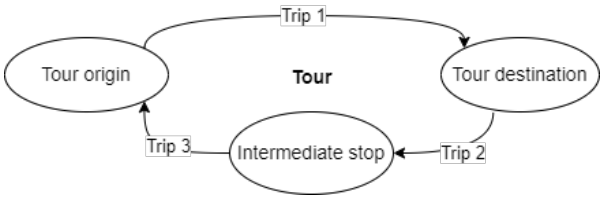


Figure 2.2: Basic representation of a tour, consisting of a series of trips that start and ends on the same location. Tours have a main destination and can have intermediate stops. Adapted from Castiglione et al. (2015)

formation, explicitly include time and space dimensions, model choice behavior explicitly, account for interactions between trips and household members, and provide a flexible simulation framework that can be easily integrated with other domains such as land use and energy use (Castiglione et al., 2015; Katoshevski-Cavari et al., 2011; Knapen et al., 2021). The advantage of this conceptualization is that it is able to show emerging travel behavior from a range of inputs such as socio-economic changes, environmental changes, travel congestion, and emergency situations (Han et al., 2021; Shiftan & Ben-Akiva, 2010), but perhaps more relevant in this case is that it provides a model that is responsive to innovative policies (Arentze et al., 2005; Dianat et al., 2020; Shiftan & Ben-Akiva, 2010).

Activity-based models can be conceptualized as a series of choice models, regardless of the specific choice modeling technique used. They explicitly model the choices that individuals make to plan their activities and the travel derived from them, as opposed to the four-step model, where the trip production and attraction already forces the modeler to make assumptions about activities and generate the trips and their destinations from it (Ortúzar & Willumsen, 2011).

According to Castiglione et al. (2015), activity-based models use a synthetic population, which is a computer-generated population that closely matches the characteristics of the real population as the individuals that will make the choices. Typically, this population will then make individual choices on long term and mobility aspects like school and work location, car ownership, and possession of a transit pass. Then, day activity patterns are defined based on the activities that individuals need to perform, and from there tours are formed on a schedule, with a primary destination and mode; in this regard tours and trips and classified into mandatory and non mandatory, where mandatory tours are those that need to be done and cannot be postponed like work and school, and non mandatory are those that can be rescheduled or skipped altogether. Further, possible stops are added in the tours, breaking them into smaller trips, for which a destination, mode and schedule are also defined. From household interactions, it is also possible to obtain activity patterns and joint travel. Figure 2.3 describes the structure of an activity-based model.

Castiglione et al. (2015) also details that the data necessary as input to develop activity-based models is in principle very similar to the data needed for its predecessor, the four-step model. In fact, activity-based models also use a zoning system, population and household data obtained from travel surveys, land use data, and network data. However, this data is subjected to much more scrutiny, as the level of detail needed is greater and internal consistency needs to be guaranteed. Likewise, the level of detail needed in the time dimension also warrant that the skim matrices obtained from the network data are also obtained for a greater number of time windows, each with their own congestion conditions that reflect on travel times.

The flexibility of the simulation framework used in activity-based models also allows for integration with

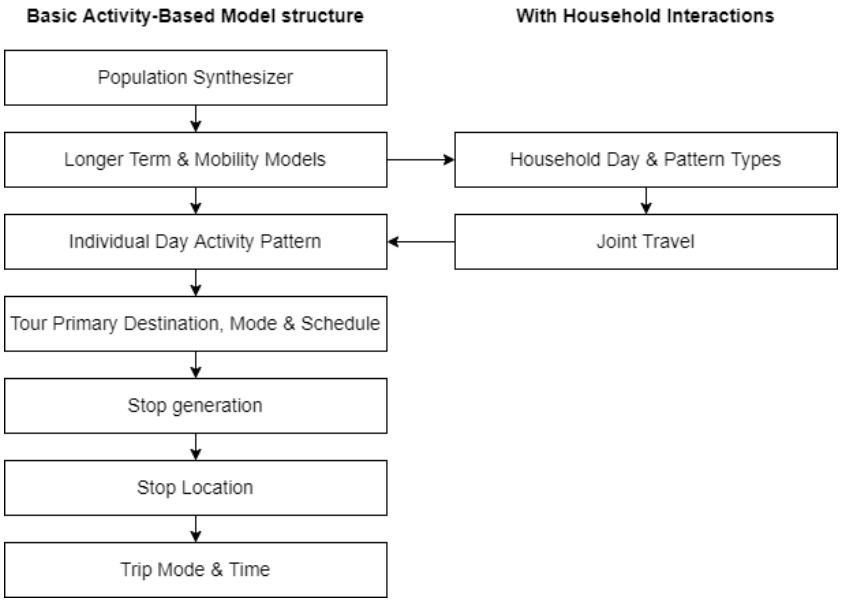


Figure 2.3: Basic structure of an activity-based model. Extracted from Castiglione et al. (2015)

other models to obtain input data, for example, the use of economic models and land use models to obtain forecasted inputs for future years. Likewise, the level of detail provided by activity-based models also allow for the integration of model outputs to other models in different areas of expertise, for example, by relating transport network performance to emissions (Castiglione et al., 2015).

These properties also enable activity-based transport models to answer questions that other models cannot, such as what happens when major disruptions occur (Han et al., 2021), and offering more detail in transport demand management and land use planning and interventions (Arentze et al., 2005; Daisy et al., 2020). Likewise, activity-based models are also able to generate performance indicators beyond the transport domain (Tajaddini et al., 2020). Because of these properties, activity-based transport models seem better suited to provide insights in the implementation of modern policies for sustainable mobility than their predecessors, as illustrated by Knapen et al. (2021). Additionally, the choice models embedded in activity-based transport models can also provide good explainatory value, making them more useful to policy evaluations, especially when traveler behavior is deeply involved as is the case with modern sustainable mobility policies. This then directs this research to focus on activity-based models.

For these reasons, and to understand the state of knowledge in activity-based modeling and identify attempts at deploying them in the Netherlands, it is worth performing a literature review on the topic so that research can be performed to cover the gaps identified.

3

State-of-the-art research and applications of activity-based models

To identify the relevant gaps that need to be addressed in the current knowledge about transport modeling, a literature review was performed. This literature review covers three aspects: the models that are currently available to evaluate (sustainable) mobility policies and their possible shortcomings, the new developments done in activity-based modeling to provide better information, and practical considerations on the deployment of activity-based models.

This chapter describes the method used to search and filter the literature based on the needs of this research, provides an overview of the findings in the literature, and discusses the implications of these findings for the research goals.

3.1. Method

To guide this literature review and make it relevant to the research, its own research question was devised: *What are the knowledge and practical barriers that limit the deployment of activity-based models?*

Additionally, three subquestions helped guide the search:

- *What are the state-of-the-art methods that have been developed in choice modeling for use in activity-based models in recent years?*
- *What are the advantages and limitations of the choice modeling techniques used in activity-based models?*

- *What are the practical barriers to the deployment of activity-based models?*

To find the relevant material, a search strategy was designed, where the search queries used were:

- ”Activity Based Modeling”
- ”Activity Based Modeling” AND Netherlands
- ”Activity Based Modeling” and ”Choice behavior”
- ”Activity Based Modeling” and ”Choice modeling”
- ”Discrete Choice Modeling” AND Transport AND (”Stated preferences” OR Survey OR ”Activity Based Modeling”)

Additionally, ”snowballing” to other papers, or reviewing the citations in the papers found by these search queries was also used as part of the strategy if the content was found interesting and relevant, as well as consulting recommended literature by experts knowledgeable in transport and choice modeling.

These queries were used in a variety of databases including the TU Delft repository, ResearchGate, Elsevier, and Scopus. Since the research question of this project pertains state-of-the-art methodology and new developments, the results were filtered by date, including newer results first, and of those results, the ones that focused on methodological discussions were examined in greater detail. Likewise, articles whose authors are reputed and have many citations were also favored after filtering by dates.

The resulting literature is summarized and discussed in the remainder of this chapter.

3.2. Choice in activity-based transport modeling

There seems to be a general agreement in that the modeler needs to prioritize the aspects of the model that affect and reflect the desired set of policies (Arentze & Timmermans, 2004b; Shiftan & Ben-Akiva, 2010) for it to be useful to inform transport policy, as made explicit in the case of the models developed by Outwater and Charlton (2008), Popuri et al. (2008), and K. Vo et al. (2021). This needed sensitivity places special importance in the way in which choices are modeled for every agent (or individual from the synthetic population) (Diana, 2012; Ortúzar & Willumsen, 2011). This means that choice modeling is at the core of the usability of activity-based transport models in policy analysis, especially the modeling of interdependent choices to make modeled individuals react in a more accurate way to proposed policies (Alonso-González et al., 2018; Diana et al., 2007; Ortúzar & Willumsen, 2011).

Given this relevance of choice modeling in activity-based transport models, it seems reasonable to give special focus to it and spot trends and gaps in the existing knowledge and current practice. It appears that there are three main ways to formalize the different choices of individuals, one is through the use of discrete choice models derived from economic theory (logit models), another is through the use of rule-based computations (Arentze et al., 2000; Shiftan & Ben-Akiva, 2010), and a third one is an emergent technique of discrete choice that makes use of a formalization of choice alternatives called

multistate supernetworks (Arentze & Timmermans, 2004a). This is reflected in the works of Arentze et al. (2005), Daisy et al. (2020), Dianat et al. (2020), Philip et al. (2013), Shiftan and Ben-Akiva (2010), and K. Vo et al. (2021), who use logit discrete choice modeling, the works of Arentze et al. (2000), Arentze and Timmermans (2004b), Eluru and Choudhury (2019), and Han et al. (2021), who instead use rule-based models, and the works of Fu and Lam (2014), Liao (2016), and Liao et al. (2010, 2014), that use multistate supernetworks to model interdependent choices simultaneously.

3.2.1. Discrete choice (logit) models

Logit models are derived from the theory of random utility maximization, where it is assumed that individuals compare choices based on a number of attributes and assigns probabilities to them by maximizing a measure of disutility or satisfaction (Arentze et al., 2000; Castiglione et al., 2015), and thus, allows to predict behavior given the relevant attributes. Discrete choice models, or logit models, incorporate a random component to the (dis)utility of alternatives (and thus having a systematic and a random component) to capture the concept of bounded rationality (Arentze & Timmermans, 2004b; K. Vo et al., 2021), and formulations such as nested logit models allow to capture to some degree the interdependence of the different choices, for example, between mode and destination (Alonso-González et al., 2018; Castiglione et al., 2015; Ortúzar & Willumsen, 2011; Shiftan & Ben-Akiva, 2010), by subordinating one choice to the other. This interdependence is achieved by obtaining the expected utility of the subordinate choice, by means of calculating the ”logsum” of the utilities of the alternatives of the subordinate choice (Castiglione et al., 2015).

This kind of model is widely accepted among modelers, as their output is easy to interpret for modelers and decision-makers, and their estimation is done with well established methods (Castiglione et al., 2015; Ortúzar & Willumsen, 2011).

Logit models, however, can very quickly grow in complexity due to the need to calculate the utilities of every alternative (Shiftan & Ben-Akiva, 2010), especially when logsum calculations are involved (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b), which complicates the implementation of large models with very large choice sets, as could be the case with location or destination choice sets in transport models that cover extensive geographical areas with high granularity.

Also, when modeling interdependent choices in nested logits, the order of main and subordinate choice in the model can be context-dependent. Which choice, mode or destination, should have the higher order in a model has been suggested to vary depending on circumstances such as the country (Castiglione et al., 2015; Kitamura et al., 1997) or the trip purpose and the time frame of the decision (Leite Mariante et al., 2018; López Díaz et al., 2020; Zondag & van Grol, 2021).

Additionally, software implementations that make use of logit models such as ActivitySim, add additional rules to the formulation, such as long term destinations (school and work) being determined in advance for every agent and remain fixed during the runs. This also means that for these trip purposes, destination is necessarily chosen before mode of transportation (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b).

This implies that despite the interest in logit models, this formulation still requires testing to ensure that the structure used is appropriate for the context in which is used, in addition to recalibrating the model

and dealing with the complexity of large choice sets.

3.2.2. Rule-based approaches

As an alternative to discrete choice modeling, ruled-based approaches try to use algorithms to mimic the choice behavior of individuals. In this regards the options are very varied, Arentze et al. (2000) uses decision trees, Arentze and Timmermans (2004b) derives the rules from the choice heuristics of consumers by also adding constraints, Dianat et al. (2020) proposes the use of a skeleton schedule that deals first with mandatory activities that happen on a more strict basis to then restrict the scheduling of more optional activities, and Hafezi et al. (2019) clusters the population to then predict attributes of individuals based on them.

There are already developed algorithms available to induct decision trees from data sets, and the decision tree formulation can be computationally efficient and its results very easy to explain to decision-makers while also allowing for interdependent choices. However the accuracy of decision trees can suffer greatly with the lack of data, where it is possible that the sample is too small and does not represent the heterogeneity of the real choice giving way to very biased outcomes. It is also worth noting that the decision tree formulation also loses accuracy with very large choice sets, such as is often the case with location alternatives (Arentze et al., 2000; Arentze & Timmermans, 2004b).

Rule-based formulations, including but not limited to decision trees, also have the issue of being deterministic in their output which can be problematic if a high confidence value is not guaranteed (Arentze et al., 2000). Additionally, while the decision rules induced by these formulations can be very easy to interpret and explain to decision-makers, the implementation of rule-based formulations can be less widely understood by modelers and thus considered less straightforward to implement than their discrete choice counterparts, as modelers would require new knowledge to do so (Castiglione et al., 2015; Shiftan & Ben-Akiva, 2010).

3.2.3. Multistate supernetwork extension to discrete choice models

An extension to discrete choice (logit) modeling, a multistate supernetwork, has received some attention in recent times. A multistate supernetwork, or ”network of networks”, is a representation of the transport system that includes a copy of the physical transport network for each possible activity–vehicle state during a tour, where the link costs are the generalized utilities from using the network, and a path searching algorithm finds solutions that minimize costs and that transitions between the different states or networks (Arentze & Timmermans, 2004a) to obtain the systematic utility to be used in a logit model. Further development of this formulation include more efficient supernetwork creation (Liao et al., 2010), stochasticity to account for uncertainty and bounded rationality (Fu & Lam, 2014), the inclusion of more choice dimensions (Liao et al., 2014), and accounting for household interactions (K. D. Vo et al., 2020).

The multistate supernetwork formulation has the advantage of formalizing interdependent choices with ease, and in a much less simplified way than logit models without this extension. However, despite attempts to achieve more efficient network representations, the inclusion of more choice alternatives or choice dimensions makes the supernetwork explode in size, which imposes practical limitations on

the size of the problems that can be handled with this approach (Arentze & Timmermans, 2004a; Liao et al., 2014). Additionally, the estimation of utility function parameters for multistate supernetworks can be more complex and new estimation methods and algorithms might be necessary (Arentze & Timmermans, 2004a).

In practice, however the logit and multistate supernetwork approaches have incorporated elements of rule-based approaches to overcome some of its trade-offs, and increase efficiency and accuracy (Arentze & Timmermans, 2007; Association of Metropolitan Planning Organizations Research Foundation, n.d.-b; Dianat et al., 2020; Hafezi et al., 2019; Shiftan & Ben-Akiva, 2010; K. D. Vo et al., 2020).

An issue that seems to be common to all methods described is difficulties in dealing with large choice sets. Logit models require the calculation of more utilities which is computationally expensive, rule-based models decrease in accuracy, and the multistate supernetwork approach, even if it can model interdependent choices with greater ease, explodes in size also imposing greater memory requirements despite attempts to make them more efficient. This is particularly relevant as large choice sets can materialize when modeling destination choice, a step that can be important for sustainable mobility not only when evaluating its direct impact with policies that change the land use, but also because mode choice (and policies that tackle it) can be dependent on the destination and thus also relies on efficient and accurate outputs in this regard.

Despite the additional attention required in the appropriate model structure, the interest in dealing with this issue in the much more established logit models has given way to the development of sampling methods that aim to reduce the number of utilities that need to be calculated in large choice sets without compromising accuracy, which warrant exploring.

3.2.4. Sampling large choice sets in choice models

Even when logit models are the most established in literature and practice, they still present shortcomings in the form of long run times in complex and disaggregate models, that stem from the need to calculate utilities from every possible alternative in large choice sets. When facing large choice sets, simulations that make use of logit models need to calculate the utilities of every alternative requiring a lot of calculations and slowing down the simulation. Additionally, in cases like destination choice models it is also questionable whether or not considering every alternative is realistic, as it is unlikely that a person knows every possible alternative that exists or has the ability to consider them all. Hence, some attempts have been done at sampling choices, with varying degrees of impact to result accuracy.

The relatively naive Simple Random Sampling method has been tested for multinomial logit models (McFadden, 1977; Pozsgay & Bhat, 2001) and it achieves consistent parameter estimates at the expense of predictive capability, and significant reductions of accuracy occur with smaller sample sizes. Nerella and Bhat (2004) suggest that to achieve acceptable results, a minimum of one eighth, and a recommended one fourth of the data must be sampled, which can still be computationally challenging with very large datasets.

Ben-Akiva and Lerman (1985) proposed Importance Sampling, which also yields consistent estimates, but it relies on the modeler's intuition to define a measure of importance or probability of utility maximization to sample, which introduces bias and difficulties in systematically implementing it (Lemp &

Kockelman, 2012). To counter these issues, Lemp and Kockelman (2012) propose Strategic Sampling, which initializes with Simple Random Sampling and then performs a series of iteration to achieve accuracy in a systematic way that would also reduce bias; however, it poses a hard constrain on the independence of relevant alternatives property, meaning that the method cannot be used for modeling dependent choices, a quality that could be desirable in activity-based transport models, while the need for various iterations also makes it more impractical.

Stratified Importance Sampling has also been proposed as a method to sample destination choices decreasing bias by defining strata of destinations based on a fixed distance from any origin point which is informed from observations in the data (Berjisian & Habibian, 2019), and extensions to this method have been developed using the concept of activity spaces to increase efficiency and account for the spatial awareness of the modeled agent (Leite Mariante et al., 2018; Tsoleridis et al., 2022). The latter extension also has the advantage of being conceived for destination choice dependent on mode choice, although it has been tested only on a small dataset or for a simplified pedestrian case, which means its performance in larger scale models remains unproven.

ActivitySim, a software package for the development of activity-based travel models (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b), has a form of presampling built into it, which first estimates utilities on a higher level of aggregation to define likely alternatives. However, this relies on the definition of a zonal system that can be aggregated to a higher order. Additionally, when estimating destinations, the package uses a simplified version of the utilities of the alternatives that ignores the logsum (expected utilities) of interdependent choices and obtains a sample by simulating decisions with the corresponding logit probabilities. The final real utilities are then only calculated for the alternatives in the resulting sample.

While the simplest method of sampling that could be implemented is Simple Random Sampling, the suggested sample size of one eighth to one quarter of the choice set is too prohibitive for this model, as the sample would still be too large and have massive memory requirements.

It is also easy to see the challenges in implementing Importance Sampling. While the consistency over the years of the format of the data would allow for some systematization of this method, the measure of importance which is left to the criteria of the modeler would introduce bias.

In the case of Strategic Sampling, although the literature suggests it can be effective even when selecting relatively small sample sizes, the requirement for different iterations, especially on a step that is hard to perform without human intervention such as the re-estimation of parameters, make this method a time-consuming alternative and only raises the barrier to develop an activity-based model.

Sampling by using simplified utility functions (with no interdependence) could be a practical solution, however, the impact of this method on accuracy and possible bias is unknown. Likewise, Stratified Importance Sampling with activity spaces could prove a useful alternative that overcomes some of the limitations of other methods while not requiring to ignore interdependence of choices for the sampling of alternatives, an advantage that could be relevant given the importance of mode choice behavior in the policies described in section 2.1 and the interdependence of mode choice and destination choice (a large choice set) as described in section 3.2.1. Testing is needed to determine the accuracy of these methods.

Logit models appear in general to have several advantages over rule-based models and the multistate supernetwork extension for their use in activity-based models. While rule-based models still suffer in accuracy with large choice sets, and multistate supernetworks are difficult to manage and implement, modelers are much more familiar with logit models which already provide good explainatory value, have well established estimation methods, and as previously detailed have seen considerable research in dealing with the difficulties imposed by large choice sets. However, the accuracy of logit models that are estimated using these methods and for the interdependent choice structure needed warrants testing.

Beyond the state-of-the-art methodology, the replicability and adaptability of the models eases their deployment by means of improving their verification (Hörl & Balac, 2021), and reducing the time and costs associated with developing them (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b; Hörl & Balac, 2021).

In this regard, it is useful to evaluate practical considerations in the use of activity-based models such as the availability and suitability of data (Shiftan & Ben-Akiva, 2010), and the use of open software (Hörl & Balac, 2021). Hence, these aspects are also further explored in this review.

3.3. Replicability and adaptability

Some attention has been put in the replicability and adaptability of activity-based transport models, especially in the context of the data and tools used. Hörl and Balac (2021) highlight that the use of proprietary data and tools does not allow other parties to examine and use the model, therefore limiting the reach and usability that they could otherwise get, while Shiftan and Ben-Akiva (2010) highlights that new data collection as opposed to using available data would present additional hurdles.

Likewise, the Association of Metropolitan Planning Organizations Research Foundation (n.d.-b) and Castiglione et al. (2015) remark that the time and costs associated with deploying an activity-based model are hard to predict upfront and can pose a significant constraint, therefore, we can argue that using available and flexible tools to reduce them can be beneficial.

3.3.1. Data requirements of activity-based transport models

Trends can be identified in the way activity-based models make use of data. As highlighted by Ortúzar and Willumsen (2011) and Shiftan and Ben-Akiva (2010), activity-based transport models have greater data requirements as they model choice in a disaggregate manner, contain a larger number of alternative choices and more unknown parameters than less advanced models, which means that use of existing data needs to be maximized, or new data generated. This is specially true considering the lack of purpose-collected data when evaluating innovative policies.

In past works, very often data is collected with the sole purpose of building activity-based transport models, especially in the form of travel diaries. This is the case of the works of Arentze et al. (2000) and Arentze et al. (2005), Philip et al. (2013), while Daisy et al. (2020) and Hafezi et al. (2019) even go to the lengths of using GPS data to augment and verify these travel diaries.

However, according to Shiftan and Ben-Akiva (2010), even when the lack of detailed data collection can present a considerable hurdle in the development of activity-based transport models, the problems present in collecting detailed data prompts researchers to keep surveys to a minimum level of complexity and instead make use of already available data. In this regard, Arentze and Timmermans (2004b) makes use of travel diaries from the Rotterdam region in 1997, while Dianat et al. (2020), Han et al. (2021), Hörl and Balac (2021), and Knapen et al. (2021) simply make use of the data available in existing travel surveys.

In the context of the Netherlands, the availability of data from a Dutch travel survey, ODiN, that is collected annually (Centraal Bureau voor de Statistiek, 2021) paints a favorable picture towards models that are easy to estimate, update, maintain and replicate.

3.3.2. Software tools for activity-based models

In addition to available data like existing travel survey data, Hörl and Balac (2021) also recommend the use of open source software tools that make their code available, which improves the adaptability of the modeling process in different contexts.

The possibility to use these tools and data also gains particular importance when accounting for the cost and time needed to develop activity-based models. Castiglione et al. (2015) mentions that while the cost and time needed to develop these models has greatly decreased in recent times, the implementation of new features and software can still prove significant, and the needed calibration for models can have uncertain costs and schedule. Open source tools and available data could provide a way to mitigate this.

Open source tools such as MatSim have been already used for activity-based models in the works of Eluru and Choudhury (2019) and Han et al. (2021), which comes in stark contrast to the case of the Melbourne Activity Based Model from Infrastructure Victoria (2017), who exclusively makes use of proprietary data and tools and whose model details are not available to the public, possibly increasing the costs associated to the model and eliminating the possibility of outside verification.

Besides MatSim, ActivitySim is presented as another open source software option to make activity-based models. ActivitySim is actively maintained, expanded, and explicitly developed by a consortium of planning agencies in the United States for reusability across different contexts (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b), which is possibly an advantage over MatSim, which has less built-in integrations and whose extensions are driven by research projects without much regard for reusability (ETH Zürich, 2016). However, it is worth noting that because of ActivitySim being developed or use in the United States, a quick look into its needed inputs it is evident that it was developed to use information from a different travel survey design in mind; travel surveys used in the United States have discrepancies from those used in the Netherlands (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b; Centraal Bureau voor de Statistiek, 2021).

There are also open software options for the estimation of logit models. Biogeme and Larch are examples of this, specializing in making model specification and parameter estimation simpler (Bierlaire, 2018; Newman, 2021). In this regard, it is also worth noting that ActivitySim has integrations with Larch, with the explicit intention of improving the reusability, replicability and adaptability of the model

(Association of Metropolitan Planning Organizations Research Foundation, n.d.-b).

In general, the availability of data and open source software seem to present an opportunity for deploying activity-based models in the Netherlands with relative ease.

3.4. Activity-based transport models in the Netherlands

In this context, a couple of important activity-based transport models developed for the Netherlands deserve attention. One is a model for the Eindhoven region developed by Arentze et al. (2005), and another is the ALBATROSS model for the Dutch Ministry of Transportation, Public Works and Water Management developed by Arentze and Timmermans (2004b), which is often integrated in the not explicitly activity-based Dutch National Model System (LMS) as needed (de Jong & Kroes, 2008). Both models made use of purpose-built travel diaries and a rule-based approach, and while the data that had to be purposefully collected at the time is nowadays more widely available, the rule-based approach still means that they would require more significant extensions to add sensitivity to new policies. The ALBATROSS model also has the limitation of assuming a single mode for an entire tour (Hasnine & Habib, 2018), which introduces limitations on the testing of policies specifically designed to favor multimodal travel. Additionally, these models also have little verification and their forecasting power and transferability are unaccounted for (Dianat et al., 2020).

The newer model by Knapen et al. (2021) was explicitly developed to be sensitive to mobility hubs in the Netherlands, and it makes use of discrete choice modeling, hence, it does not suffer from the same problems that its rule-based counterpart does. However, the methodology described still leaves the previously explored issues of choice models with large choice sets unresolved, an issue that Bao et al. (2018) note that the used model formulation has. Likewise, the need to use multiple tools with purpose-made integrations make the maintenance and replicability of the model less straightforward, even if it uses readily available travel survey data.

3.5. The gap in the research

In the trade-offs between accuracy and practicality for real life applications, logit models seem to be the most promising alternative to model choices in activity-based models as they provide good explainatory power and do not seem to lose accuracy with large choice sets (like it is the case of destination alternatives in large areas) nor explode in size with the number of choice dimensions modeled. Logit models are also well established in the literature and their estimation procedure well known, to the point that as software packages exist that make model specification and parameter estimation simpler (Bierlaire, 2018; Newman, 2021).

However, logit models do still face issues when dealing with these large choice sets, especially when there is interdependent choices, as the need to calculate all the possible utilities can significantly slow down model runs. For this, sampling methods for large choice sets have been developed in the literature, of which the use of simplified utility functions, and Stratified Importance Sampling could be interesting to test and evaluate their accuracy. This also places special importance in destination choice modeling.

The possibility of using existing data in the form of the National Travel Survey is also important, as it further simplifies model estimation and increases replicability and adaptability of the model, especially as it makes it easy to update the model parameters when new data is released. Previous models in the Netherlands have already made use of the survey data, however, they stop short of providing any guidance that helps in the replicability of the data preparation nor the estimation of the parameters derived from this data.

Additionally, these models also do not make use of frameworks that can easily be reused, opting instead for their own custom-made submodels and integrations without a focus on being able to reproduce the work either. Open source software packages such as ActivitySim seem to unveil a potential to do this, making activity-based models much simpler to apply, maintain, and replicate, especially with useful and reusable integration of the different modeling steps.

3.5.1. Research questions

Dutch decision-makers in the arena of transport planning could benefit from an activity-based transport model formulation that uses discrete choice modeling, which has proven easy to calibrate, integrate, and to incorporate sensitivity to relevant policies to it (Castiglione et al., 2015), so that it could have an advantage over existing alternatives. A possibility is to use ActivitySim, an open source software package developed to estimate activity-based models in the US as a starting point to estimate such a model in the Netherlands, and use the already available Dutch travel survey data (Centraal Bureau voor de Statistiek, 2021), to aim for a model that is adaptable and easy to update.

Such a formulation needs to overcome the data requirements of the software, which were developed for a differently designed travel survey. It should also consider advances in the sampling of alternatives for computational efficiency and test the suitability of the built-in sampling method used in the software and of other recently developed sampling methods to optimize model run times and predictive capability.

Such formulation, however, remains to be implemented and tested, which brings brings us to the main research question:

What is the accuracy and efficiency of an activity-based model developed using travel survey data and open source software?

This research question is also broken down into the following subquestions:

- *What are the advantages and limitations of using ODiN data and the ActivitySim software to estimate activity-based model parameters?*
- *What is the accuracy of destination choice models developed using ODiN data and ActivitySim?*
- *What is the impact of Stratified Importance Sampling on model accuracy and performance?*

4

Research approach

This research aims to develop an activity-based transport model for the Netherlands, in which the gaps identified in Section 3.5 are addressed, and the performance of the model is benchmarked.

This project will estimate the choice model parameters to be used in an activity-based transport model, and then test the usability of such a model for policy planning, while ensuring that the model can be easily be updated with new travel survey data, which makes the framework replicable and reusable.

In this chapter, the research approach followed for the research is outlined, the data and tools needed are listed and described, and the method to answer the research questions is explained.

4.1. Research approach

This research project will follow an empirical analysis and process design approach, as it focuses on estimating efficient and accurate choice models using existing data and software, and then on devising a useful procedure to build a reliable and usable activity-based transport model in the Netherlands with it. The main outcome of this project is the procedure itself, and the performance evaluation of the model. This model, or others that follow it, can then be used by analysts and researchers to make sense of travel patterns and evaluate policy impacts.

4.2. Data requirements and data sources

A first step in the model development consists in identifying data requirements. Modeling the choice behavior of travelers requires a significant amount of data about the travel destinations (zonal data), the network, and the travelers themselves. Additionally, the policies that will be tested need to be considered to make sure that the model is sensitive to them, for which additional data requirements

might be imposed. For this project, TNO made available a number of data sources that are used in models for the Metropolitan Region Rotterdam The Hague, hence, for practical purposes, this project will focus on this region.

4.2.1. Zonal (land use) data

Zonal data is also in the possession of TNO and it also corresponds to the data used for the transport models of the Metropoolregio Rotterdam Den Haag (2021). Zonal data defines a number of properties for every TAZ (Traffic Analysis Zone), like the number of jobs (broken down per type of job), number of study places (broken down per type of students or level of education), population, area size, urbanization level, and other characteristics that may be needed for the model.

This zonal system splits the surface of the Netherlands into 7786 different TAZs (MRDH zones), which are much smaller in size in the Metropolitan Region Rotterdam The Hague than in the rest of the Netherlands. This has to do with the level of aggregation desired in the model, where outputs are needed in much more detail for the area of interest (the Metropolitan Region Rotterdam The Hague) that they are needed in the area of influence (rest of the Netherlands). The area of influence is, however, still included since as its name indicates, it can still affect model outcomes, namely by serving as origin and destination of trips that also traverse the area of interest.

In rough terms, the numbering of the zones increase the further away they are from the Metropolitan Region Rotterdam The Hague, and thus their area size also tends to increase with the numbering.

4.2.2. Network data

Network data can be obtained from sources like OpenStreetMap (2021) or publicly available resources like the data register of the Rijkswaterstaat (2018), and then processed into skim matrices that reflect the generalized cost or disutility of traveling at different times during the day. They can be processed with the use of software tools like OmniTRANS or MatSim, however, in this case, TNO has made available already processed skim matrices that were obtained from the transport models of the Metropoolregio Rotterdam Den Haag (2021) which will be used in this model.

This data has been processed in the same system as the land use data, that is, into MRDH zones, which means that each skim matrix contains network level of service information for 7786 x 7786 origin and destination pairs. This data set contains the generalized travel costs between every origin and destination pair, per every possible mode of transportation modeled, and for different times of day during weekdays (morning peak from 7 to 9 am, afternoon peak from 4 to 6 pm, and rest of day).

4.2.3. Survey data

Data from the travelers is obtained in the form of the Dutch National Travel Survey, collected by the Centraal Bureau voor de Statistiek (2021). This survey contains revealed travel preferences for the Netherlands, and since 2018 a new survey methodology was introduced to create the Onderweg in

Nederland (ODiN) survey, which makes current data not comparable to prior data, which means that special attention will need to be placed when processing the data using methods from previous years to make sure that the data remains logical and consistent. Additionally, this makes it more complicated to compare with previous models that used data in the old format. Demographic data can also be used to complement this set, obtained from the CBS (Centraal Bureau voor de Statistiek, 2019).

This data set contains information of all the trips performed by respondents on a specific day, like origin and destination, time of departure and arrival, purpose of the trip and mode of transportation; as well as personal information about the respondent such as home zone, income level, vehicle ownership, driving permit possession ownership, employment, age, and gender. A description of all the fields in the survey data can be found in Appendix A.

This data codes origins and destinations by post code zones (PC4), while the zonal and network data obtained from the MRDH model are coded in MRDH zones. This means a translation from PC4 to MRDH zones is needed, which in this case, was performed based on travel purpose and zonal data, that is, a PC4 destination for a trip with a certain purpose is mapped to the MRDH zone that overlaps the PC4 zone and has the most relevant zonal characteristics, for example, the most jobs in a work trip.

Data from 2019 is used, as the latest available data (2020) is deeply influenced by the restrictions on mobility put in place to curve the spread of the COVID-19 pandemic, rendering it unusable for our purposes. The ODiN survey for 2019 contains 179091 entries from 53380 different respondents, each from a different household. Each respondent is asked to fill in information about all their trips (travel diary) performed on a single day of the reporting year.

4.2.4. Other data

Additionally, a synthetic population is used as the population for the simulation runs after the model is estimated. This synthetic population reproduces the characteristics of the aggregated land use data to provide a close match to what the real population looks like. For this project, TNO provided an already processed synthetic population, which is already formatted to the requirements of the model, containing the fields like age, sex, employment, household, person type, student type, possession of driving license, education status, and possession of vehicles.

Finally, there are data requirements to validate model results. Possible options include outputs of previous models, and alternative data sources like commercial GSM data.

4.3. Software and tools

The main tool to be used is ActivitySim, an open source activity-based modeling tool developed using the Python programming language (Association of Metropolitan Planning Organizations Research Foundation, n.d.-b). Python offers as an advantage that many useful open source libraries have been developed for it, like various tools such as Pandas for data processing, Larch for choice model estimation (which already has some coupling with ActivitySim), and openmatrix and pyyaml to handle data

inputs.

ActivitySim first performs a series of steps to setup the data that it needs, to then execute choice submodels in a structure equivalent to that described in section 2.3, and finally export the results. Long term choices (workplace and school location) are decided first for every person in the population, then medium term choices like car ownership and whether or not the person has access to free car parking, subsequently activity patterns for the day are decided (tour generation, scheduling, mode choice, etc), and finally trip level decisions are made (adding additional stops in tours, trip purpose and destination, and mode choice). The model structure is visually represented in Figure 4.1.

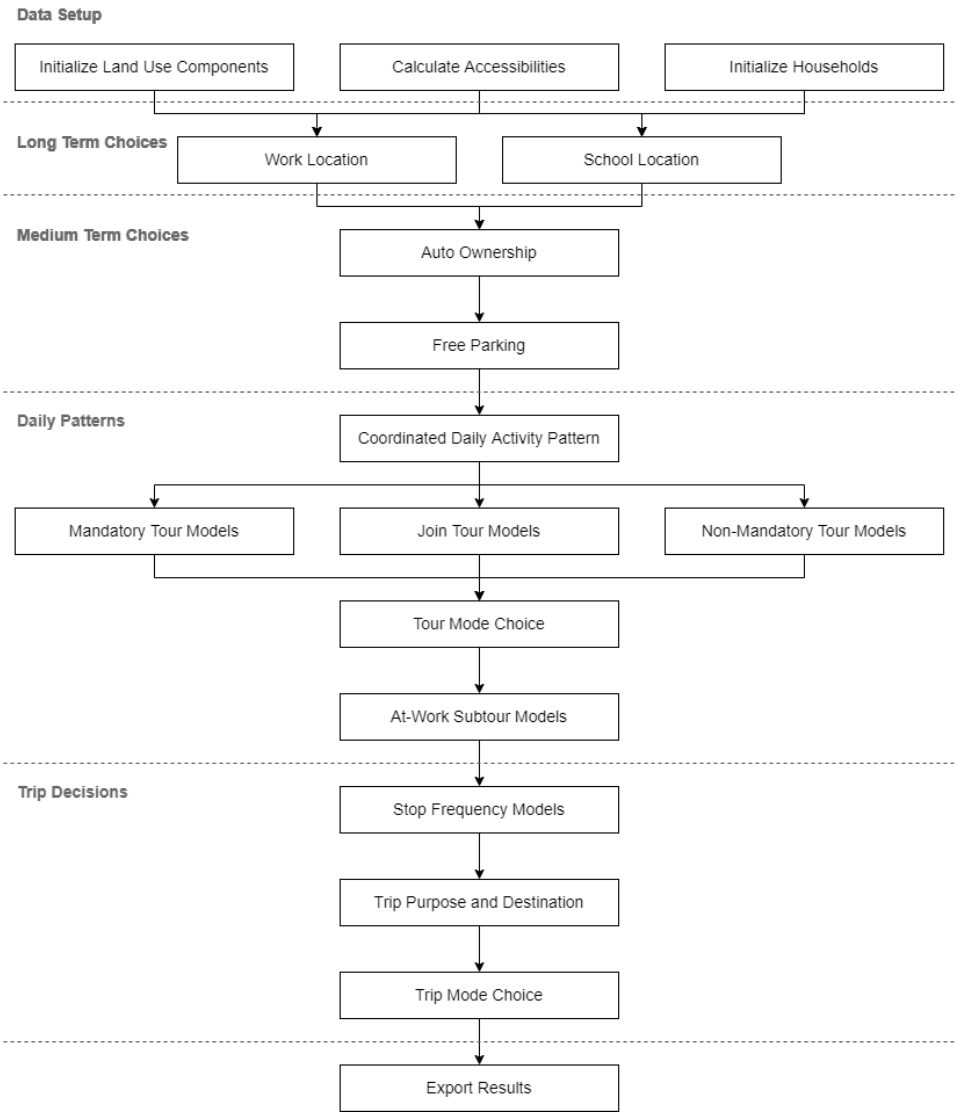


Figure 4.1: ActivitySim submodels, based on Association of Metropolitan Planning Organizations Research Foundation (n.d.-a) and Zephyr Transport (2020).

The use of the ActivitySim framework also presupposes the use of logit models, as one of the big advantages of this software package is that it has support for it. The ActivitySim framework was designed to be reusable and relatively easy to set up, and as such it has functionality added to estimate model parameters. Inputs are easily loaded in the form of the data previously described in csv tables and omx matrices, model settings are written in yaml files, and choice model parameter values are written

in another set of csv files. In some cases, some data processing and the calculation of additional fields might be necessary between model steps, which is made possible by specifying the different calculations on additional csv files. This becomes relevant for example when calculating the expected utilities of a choice to be used in another choice model, for example, the expected utility of the mode choice for the destination choice which is calculated as the logsum of the utilities.

To be able to estimate model parameters, ActivitySim can first perform a run in ”estimation mode” using any initial parameter values for the first iteration. ActivitySim runs the simulation using these parameters and running the choice models for the population of the survey. The explainatory variables used, the simulated choice, and the choice that was actually made (known or inferred from the survey data) are stored in what the developers of the software call an ”Estimation Data Bundle” (EDB), as this contains the data that would be needed to re-estimate the model parameters and evaluate the fitness of the model.

Here, the Larch integration of ActivitySim comes in handy. Larch is a library that was developed to perform maximum likelihood estimations for logit models, that is, it estimates the model parameters, which in this case does so by using the Estimation Data Bundles as input. The Larch integration produces the new parameters and measures of likelihood and fitness of the new choice model, as well as statistic significance for each parameter. Based on these results the modeler can gauge it a satisfying model has been produced or if a new iteration needs to be run in estimation mode.

After satisfying models have been estimated, ActivitySim can then run a full simulation using the synthetic population.

Additionally, an SQL server is set up to host, access, and query the survey data in an accessible and easily manageable way, with the help of software like HeidiSQL and connections with the Pandas Python library.

While it will not be used directly throughout the course of this research project, it is also worth mentioning that OmniTRANS (Goudappel, 2022) was used in the processing of the skim matrices.

Finally, to handle the computationally expensive model, a remote server with enough processing power will be used to achieve manageable runtimes. This server has a 40 core 2,4 GHz CPU and a memory of 130 GB.

4.4. Method

In this section, the method used to execute the research and answer the research questions is described.

4.4.1. ODiN survey data processing

While network and zonal data require to be put in the right format for ActivitySim to take it as input, travel survey data requires considerable processing to fit the input requirements that the package imposes

to estimate the model, as seen in Figure 4.2.

In an ActivitySim simulation, the zonal data, network data and a synthetic population would be passed as inputs and the package would give a series of trips and tours as an output. In this case, however, ActivitySim offers a useful functionality, called estimation mode, to estimate the parameters of the model from a travel survey that is formatted as the population table but also as the outputs, so that the model can compare its own output with the expected output in what is called an Estimation Data Bundle. This is then used by a logit model estimation tool such as larch, which has some integration with ActivitySim, and the new parameters for the model are estimated, which are then fed to ActivitySim to run a new simulation.

To obtain the Estimation Data Bundles, the survey data needs to be processed into a set of different tables, namely households, persons, tours, joint tour participants, and trips tables, with the expected fields mentioned in Figure 4.2. The challenge lies in that ActivitySim was designed for US survey data, which is different from the ODiN data set, and thus, many transformations and the inference of implicit data need to take place.

This step needs to have as a result the formatted survey data and the replicable methodology to process the ODiN data from subsequent years, so that the model can be easily re-estimated in the future.

4.4.2. Discrete choice modeling

After obtaining the Estimation Data Bundles, the model parameters can be estimated. These parameters correspond to the arguments in the utility functions of the different choice alternatives that are defined by the user in the settings files.

The estimation process yields as a result a measurement of fitness, and a measurement of statistical significance per every argument in the utility function. The utility functions can be changed using this information and repeating the previous steps until a good fit is achieved. Special attention needs to be placed in arguments that introduce sensitivity to desired policies, as if these are removed, so is the sensitivity and the usefulness of the model.

Then, ActivitySim can be used to run a simulation using these parameters and a synthetic population. These simulation results need to be evaluated and validated. Figure 4.3 details the workflow to process the data and arrive to a re-estimated model.

4.4.3. Sampling of destination choices

As described in subsection 3.2.4, ActivitySim already implements a form of destination choice set sampling that is based on a simplified simulation of the utilities of the alternatives. Based on the model outputs, this method needs to be tested for the Netherlands, and if necessary, improved or changed and compare the performance of the alternative.

Stratified Importance Sampling can be useful to compare against ActivitySim's default sampling. The

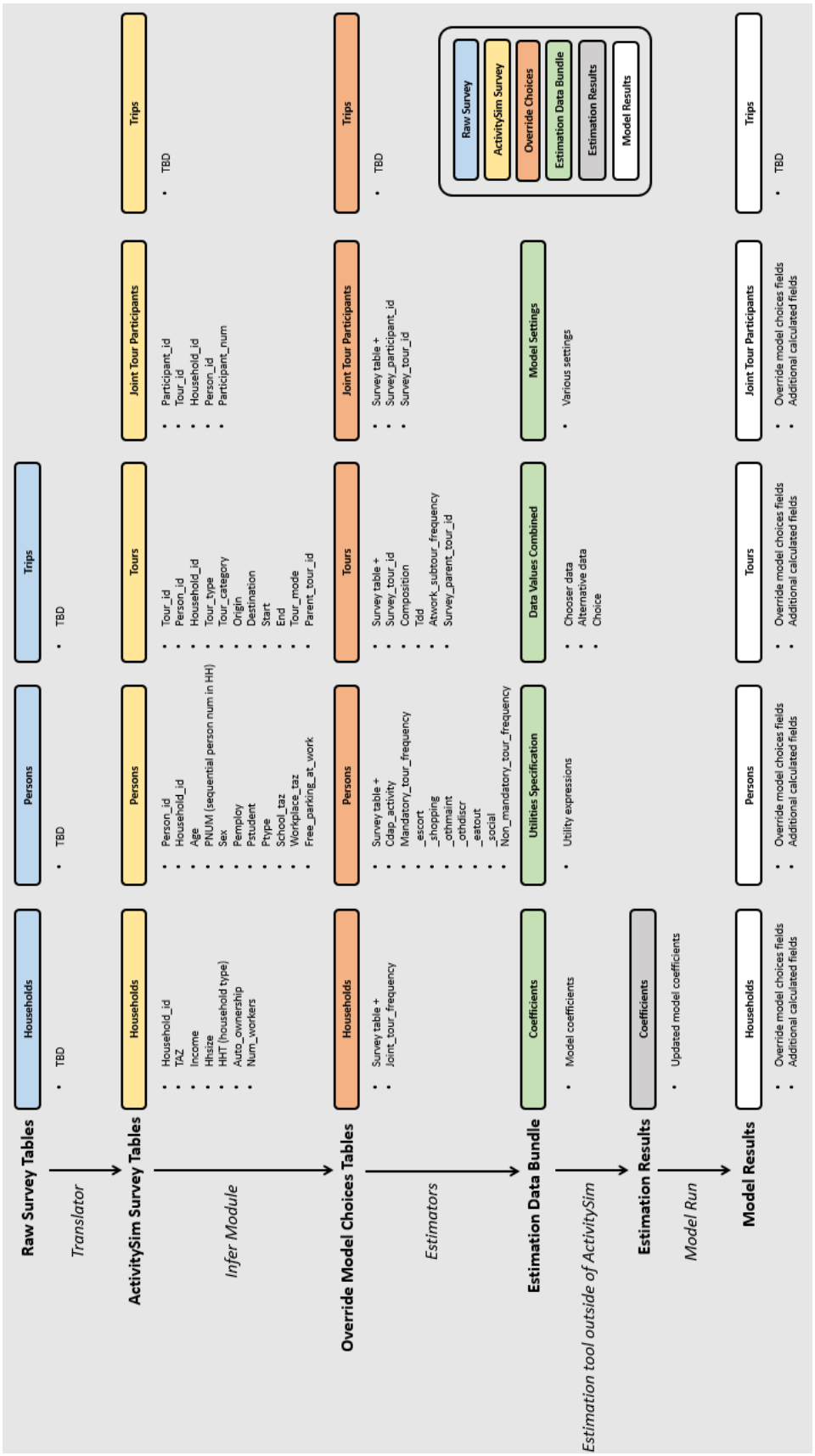


Figure 4.2: Data flowchart for ActivitySim, extracted from Association of Metropolitan Planning Organizations Research Foundation (n.d.-b)

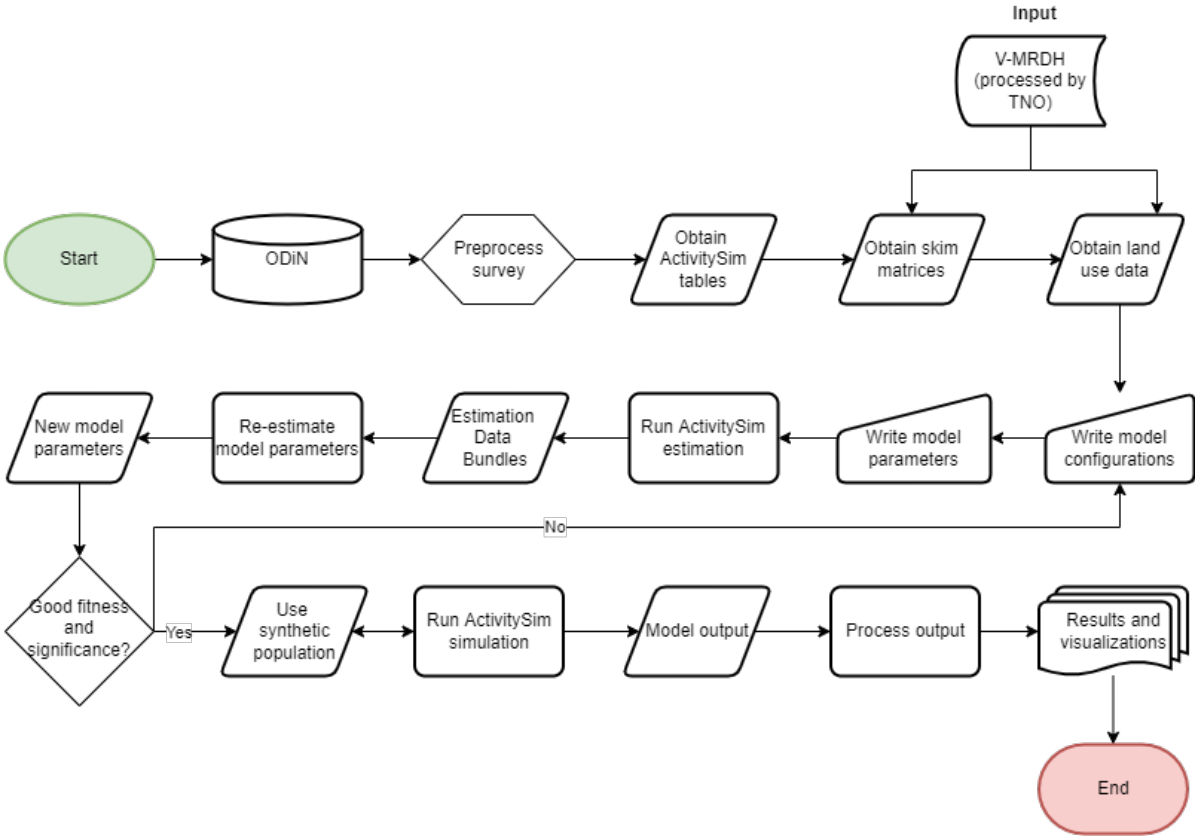


Figure 4.3: Workflow for re-estimation of the model parameters using ActivitySim and ODiN data.

samples can be systematically obtained from the survey data, by looking in the origins and destinations for the specific trip purpose present in the data, obtaining the distances between them, and using that to define a distance search radius from each possible origin. The zones that fall within the radius have equal probability, while the zones outside have a probability of zero. This can then be swapped instead of the default sampling, meaning that the possible destinations will be drawn form here, and the new results can be evaluated and validated.

4.5. Validation and performance measurement

The model output will be evaluated in a series of metrics. The first and most important is the model accuracy, where the predictive power of the logit destination models needs to be measured. An initial measurement is provided in the estimation step, where the fitness ρ^2 value gives an idea for every choice model.

This however, does not provide the full picture, and simulation outputs need to be compared with real data to provide insights in its usability in predicting mobility across the Netherlands. The choice of workplace and school locations for the synthetic population can, for example, be compared with the jobs and student capacity known from the zonal data, and the distance to work and school can be compared between simulation output and what is known from the survey data. It is also important to take a look into specific cities and not just the aggregated result.

Additionally, since efficiency also has an important impact, runtimes in the available hardware will be measured.

5

Results

Following the outline of the methodology described in Chapter 4, in this chapter the results of each step will be presented and interpreted in light of the research questions.

5.1. Data processing

While being the most time intensive stage in the project, the steps taken to process the ODiN survey data resulted in the data in a format that matches most of the requirements imposed by ActivitySim, even if the package was not designed for compatibility with ODiN.

Since the policies conceived in section 2.1 mainly focus on travel that is not performed as a professional activity (i.e. driving a truck), professional drivers needed to be removed from the data. Likewise, to obtain the data for a ”typical” day in the Netherlands, the survey first needed to be filtered to contain only information from weekdays between 5 am and 10 pm and to remove respondents with missing information in a needed field (i.e. missing information about employment status). This proved easy to do by querying the ODiN data from the SQL database that was previously set up. Additionally, the focus on the Metropolitan Region Rotterdam the Hague also requires to filter out respondents that do not perform any traveling on this region.

Second, all origins and destinations needed to be mapped from postcodes to MRDH zones. To create a mapping that still provided reasonable destinations for trips based on purpose, the shapefiles of both zone systems were laid on top of each other and new ”fragments” are created with the different overlaps. The zonal data from the MRDH zones is then also split proportionally to the area of the fragment, that is, if a fragment has a percentage of the area of the MRDH zone, then it gets assigned the same percentage of the zonal data values. Then, a certain postcode location could be mapped to either of the zone fragments that originated from an overlap with said postcode, and the fragment with the greatest value for the relevant zonal data field according to the trip purpose (i.e. employees for

workplace location and population for home).

The next step was obtaining the workplace location and school location. Workplace and school locations are not explicitly defined in ODiN, however, it can be assumed that they correspond to the destination of trips that have as purpose to go to work and school respectively, for every survey respondent.

Then, the different tours and subtours were identified in the data. Tours are home-based and subtours are work-based, which means that all trips counting from the departure from the workplace location until its return and without going home first are a subtour, and all the trips since the departure from home and until arrival back at home, excluding those that are part of a subtour, are a tour (and the parent tour of any subtours that may have been performed). An ID is given to every tour and the respective parent tour ID is given for every subtour.

After, the outbound and inbound trips of every tour, defined as the trips before and after arriving to the main destination of the tour, were respectively identified.

The data was then subset to obtain the households table as required by ActivitySim (figure 4.2). It was assumed that there was only one unemployed person in the household if the respondent was unemployed, and non otherwise. All other adults in the household are employed. The median income deciles from ODiN were mapped to median monetary values using CBS data, and the household type was obtained based on the household composition and gender of the respondent according to specifications obtained from Association of Metropolitan Planning Organizations Research Foundation (n.d.-a).

For the persons table, the employment status, student status and person type were also defined according to the specifications of Association of Metropolitan Planning Organizations Research Foundation (n.d.-a), based on the age of the respondent, paid work status, and possession of a student OV chipkaart, as well as the age of retirement in the Netherlands. Since children can have some information unavailable from the survey due to privacy concerns, it had to be assumed that all children from 12 years of age and older have finished primary school. Coordinated daily activity patterns (CDAP) were obtained by obtaining the tours per person, and identifying persons who took mandatory tours, persons who only took non-mandatory tours, and persons who stayed at home on the reporting day. Mandatory and non-mandatory tour frequency were obtained by counting the number of corresponding tours (per purpose) and mapping the counts to alternatives specifying a combination of tours per purpose.

For trips, modes were grouped into walk, car, car passenger, bike, ebike, public transport (with walk as access and egress mode), and DRT. Additionally, the survey contained stay at home ”trips”, which are false empty trips that occurred when the respondent did not leave home for the entire day of reporting, and thus have to be removed.

Subsequently, for tours, the tour purpose was selected from the purpose of all its trips according to an assumed hierarchy, and the destination of the tour was set as the destination of the trip whose purpose matched the tour purpose. The origin and departure time were set as those from the first trip in the tour, and the end time was that of the last trip. Tours were then categorized as mandatory, non-mandatory, and at work. The stop frequency was inferred by counting the stops per (outbound and inbound) part of the tour and likewise the at work subtour frequency was obtained by counting the subtours of work tours. The tour departure and duration (scheduling) was obtained from departure and arrival times and

mapped to a single combination ID that would be used by the model for ease. Lastly, the tour IDs were redefined by using what is known as ”canonical tour IDs” by the developers of ActivitySim, which provide constant and replicable IDs based on the person ID, tour purpose, and tour number for the person.

The processing of ODiN survey data yielded as a result the processed data for 5514 persons, who performed 16545 different trips as part of 7204 tours (of which only 24 are at work subtours, and none identified as joint tours). These will be used as input to estimate model parameters in ActivitySim.

The procedure developed to process the data is summarized in the flowchart in figure 5.1. No joint tour participants table was possible to extract as very limited information exists about joint tours. Likewise, no tours were classified as joint because of the same lack of information. Other fields needed on these tables were directly available from the survey data. A summary of the data processing decisions and assumptions made and the calculations performed can also be found in a table format in Appendix B.

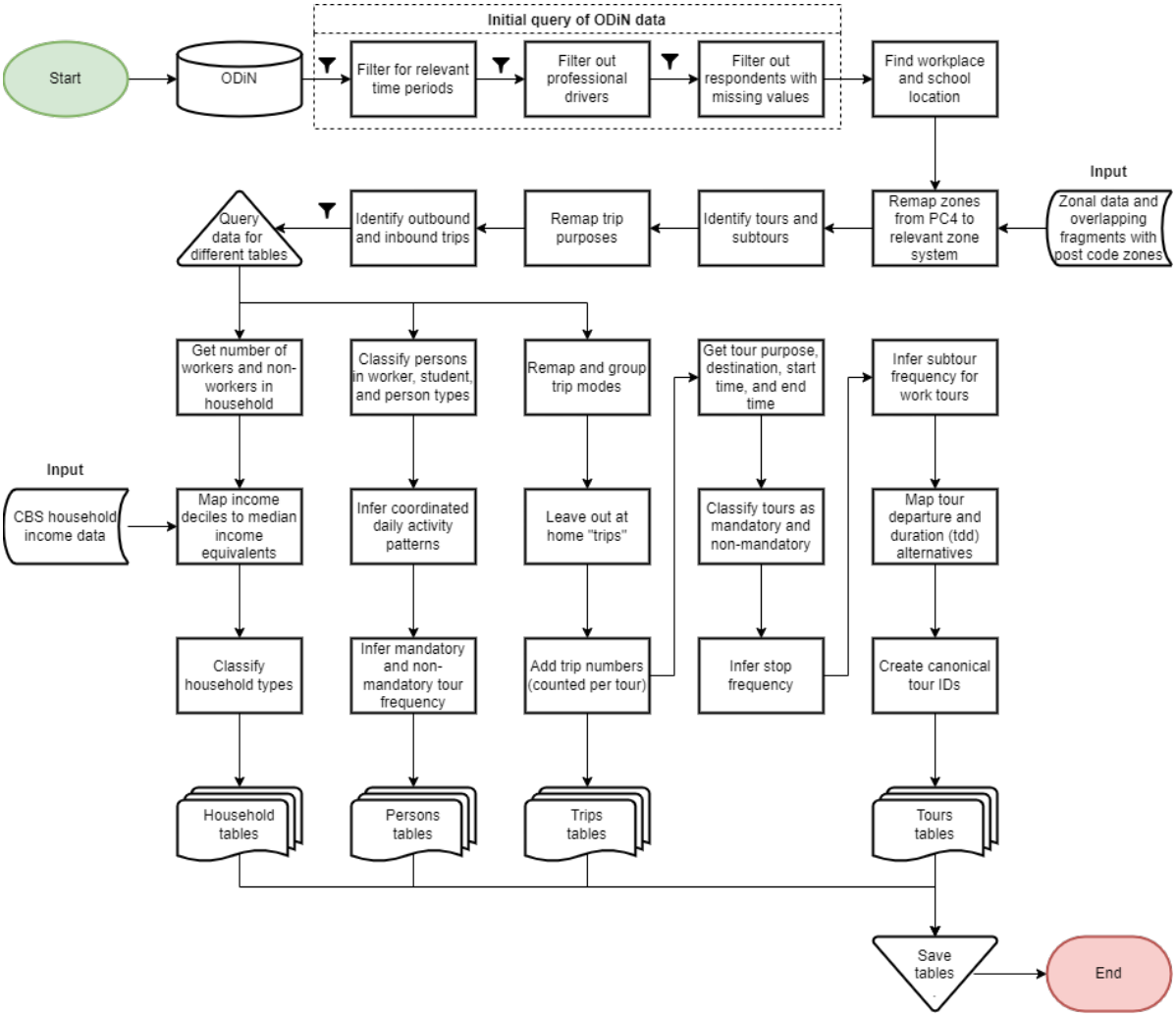


Figure 5.1: Flowchart diagram depicting the procedure to obtain the tables required by ActivitySim for model estimation from the ODiN travel survey.

5.1.1. Data completeness

Although as previously mentioned most of the information needed was present in the data, some exceptions exists, not only in the need to draw CBS income data but also regarding joint tours and household members. Joint tours information is largely unavailable in the ODiN data, as respondents are not asked about other participants in their tours beyond information requested about their role in a vehicle (driver or passenger) and if the purpose of a tour or trip was to escort someone else. This affects the joint tours table and the tour composition field in the tours table which describes the composition of joint tours (see figure 4.2), and which are needed to calibrate the model.

Likewise, only one person is surveyed from a given household, and while household information is requested in the survey, including the household composition, very little disaggregated information nor trips and tours performed is obtained for individuals other than the respondents themselves, limiting the possibility of evaluating household interactions in the model.

Both these issues occur due to the differences in the survey design between the Dutch travel survey (ODiN), and the US travel survey for which ActivitySim was designed. Contrary to travel surveys in the US, the ODiN survey is applied to a single respondent per household, and only travel information from that respondent is asked.

Additionally, as mentioned in section 5.1, the school and workplace location is determined by obtaining the destination of trips with the corresponding purposes per person. The absence of this data directly from the survey, however, introduces an important limitation in that even if the person is known to be a student or employed, if said person did not perform a trip to the relevant location on the reporting day the location is unknown from the survey. These entries, however, are kept in the data since they still provide valuable information for other trips.

The number of subtours obtained in the data also seems to be rather low in comparison to the total number of tours. This might put into question the suitability of this data to predict these subtours.

In light of the policies previously outlined in section 2.1, it relevant to know that there is available information in the survey about the fuel type of cars owned in a household, including electric and hybrid vehicles, and information about engagement in car sharing (although the sharing of other modes of transportation remains missing), which is needed to include more detail about these modes of transportation in the model.

Another additional bit of information that can be useful in future models is related to remote working. When a respondent is asked to fill in their travels for a given day, but that respondent did not leave home during the whole day, more information is asked as to for what reason this was the case, and one of the possible answers is working remotely. While not directly related to any sustainable mobility policy mentioned in this work, it is reasonable to think that if remote working changes travel patterns this will have to be taken into account in future models to obtain realistic outputs.

5.1.2. Fitness for purpose of the data and documentation available

The data from ODiN could be easily split in persons data, households data, trips data, and tours data. Moreover, almost all the fields required by ActivitySim were able to be obtained from the survey after performing a series of transformations and calculations.

This was done using minimal additional sources of data, the exception being household income. ODiN does not directly provide household income values in euros, and instead splits households in income groups (deciles) by using the income distribution of the Netherlands and splitting it into groups of 10% of households each. To get back the income value in euros, this process needs to be somehow reverted, for which the median income of each income group was obtained from the Centraal Bureau voor de Statistiek (2019) and mapped to every household of the corresponding income group. This still leaves 10 income groups, but the field is expressed in a monetary amount.

Much of the information needed to arrive to the correctly formatted data is available in the ActivitySim documentation, such as the definition of student types, worker types and person types, which were then adapted to the context and availability of information of the Netherlands, like a different age of retirement and the lack of detailed data for small children.

Following this, the fields required to override the model choice in the estimation process could be inferred in a way very similar to that used to obtain them from the US survey (for which ActivitySim was conceived). The assumptions made to process the data are detailed in Appendix B.

There were, nonetheless, some unexpected complications in processing the data that came from field definitions that are left implicit in ActivitySim's documentation, and had to be deduced by looking into the example cases and the source code. One example of this issue is the use of something called "canonical IDs" to identify tours, which essentially make tour IDs consistent across the different processes in the model, and even across simulations. The model required such an ID scheme to be used in the data that was input for estimation, but no explanation is provided in advance, and only looking at the source code directly provides clear answers.

5.1.3. Implications for the modeling process

Since the ODiN data has the same format every year, this procedure, and the script in which it was performed, can be reused to obtain the necessary data from any other year with ease, which would eliminate the need to spend considerable time in processing this data for ActivitySim, as well as lowering the barrier to develop and calibrate an activity-based transport model for the Netherlands. The steps needed to process the data are described in the relevant script.

Still, minimal modifications need to be made in the programming of the ActivitySim package to deal with the missing information for joint tours and joint tour participants, which in this case is only needed in one step and thus can be easily changed.

After obtaining the processed data inputs, writing model settings, and writing initial parameters on the utility functions, ActivitySim can be run in estimation mode to obtain the Estimation Data Bundles to be

used in the estimation of the parameters.

However, the missing workplace and school locations from the data warrant additional processing after obtaining the bundles. The estimation procedure adjusts the parameters of the utility functions of choices by comparing the choices made by the model and the choices known for the survey, but when there are unknown choices in the survey, these must be removed to be able to proceed with estimation.

The issue of missing documentation described also has as a consequence an increase of the time needed for the already time consuming data processing, as considerable time is spent in an attempt to reconstruct the needed format without guidance from the developers of the software.

5.2. Discrete choice modeling

ActivitySim operates under the assumption that the destination for mandatory tours (school and workplace location) is selected only once per person and does not change over time. Additionally, this choice is not done in conjunction with mode choice, and the only aspect of the mode choice that could affect the location choice for these purposes is as part of the utility functions through the use of logsums. To test the suitability of this modeling choice, the data from the ODiN survey is used to run ActivitySim in estimation mode and obtain the Estimation Data Bundles.

The Estimation Data Bundles are used to estimate the model parameters with the help of the Larch library (and its integration with ActivitySim). These parameters are then swapped in for the previously used parameters in the model, and ActivitySim is run in simulation mode using a synthetic population that accurately represents the population according to the land use data. The output from this step is described in this section.

5.2.1. Estimated logit models

The Estimation Data Bundles needed to re-estimate the parameters of the utility functions in the logit models for ActivitySim are obtained by running ActivitySim in estimation mode using the prepared ODiN data. The bundles contain the choices simulated using the survey data and the choices known from the same data, as well as the attributes and parameters that defined those choices. This information is read with the help of the integration between Larch, an estimation software package, and ActivitySim, and the computations are performed to obtain the new parameters and statistics that indicate their suitability, such as the model fitness, and the statistical significance of each parameter. The utility functions are tweaked progressively and this process iterated until utility functions that contain only statistically significant parameters and which provide good model fitness are obtained.

For school location, ActivitySim segments the population into school students, high school students, and university students, and the utility functions that define the school location choice after performing the iterative process can be found in equations 5.1, 5.2, and 5.3, obtaining a fitness value $\rho^2 = 0.85$. In these equations it can be observed that the distance measurements are made in segments, which was done to allow for the distance between home and school to be weighted in a way that is not completely linear (as a constant parameter for the whole distance would imply).

It is worth noting that high accuracy was obtained despite not including demographic variables, probably because they are already implicit in the segmentation and each segment behaves in a relatively uniform way. Additionally, ActivitySim takes into account from the zonal data whether or not an alternative locations is viable, that is, if it has education places, which means that the choice set is constrained which increases accuracy as well.

$$V_{school} = coef\_grade\_dist\_0\_1 * dist\_0\_1 + coef\_grade\_dist\_5\_15 * dist\_5\_15 + coef\_mode\_logsum * mode\_logsum \tag{5.1}$$

$$V_{highschool} = coef\_high\_dist\_1\_2 * dist\_1\_2 + coef\_high\_dist\_2\_5 * dist\_2\_5 + coef\_high\_dist\_5\_15 * dist\_5\_15 + coef\_high\_dist\_15\_up * dist\_15\_up + coef\_mode\_logsum * mode\_logsum \tag{5.2}$$

$$V_{university} = coef\_univ\_dist\_1\_2 * dist\_1\_2 + coef\_univ\_dist\_2\_5 * dist\_2\_5 + coef\_univ\_dist\_5\_15 * dist\_5\_15 + coef\_univ\_dist\_15\_up * dist\_15\_up + coef\_mode\_logsum * mode\_logsum \tag{5.3}$$

$Where$

$dist\_0\_1: distance\ segment\ from\ 0\ to\ 1\ miles$
$dist\_1\_2: distance\ segment\ from\ 1\ to\ 2\ miles$
$dist\_2\_5: distance\ segment\ from\ 2\ to\ 5\ miles$
$dist\_5\_15: distance\ segment\ from\ 5\ to\ 15\ miles$
$dist\_15\_up: distance\ segment\ from\ 15\ miles\ and\ up$
$mode\_logsum: expected\ utility\ of\ the\ mode\ choice$

The estimated parameters of the school location model, their values, and their t statistic (significance) can be found summarized in table 5.1. All the resulting parameters are highly significant with the absolute value of the t statistic higher than 1.96 (over 95% confidence).

Table 5.1: Estimated parameters for school location model

Parameter	Value	t Stat
coef_grade_dist_0_1	-6.13	-24.68
coef_grade_dist_5_15	-1.86	-36.11
coef_high_dist_15_up	-0.287	-19.96
coef_high_dist_5_15	-0.876	-23.14
coef_high_grade_dist_1_2	-3.60	-24.65
coef_high_grade_dist_2_5	-2.99	-52.62

coef_mode_logsum	0.201	3.67
coef_univ_dist_15_up	-0.154	-22.33
coef_univ_dist_1_2	-2.79	-8.40
coef_univ_dist_2_5	-1.91	-20.64
coef_univ_dist_5_15	-0.756	-24.60

The same procedure was applied to the workplace location model, where the population is segmented into low income, medium income, high income, and very high income; and the corresponding utility functions can be found in equations 5.4, 5.5, 5.6, 5.7. Once more, the segmentation means that no accuracy is gained from adding demographic variables in the function, however, this time the fitness value is lower ($\rho^2 = 0.45$) probably because the choice set is much less restricted based on employment availability.

$$\begin{aligned} V_{low} = {} & coef\_dist\_0\_1 * dist\_0\_1 + coef\_dist\_1\_2 * dist\_1\_2 \\ & + coef\_dist\_2\_5 * dist\_2\_5 + coef\_dist\_5\_15 * dist\_5\_15 \\ & + coef\_dist\_15\_up * dist\_15\_up + coef\_mode\_logsum * mode\_logsum \\ & + log(e^{work\_low\_MWTEMPN * MWTEMPN} + e^{work\_low\_OTHEMPN * OTHEMPN}) \end{aligned} \tag{5.4}$$

$$\begin{aligned} V_{medium} = {} & coef\_dist\_0\_1 * dist\_0\_1 + coef\_dist\_1\_2 * dist\_1\_2 \\ & + coef\_dist\_2\_5 * dist\_2\_5 + coef\_dist\_5\_15 * dist\_5\_15 \\ & + coef\_dist\_15\_up * dist\_15\_up + coef\_mode\_logsum * mode\_logsum \\ & + work\_low\_OTHEMPN * OTHEMPN \end{aligned} \tag{5.5}$$

$$\begin{aligned} V_{high} = {} & coef\_dist\_0\_1 * dist\_0\_1 + coef\_dist\_1\_2 * dist\_1\_2 \\ & + coef\_dist\_2\_5 * dist\_2\_5 + coef\_dist\_5\_15 * dist\_5\_15 \\ & + coef\_dist\_15\_up * dist\_15\_up + coef\_dist\_0\_5\_high * dist\_0\_5\_high \\ & + coef\_dist\_5\_up\_high * dist\_5\_up\_high + coef\_mode\_logsum * mode\_logsum \\ & + log(e^{work\_high\_MWTEMPN * MWTEMPN} + e^{work\_high\_OTHEMPN * OTHEMPN}) \end{aligned} \tag{5.6}$$

$$\begin{aligned} V_{very\_high} = {} & coef\_dist\_0\_1 * dist\_0\_1 + coef\_dist\_1\_2 * dist\_1\_2 \\ & + coef\_dist\_2\_5 * dist\_2\_5 + coef\_dist\_5\_15 * dist\_5\_15 \\ & + coef\_dist\_15\_up * dist\_15\_up + coef\_dist\_0\_5\_high * dist\_0\_5\_high \\ & + coef\_dist\_5\_up\_high * dist\_5\_up\_high + coef\_mode\_logsum * mode\_logsum \\ & + work\_very\_high\_OTHEMPN * OTHEMPN \end{aligned} \tag{5.7}$$

$Where$

$dist\_0\_1$: $distance\ segment\ from\ 0\ to\ 1\ miles$
$dist\_1\_2$: $distance\ segment\ from\ 1\ to\ 2\ miles$
$dist\_2\_5$: $distance\ segment\ from\ 2\ to\ 5\ miles$
$dist\_5\_15$: $distance\ segment\ from\ 5\ to\ 15\ miles$
$dist\_15\_up$: $distance\ segment\ from\ 15\ miles\ and\ up$
$dist\_0\_5\_high$: $distance\ segment\ from\ 2\ to\ 5\ miles\ (only\ for\ high\ and\ very\ high\ income)$
$dist\_5\_up\_high$: $distance\ segment\ from\ 5\ miles\ and\ up\ (only\ for\ high\ and\ very\ high\ income)$
$mode\_logsum$: $expected\ utility\ of\ the\ mode\ choice$
$MWTEMPN$: $manufacturing\ and\ wholesale\ employment\ positions$
$OTHEMPN$: $employment\ positions\ other\ than\ manufacturing,\ wholesale,\ and\ retail$

The estimated parameters of the school location model, their values, and their t statistic (significance) can be found summarized in table 5.1.

Table 5.2: Estimated parameters for workplace location model

Parameter	Value	t Stat
coef_dist_0_1	-1.87	-6.35
coef_dist_0_5_high	0.538	16.70
coef_dist_15_up	-0.0880	-24.26
coef_dist_1_2	-1.92	-15.03
coef_dist_2_5	-1.81	-53.83
coef_dist_5_15	-0.657	-68.84
coef_dist_5_up_high	0.0484	12.76
coef_mode_logsum	-0.0343	-2.14
work_high_MWTEMPN	-3.08	-BIG
work_high_OTHEMPN	1.39	3.73
work_low_MWTEMPN	-3.69	-BIG
work_low_OTHEMPN	1.69	2.38
work_med_OTHEMPN	1.59	3.68
work_veryhigh_OTHEMPN	1.05	2.90

The estimated models have good explainatory value and should provide very accurate simulation results, lending credence to the technique of modeling school and workplace location in the early stages of the model as a long term choice.

5.2.2. School location accuracy

After using the newly estimated parameters and performing a simulation with ActivitySim for a synthetic population that closely matches the characteristics of the real population, the output for school location can be plotted and compared with the known school locations from the land use data on the area of the Metropolitan Region Rotterdam The Hague and on a closer level on the different cities. In figure 5.2 the geographical distribution of students across the Metropolitan Region Rotterdam The Hague can be compared between the known distribution of students from the land use data and the obtained distribution from the simulation outputs with ActivitySim.

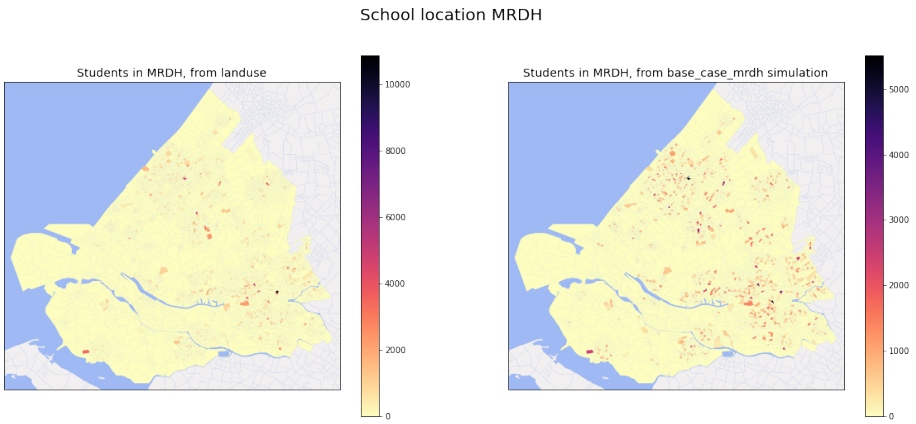


Figure 5.2: Students per MRDH zone from the land use data (left) and simulation results (right) for the entire region. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

In addition to the higher level distribution across the entire Metropolitan Region Rotterdam The Hague, a more detailed look into cities provide more insights on the reliability of the results. The comparison for distribution of students in the city of Rotterdam (including port area) can be found in figure 5.3, and figure 5.4 shows the same for The Hague.

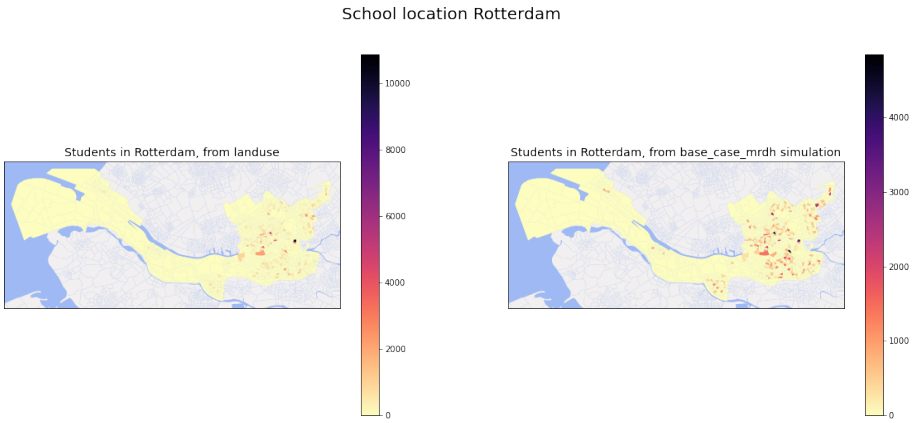


Figure 5.3: Students per MRDH zone from the land use data (left) and simulation results (right) for Rotterdam, including port area. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

For The Hague, the distribution seems to confirm that the simulation output is highly accurate, closely mimicking the distribution observed in the land use data with only some TAZs that have slight differences in the concentration of students they have. This, however, does not hold entirely true for Rotterdam or the entire Metropolitan Region Rotterdam The Hague. As it can be appreciated, the simulation output shows a distribution that is much less concentrated, as it can be seen from the slightly higher number of TAZs highlighted and the lower maximum number of students in a single TAZ. Other cities in the Metropolitan Region Rotterdam The Hague yield similar results to those shown for school locations.

Results seem to suggest some accuracy, partially confirming the validity of the high fitness value of the

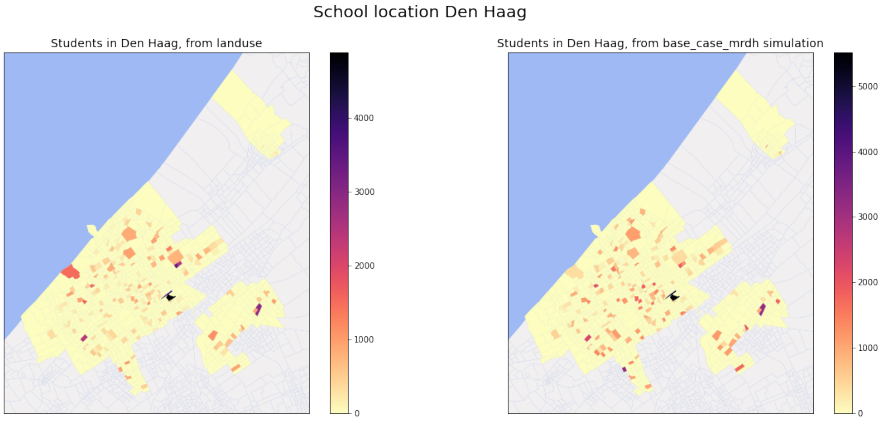


Figure 5.4: Students per MRDH zone from the land use data (left) and simulation results (right) for The Hague. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

choice model for school locations. Although, it is seen that the simulation tends to spread school location choices more than the comparatively concentrated reality, which can have an impact in subsequent outputs.

5.2.3. Workplace location accuracy

The same procedure was performed for worker location, and on figure 5.5 the geographical distribution of workplaces around the entire Metropolitan Region Rotterdam The Hague can be compared between the land use data and the simulation output from ActivitySim with the new parameters.

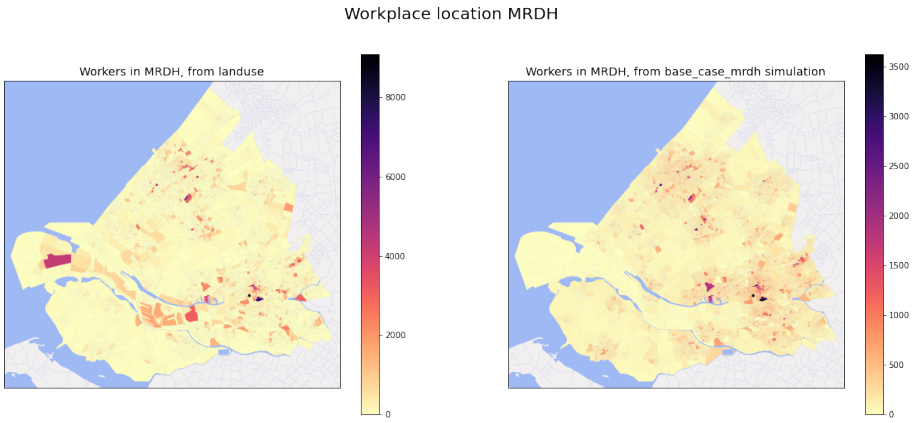


Figure 5.5: Workers per MRDH zone from the land use data (left) and simulation results (right) for the entire region. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

In a similar fashion, figure 5.6 shows the geographical distribution of workplaces for the city of Rotterdam for both the land use data and the simulation output from ActivitySim, and figure 5.7 shows the same for The Hague.

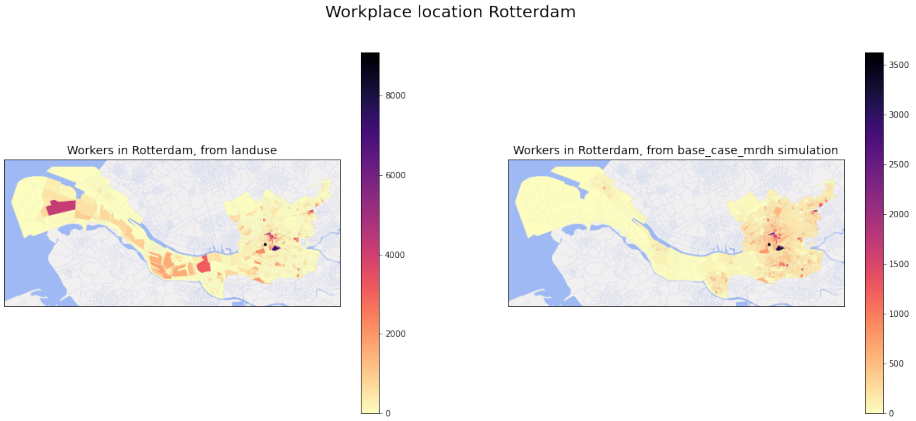


Figure 5.6: Workers per MRDH zone from the land use data (left) and simulation results (right) for Rotterdam, including port area. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

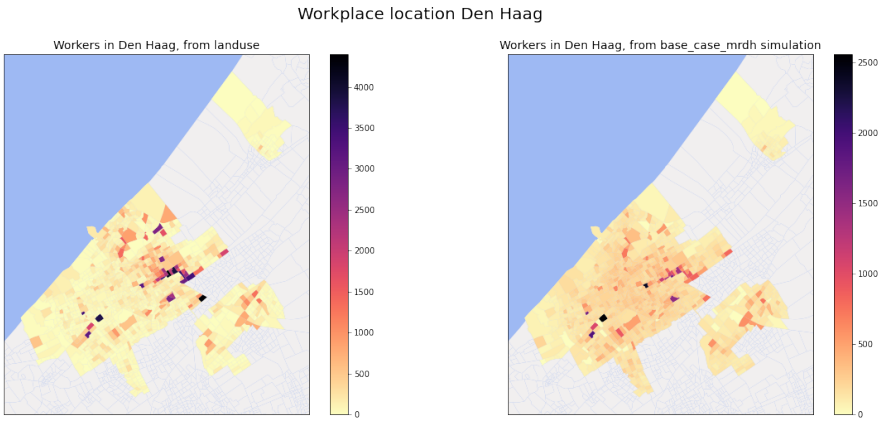


Figure 5.7: Workers per MRDH zone from the land use data (left) and simulation results (right) for The Hague. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

Once more, it can be seen that simulated output in this case tends to spread workers more than what the land use data suggests is the real situation, which can be seen by the higher number of TAZs highlighted and the lower maximum number of workers in a TAZ. Most cities in the region show the same situation.

One particularity in the case of workers is, however, the distribution of workers in the port area of Rotterdam in figure 5.6, where despite the simulation showing an overall more spread out distribution, it fails to assign workers to the port almost completely, making the results look very different in the simulation output from what is depicted in the land use data.

5.2.4. Comparison of travel distance

Given the discrepancies observed between the land use data and the aggregated outputs, it is valuable to look into the results in greater detail, especially disaggregated results. For this, the distances between home and the chosen locations for school and workplace were obtained per every person by obtaining the distances from the network data. Figure 5.8 shows the comparison of kernel density estimations (a smoothed curve that shows the distribution of travel distances, analogous to a histogram) for the distances to school and workplace locations from the survey data and the simulation output.

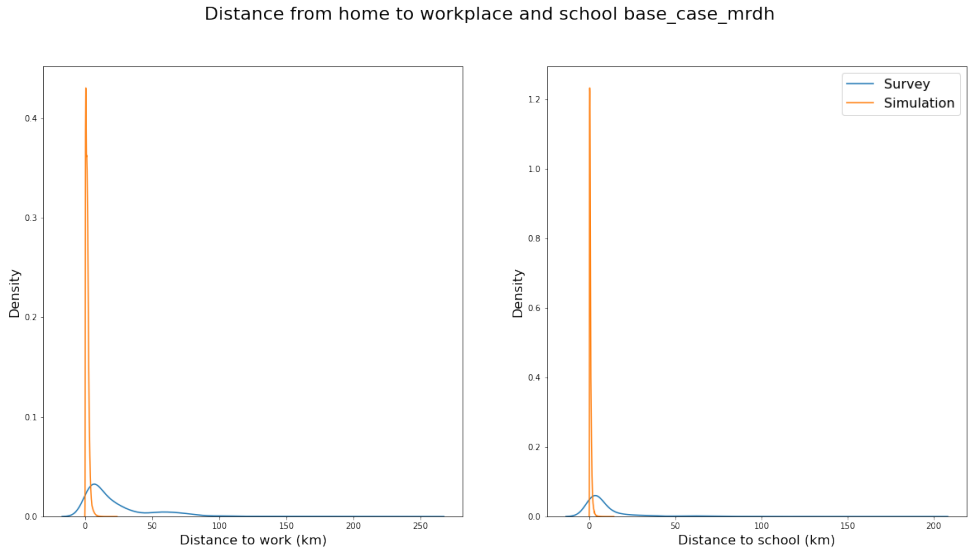


Figure 5.8: Distance kernel density estimation to work (left) and school (right) for survey data and simulation output.

It can be appreciated that the distances from the survey tend to be significantly larger than those obtained from the simulation in both cases, even when the aggregated results for school and workplace location looked promising. It appears that for both school and workplace location the model tends to place the output in areas very close to the home zone of the agent, where the closer match seen in school location is possibly due to there being fewer school location alternatives in the data and the model being forced to chose a distribution closer to the aggregated data. The spread of the travel distances is also appreciably smaller in the simulation than in the survey data, revealing a more homogeneous output than expected. This last issue can possibly also be due to inadequate sampling, where not enough, or not the right alternatives are considered.

There is an apparent discrepancy between the tendency of the model to predict locations closer to home as evidenced in the disaggregated output, and the higher spread of school and workplace locations in the aggregated output. Inadequate sampling could also be an explanation for this, meaning that while some alternatives are in reality much more likely to be chosen (the ones were there is a higher concentration in the land use area), these are left out of the sample and out of consideration for the modeled individual, forcing the model to chose a less likely alternative that is within the sample. This could also explain the lack of workplace locations assigned to the port of Rotterdam.

5.2.5. Usability of model

Given the high fitness values obtained from the model estimation process, it can be assumed that the configuration of the logit models for destination choice (where school and work destinations are determined in advance) and its parameters are appropriate. However, the output given by ActivitySim still has space for improvement. One possible area to look into is the sampling method used by ActivitySim by default, as this conceptualization might give too much weight to travel distance (cost) when sampling the alternatives and thus resulting in the observed distributions.

5.3. Sampling of destination choice set

An alternative sampling method for the destination choice set was defined inspired by Stratified Importance Sampling based on the works of Berjisian and Habibian (2019) and Tsoleridis et al. (2022). This method originally uses a fixed radius from origin zones to search for destination zones, and said radius is informed from the data to prevent bias. However, the remapping of PC4 zones to MRDH zones, which as previously described are more granular in the area of interest than in the in the area of influence, can produce distortions on the distances between zones because of the very different zone sizes. This made the use of a fixed radius to search for the sample questionable at best, and instead, a more flexible formulation was devised which obtains a search radius for every possible origin zone.

The new formulation consists on listing the origin (home) and destination pairs per tour purpose and using the maximum distance between origins and destinations to inform a search radius for that origin and purpose (for example, the maximum distance traveled from a given zone for work trips). For origin zones and purpose combinations not present originally in the survey data, the radius is obtained from the next available zone (backfilling), taking advantage of the fact that the radius obtained from that zone is likely not smaller than what would correspond (based on the fact that the zones tend to grow in size as explained in section 4.2.1).

The search radius is then finally used to find all the possible destinations that are within that distance, including the origin itself, and a uniform probability is assigned to every alternative within the radius, while alternatives that are further away are given a probability of zero. While the backfilling used for missing data could marginally increase runtimes by adding additional zones to the sample, this is not expected to have a negative impact on accuracy. The sampling for every origin is then saved to a file that can be read by ActivitySim when performing the simulation, saving the need to produce the sample during the run.

The ActivitySim source code was then modified to substitute the default sample with the newly obtained sample, by specifying the relevant file name from the model settings. A new simulation was then performed using the synthetic population with the new sampling method and the results for school and workplace location choice are presented here.

5.3.1. School location accuracy

The aggregated outputs for school location for the simulation with Stratified Importance Sampling are plotted for the entire Metropolitan Region Rotterdam The Hague in figure 5.9, for Rotterdam in figure 5.10, and for The Hague in figure 5.11, where the geographical distribution of school locations on the land use data and the simulation output can be seen. The output for other cities in the region is similar.

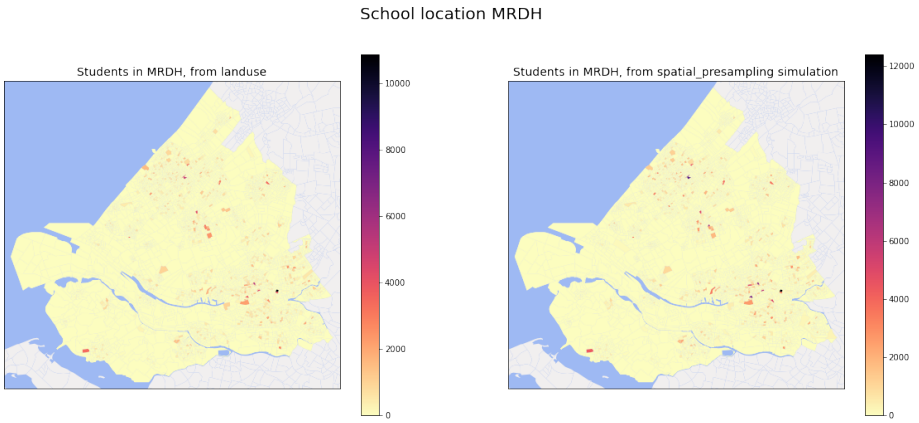


Figure 5.9: Students per MRDH zone from the land use data (left) and the new simulation results (right) for the entire region. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

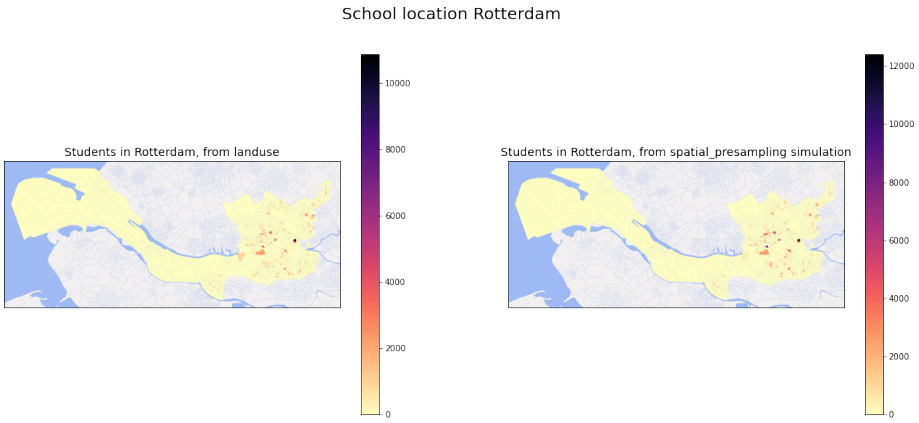


Figure 5.10: Students per MRDH zone from the land use data (left) and the new simulation results (right) for Rotterdam, including port area. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

With Stratified Importance Sampling, the geographical distribution of students obtained from the simulation is still very close to that of the land use data, and in fact, it seems to more closely match the concentration of locations seen from the land use, which is an improvement from the simulation with default sampling. However, the case of the Hague is worth highlighting, as it appears that the maximum number of students in a single TAZ is significantly higher, indicating that the concentration effect might be too strong with the alternative sampling method.

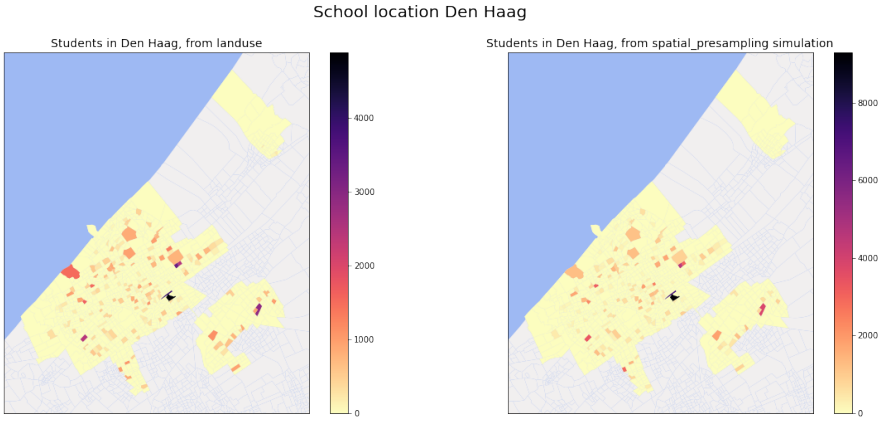


Figure 5.11: Students per MRDH zone from the land use data (left) and the new simulation results (right) for The Hague. The visualizations highlight where students are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

5.3.2. Workplace location accuracy

Likewise, the aggregated geographical distributions for workplace location for the land use data and the simulation with the Stratified Importance Sampling method are plotted for the entire region in figure 5.12, for Rotterdam in figure 5.13, and The Hague in figure 5.14; while the output for other cities in the region is similar.

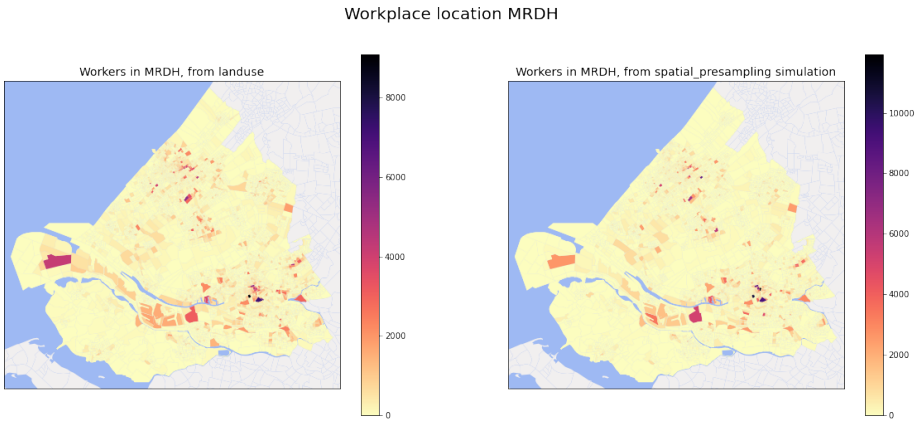


Figure 5.12: Workers per MRDH zone from the land use data (left) and the new simulation results (right) for the entire region. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

This time, it can be appreciated that with Stratified Importance Sampling, the aggregated distribution of workers much more closely matches that of the land use data when compared with the simulation with default sampling, allowing for workers to be assigned across a larger region in a more realistic way, which is specially noticeable along the port of Rotterdam, an improvement from the previous case. The situation is similar for other cities in the region.

Once more, however, it seems that the concentration effect is too strong in The Hague, where the

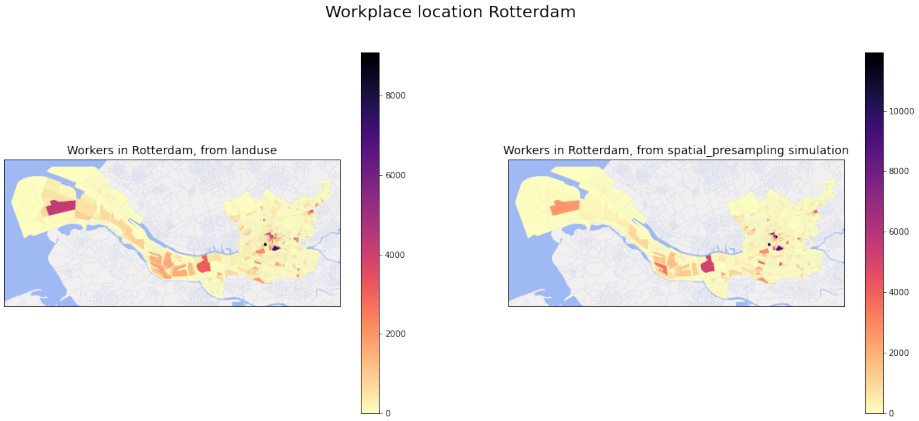


Figure 5.13: Workers per MRDH zone from the land use data (left) and the new simulation results (right) for Rotterdam, including port area. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

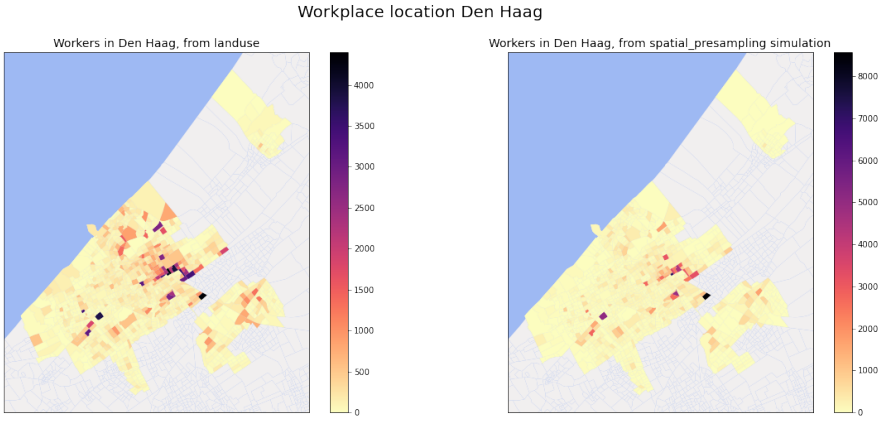


Figure 5.14: Workers per MRDH zone from the land use data (left) and the new simulation results (right) for The Hague. The visualizations highlight where workers are concentrated the most, and the legend to the right of each visualization gives an idea of the maximum number of students that are in a single TAZ.

workers are assigned on fewer TAZs on the simulation, and there is a higher number of workers in a single TAZ.

5.3.3. Comparison of travel distance

While the results for using Stratified Importance Sampling are promising, it can still be valuable to look into disaggregated results once again. Figure 5.15 shows the kernel density estimation for the distances to school and workplace compared between the survey data and the simulation output for the case with Stratified Importance Sampling.

The Stratified Importance Sampling method not only reproduces the land use data effectively, but it also matches the distances between home and destinations much more closely, which can be considered

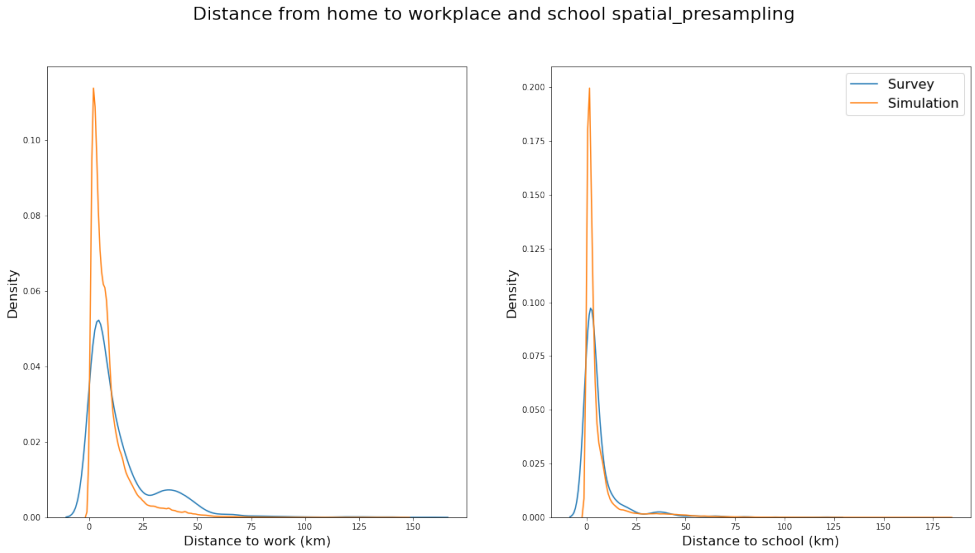


Figure 5.15: Distance kernel density estimation to work (left) and school (right) for survey data and simulation output.

an improvement on the output.

5.3.4. Usability of model

In this case it is highlighted that the accuracy of the output given by ActivitySim is influenced by the sampling method chosen for destination choice sets. The Stratified Importance Sampling method implemented here proves that it is possible to improve model output beyond what is offered by default in the ActivitySim framework, meaning that while the framework has enough built-in functionality to be advantageous to use as opposed to a new framework, it can still benefit from the customization enabled by its open source code, where improvements and extensions can be easily introduced.

The model run times were not significantly different using either sampling method, thus, accuracy is achieved without sacrificing performance.

5.4. Model framework

The ActivitySim framework for activity-based modeling has proven useful to create transport models in the Netherlands. The parameters for these models can be estimated from available survey data, and while processing such data is a time-consuming activity, the reusability of the procedure means that after the initial time investment, the model can be easily re-estimated with new data.

The choice model structure used in ActivitySim was able to achieve a high degree of accuracy for the Netherlands using the previously prepared data, yet the discrepancy in distances between home and school and workplace locations between the survey data and the simulated output leave room for improvement. The adaptation of the model to use Strategic Importance Sampling yielded more accurate results in this regard, highlighting the influence of choice set sampling methods on the accuracy of

model outputs. One observation is that the accuracy gains seem to diverge from city to city (or at least from the Hague to the rest of the region).

The flexibility of open source software like ActivitySim also means that the model can be modified to better adjust to the conditions of the Netherlands, such as the lack of joint tour data, and alternative choice set sampling methods.

The workflow previously detailed in figure 4.3 represents a good way to develop an activity-based model for the Netherlands, with the addition of obtaining the sampling from the survey data and using it as input for the simulation, obtaining results that are very accurate while being able to provide explanations from the behavior by looking into the parameters that define the logit models.

6

Conclusions and discussion

The research approach outlined on chapter 4 contemplated the development of a procedure to obtain efficient and accurate choice models to be used in activity-based transport models in the Netherlands, with the use of readily available data. Chapter 5 shows the results of developing and evaluating such a procedure, and in this chapter conclusions will be drawn from those result. This research will also be placed in the context of the existing scientific literature, its implications for society will be outlined, the limitations faced will be discussed, and lastly, recommendations for further implementation and research will be drawn.

6.1. Conclusions

It is possible to successfully process ODiN survey data into the format needed the ActivitySim open source software to develop activity-based models. The time intensiveness of the procedure is mitigated by the possibility of effortlessly reusing the process to re-estimate the model with data from future years, as the formatting, calculation and inference of the needed data fields can be easily automated with a script. The missing data from joint tours, however, warrant doing modifications on the source code of ActivitySim to bypass the impossibility of running joint tour submodels using only ODiN data. The implication of this is not only the need to perform additional work, but also that the model is not able to capture household interactions properly without using additional data.

Additionally, there is data in the ODiN survey that could be better exploited, such as electric vehicle ownership and participation in car sharing. It must be noted, however, that even if much of the lack of documentation is skirted by using this method, adding additional data to the model might result in needing to revert the modifications to the code, and to look again at the (in some instances inadequate) documentation of the software to properly format it.

The steps that then need to be performed to be able to use available data and software to develop

an activity-based model in the Netherlands, as performed in this research is to filter the data to the relevant time periods to be modeled, map origins and destinations to the zonal system used in the skim matrices and land use data, obtain workplace and school locations per respondents, identify tours and subtours, identify outbound and inbound trips, map household income to monetary values, segment the population (household, student, worker, and person types), classify tours and obtain coordinated daily activity patterns, group modes and assign tour mode, classify tours by purpose, calculate tour stop frequency and subtour frequency, classify tour schedules, and use canonical tour IDs.

The choice model structure used in ActivitySim for destination choice makes use of logit models, and in the case of school and workplace location uses multinomial logits to decide these destinations in the early stages of the simulation run and only once. The resulting tables from the data processing can then be used as input for estimation, and the explainatory variables in the utility functions in the logit models can be iteratively tested and its parameters estimated until a high fitness value is reached, such as the case with the parameters in tables 5.1 and 5.2.

In this regard we can conclude that using ODiN data and the ActivitySim software provides significant advantages such as establishing a straightforward procedure to re-estimate choice models in an activity-based model, allowing updates of the model with more recent data with ease, and being able to exploit the quality data from the survey than can be used to obtain outputs that can be interesting for sustainable mobility analyses, such as using electric car ownership data to account for electric cars in mode choice.

As disadvantages, however, this methodology cannot account for household interactions on its own, one of the main improvements promised by activity-based models compared to classic models, as the survey is missing this data. On a related note, inadequate documentation might complicate the addition of new data as well.

The models estimated with this method seem to provide acceptable accuracy, as evidenced by the fitness values obtained for the school and workplace location choice logit models. However, these fitness values do not serve as fail-proof measurements of model performance, as it was seen that good aggregated destination results can still give way to relatively inaccurate disaggregated outputs. The output can be influenced by the sampling of the large choice set (which is done to achieve decent run times).

Hence, it can be concluded that the accuracy of destination choice models estimated using ActivitySim and ODiN data can be disputed. While the models themselves suggest high accuracy, when running ActivitySim simulations some flaws are exposed, which skew simulation outputs.

Changing the default sampling method in ActivitySim for Stratified Importance Sampling based on the survey data, while not necessarily driving changes in run times, can positively affect the accuracy of the simulation outputs, as was tested in this research where more accurate results were achieved.

It can be concluded that Stratified Importance Sampling can be used to sample the choice set of destinations without compromising the accuracy of an activity-based model in the Netherlands.

This research shows that the development of an activity-based transport model for the Netherlands using widely available tools and data while maximizing explainatory value and minimizing run times

can be achieved by using the data from ODiN, a travel survey that is made available every year, and feeding it to ActivitySim, an open source software package developed in the United States that uses logit formulations to model choice. The data can be processed to the requirements of the software and be used to estimate the model parameters with ease, and using Stratified Importance Sampling on the destination choice set can provide results in reasonable run times with great accuracy.

It can be concluded that an activity-based model that uses travel survey data and open source software can achieve good accuracy and efficiency, especially after addressing the flaws in the sampling method; however, the impact of missing household data in accuracy remains unaccounted for.

6.1.1. Contribution to existing knowledge

The possibility to build an activity-based transport model using available data and open source tools regardless of the large data requirements further reinforces the idea of Eluru and Choudhury (2019), Han et al. (2021), Hörl and Balac (2021), and Knapen et al. (2021) that more easily replicable and adaptable models can be built with their use.

In this case, the use of ActivitySim, which has been explicitly developed with adaptability in mind, also proves the value of similar formulations. Even when it was developed with the context of the United States in mind, this was not a major barrier in conceiving a model for the Netherlands. While other open source platforms such as MatSim exist, the adaptability of ActivitySim and absent in MatSim proved a major advantage and displays is as a better candidate for applications of activity-based models.

The accuracy achieved by the logit model formulations and the easy model parameter estimation that can be achieved with them reinforces the ideas of Castiglione et al. (2015) and Ortúzar and Willumsen (2011), establishing that logit model formulations perform very well while being easy to implement.

The discrepancy in the simulation outputs when using the two different choice set sampling methods also reinforces the arguments from Berjisian and Habibian (2019), Leite Mariante et al. (2018), Lemp and Kockelman (2012), Pozsgay and Bhat (2001), and Tsoleridis et al. (2022) in which it is claimed that the choice of sampling method has an important impact in the accuracy of the outputs, possibly introducing bias.

A Stratified Importance Sampling method that makes use of available survey data, namely the travel distances from a destination per purpose, was implemented and proposed as an alternative sampling method and it achieved greater accuracy, building on the work of Berjisian and Habibian (2019), Leite Mariante et al. (2018), and Tsoleridis et al. (2022) and using this sampling method regardless of mode choice and on a very large data set. This also further maximizes the use of the available data, something that Shiftan and Ben-Akiva (2010) advocates for, while noticeably increasing output accuracy.

The methodology used for this activity-based model also seems to have less need for manual integration of different software products than the methodology used by Knapen et al. (2021) to develop an activity-based model, nor the time and resource intensiveness of purpose-built models that is mentioned by Castiglione et al. (2015) and that is the case for the model developed by Arentze et al. (2005).

6.2. Discussion

Beyond the scientific conclusions brought forward in section 6.1, the results obtained now must be placed into the larger context for them to be usable and productive.

The methodology proposed in this research to develop an activity-based model for the Netherlands is meant to be replicated with relative ease. The use of the widely available and free to use ActivitySim software, and the use of the ODiN data that is made available every year is done intentionally to lower the barrier and costs of developing these models.

Additionally, the development of the procedure itself to process the data, and for the entire modeling process, both made explicit in this research, is meant to serve as guidance for modelers attempting to employ activity-based models and needing to develop one, a task that can be hard to surmount if done from scratch. The procedure can be completely and quickly replicated by simply using the data for the relevant year, allowing modelers to direct their efforts towards other modeling decisions.

The use of logit choice models as done in ActivitySim also banks on its wide acceptance to provide a model that can be easily and quickly interpreted by transport modelers already familiar with them, hence avoiding the need to gain significant new knowledge just to be able to operate the model.

At the same time, the use of Stratified Importance Sampling to sample the destination choice set reduces the need for sampling criteria set by the modeler as is the case with other sampling methods, which reduces bias and the need for even more modeler input, while improving accuracy. This is of particular importance when one considers that destination choices are used in other submodels further in the simulation, for example, mode choice, where one could expect the choice to be affected by the distance to travel and the availability of certain modes between the origin and destination. This highlights the importance of having accurate destination choice models as a requisite to have accurate activity-based models.

While destination choice might not be by itself of particular relevance to evaluate sustainable mobility policies, its effect on other submodels warrants the focus that was given to it in this research. Greenhouse gasses emissions, for example, can be estimated based on travel distances, travel time, and mode choice, all of which can be expected to heavily rely on destination choice. Additionally, as destination choices are partly explained by land use (for example, the availability of employment), from formulating an accurate destination choice model an analyst can assess the impact of policies that rely on land development on destination choice, such as stimulating commercial development.

With the decisions made with this method, transport modelers can focus on formalizing policy alternatives and scenarios, while providing them with an easily understandable model that can then be easily modified to suit their needs, overcoming the complex model structure and lack of appropriate documentation in some regards. Likewise, the model is easy to maintain with this procedure, as it can be easily re-estimated with data for another year or after complementing with other data sources.

Additionally, as previously mentioned in section 5.1.1, the already available additional data in the ODiN survey, which can be used in the model, also facilitates the use of this framework to develop models that are sensitive to sustainable mobility policies, for example, by expanding the mode alternatives in

the model and using the extended data in the utility functions to re-estimate the model.

The possibility of getting some information regarding remote work was also mentioned in section 5.1.1, where it was described that when a person does not perform any traveling in the reporting day, it can potentially respond that it did not do so because of remote working. This, however, has one particular weakness in that such information is only requested when the respondent did not travel for any other purpose either, which means that if the respondent performed any non-mandatory travel like social activities or doing groceries, this information is not captured. Likewise, only one reason can be given, meaning that if the respondent has more than one reason for not traveling during the reported day it is forced to chose only one, losing data in the process.

Additionally, as remote work and study has gained relevance after the COVID-19 pandemic, it is worth noting that when a person does not perform any trips on the reporting day, a reason is given for it and it is possible to respond that the reason is to perform remote work or study. While this is valuable data, it was the weakness of not being asked if the respondent performed any kind of trip regardless of purpose.

While the methodology used in this research is made for the specific case of the Metropolitan Region Rotterdam The Hague in the Netherlands, it is not far fetched to think that a similar process can be followed with data from other parts of the world where a travel survey is available. Certainly, the quality and extensiveness of the data in the Netherlands cannot be found everywhere, yet, the fields that are needed to process the data into the format required by ActivitySim is likely present in other surveys, so that an activity-based model can be implemented with the help of ActivitySim. The sampling method is perhaps even more transferable, as the survey data needed to come up with the samples is even more basic, pertaining to trip origin, destination, and purpose.

One important point of discussion is the impossibility to properly capture household interactions without the inclusion of additional data. While by the definition of Castiglione et al. (2015) illustrated on figure 2.3 this is still an activity-based model, not capturing these interactions is certainly not ideal, as this can limit the sensitivity of the model to certain policies and reduce the value of certain outputs when using the formulation proposed in this research. For example, it is not possible to account for how households share a private vehicle, or in other words, how the travel of one member of a household using a vehicle is constrained by another member of the household also needing that vehicle at the same time, which could either prompt either of the travelers to choose another mode of transportation, to not travel altogether, or for joint travel to originate. This could have an impact (unmeasured in this research) on mode choice outputs, and thus on policies that rely on influencing mode choice.

Attention also needs to be placed in the case of predicting subtours, as it was not possible to obtain many of these from the data. It is unclear if not performing many subtours is a trend in the Netherlands, a limitation of the travel survey in capturing them, or a limitation of this method to properly define them from the data. In any case, this could affect the accuracy of the subtours submodel.

6.2.1. Societal implications

The possibility of developing and calibrating activity-based transport models with ease is something that can potentially have an impact on decision-making for mobility, especially in the context of sustainability.

On subsection 2.3 it was established that policies oriented towards sustainable mobility need to be evaluated with model results that are best produced by activity-based models, hence, by making these models more accessible we can also facilitate the availability of quality information for decision support and help reach the climate goals of the Netherlands and Europe.

Moreover, as the framework here proposed takes much of the burden of developing the model away, and is meant to be easily understood by modelers already familiar with more traditional transport models, it allows modelers to focus on other modeling choices, the formalization of policies in the model, and the creation of scenarios, which in turn hopefully leads to better quality and timelier results.

6.2.2. Limitations

Some limitations were found in the elaboration of this research. First, the scope of this research was not designed to test the possibility of developing activity-based models from other available data sources for network and zonal (land use) data. These data sources are already needed in less refined transport models, such as the so-called four-step model, and thus were not deemed of particular importance, however, the time it takes to process the data and its quality also has an impact on the development of activity-based models and must be considered when planning the work to do so.

Second, the lack of more detailed household data limits the possibility of properly developing a model that evaluates household interactions in the form of coordinated daily activity patterns, at least without additional data sources or even estimating such submodel separately. This is relevant as one of the advantages that proponents of activity-based models often quote is this possibility, which accounts for households scheduling their activities around one another, and also engaging in joint travel, a behavior that could be interesting to analyze for policy impacts, especially as policy-makers strive to reduce car ownership, possibly making the resource dependence within households more relevant for the model output. As previously described, a similar problem might be present in the case of subtours.

Third, the variations in the accuracy of the results between different cities is left largely unaddressed. This can be because people behave differently across these cities, but due to time constrains and the scoping of this research project there was not more effort put into evaluating this issue.

Fourth, the sequential logit models used to define the different choices in this formulation make use of shortcuts to model interdependence, either by making the interdependence entirely unidirectional (for example, a traveler first choosing whether or not to perform a non-mandatory tour to only then consider tour destination) or simplifying with the use of expected utilities (like the mode choice logsums in destination choice). In this regard, multistate supernetworks were conceived to model these choices simultaneously in a way that captured their interdependence, but their poor manageability in complex cases is something that hindered its application in this research.

Fifth, by pre-specifying level of service for different time windows as a static dataset that does not change during model runs means that the model is not sensitive to changes in traffic and congestion patterns that could be originated from the implementation of policies. One example where this might be an issue is with the evaluation of congestion pricing policies, where the model is able to capture when car trips are substituted with trips in other modes, but it is unable to capture the effect of car trips shifting to less congested times and affecting the travel times of other travelers.

Sixth, the availability of more recent data that represents travel patterns after the trend-breaking COVID-19 pandemic, or rather the lack of it, had an impact on what could be achieved with this formulation. While recent versions of ActivitySim include submodels to account for remote working, the most recent year for which survey data is available is 2020. The data for 2019 was used instead for reasons already explained in subsection 4.2.3, and since the data for 2021 still faced some of the anomalies in mobility due to government sanctioned restrictions, it seems that to fully test this new addition we will have to wait until at least 2022 data is made available. Even then, the modeling of such choice could face difficulties in that the ODiN survey was not designed to capture this phenomenon very well, for instance, respondents only report their activities for one day of the week, which limits the understanding of hybrid working patterns across the week for individuals, with little else to offer to complement this information. Additionally, as previously described, very limited information is captured about remote work, and big information loss might occur in this regard, making it unreliable if one were to use it to model remote work.

Lastly, time constrains did not allow for more thorough testing of the capabilities of the model when using Stratified Importance Sampling based on survey data. The explainatory value of this formulation needs further testing, as it is possible that the model is overspecified and made to represent survey data too closely, which can put into question the value of such a sampling method, as this would mean the model loses sensitivity to other inputs. The explicit definition of distance as a measure of importance might also be an instance of modeler bias and not proper representation of reality, that is, while the distance radius used to define the ”strata” is directly informed by the survey data and thus does not need the modeler criteria that other methods require, the choice of a measure of distance to represent importance could in itself be biased towards modeler beliefs.

6.2.3. Recommendations

In line with the previous limitations, possible users of this model formulation must account for the time it takes to process zonal and network data. Planning agencies and companies that already make use of other transport models may, however, be able to reuse the already processed data used in those.

Additionally, to lower the barrier of entry for these kinds of models, the documentation provided with them and with dedicated software packages needs to be improved. While this procedure deals with some of the poorly documented processes, it is still recommended to provide the documentation for them.

Users of this formulation also need to be wary of applying it for 2020 and 2021. Depending on the purpose of the model, it could still be useful to understand mobility patterns with it, but the impossibility to go certain destinations such as offices for work will need to be accounted for, perhaps by setting availability conditions in the logit model at the risk of incurring in overspecification.

The lack of validation with alternative data sources also warrants caution for users of this procedure, who should perform their own validation to test the performance of their models. Possible alternative data sources include but are not limited to mobile tracking data, census data, and more specialized travel surveys.

Lastly, users of the proposed sampling formulation also need to perform analysis on the simulation

outputs, especially sensitivity analyses for the parameters that define their policies of interest.

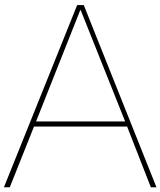
6.2.4. Further research

An initial point that deserves further research is the use of additional data to adequately model household interactions, since as previously discussed the ODiN survey is not sufficient in this regard. This would help paint a better picture of the suitability of this method to produce activity-based models, on top of producing more useful activity-based models.

Further research is also recommended to apply ActivitySim model formulations on 2020 and 2021 (pandemic restriction years) and from 2022 onward (new mobility trends such as remote work, yet hopefully, no more restrictions). The remote work extension for ActivitySim is of particular interest as it has already been developed but could not be tested in this research, however, the lack of dedicated data fields in the ODiN survey needs to be overcome.

Some knowledge gaps can also be identified for multistate supernetworks as described in section 3.2.3, however, the applied nature of this research rendered them out of scope, although they can provide an interesting avenues for future research. Namely, the efficient generation of the supernetworks, and the appropriate estimation method for its parameters are issues that need to be solved before multistate supernetworks make it into practice. Another avenue for research in the domain of multistate supernetworks could also be the modeling of more realistic multimodal travel. While other approaches require to come up with mode combinations upfront to be able to obtain the relevant level of service of the transport network between origins and destinations, multistate supernetworks could potentially model level of service in more detail and let the model choose mode combinations, only imposing restrains on unfeasible or unlikely combinations, like riding a private car after leaving home without it.

Likewise, further analyses and testing needs to be performed on the proposed Stratified Importance Sampling method, to determine if it introduces new biases or the risk of overspecification. Other sampling methods, especially those that rely on changing the logsum structure on whose basis ActivitySim already samples choice sets can be proposed.



Fields and descriptions from ODiN survey data

Table A.1: Fields available in the ODiN data set, with descriptions, identification of used fields in this research, and filters used if applicable.

Field name	Description	Used	Filter
op	New person	Yes	
opid	Unique ID for every person	Yes	
steekproef	Sample indicator	No	
mode	Response mode (all values are Computer Assisted Web Interview)	No	
hhpers	Number of people in household	Yes	
hhsam	Household composition	Yes	
hhplop	Place in household compared to hh nucleus	No	
hhlft1	Number of household members younger than 6 years	No	
hhlft2	Number of household members from 6 to 11 years old	No	
hhlft3	Number of household members from 12 to 17 years old	No	
hhlft4	Number of household members of 18 years or older	Yes	
wopc	Postcode Home Address	Yes	
wogem	Municipality of residence	No	
sted	Type of municipality of residence	Yes	
gemgr	Inhabitants of municipality of residence	No	
prov	Province of residence	No	

corop	Corop area of residence	No	
buurtadam	Neighborhood combination Amsterdam	No	
mra	Amsterdam metropolitan region	No	
mrdh	Rotterdam The Hague metropolitan region	No	
spl	Parkstad Limburg City Region	No	
utr	Utrecht Province	No	
geslacht	Gender	Yes	
leeftijd	Age	Yes	
kleeft	Age class	No	
herkomst	Migration background	Yes	No unknown values
betwerk	Paid work	Yes	No unknown values
onbbez	Unpaid activity	No	
maatspart	Social participation op	No	
opleiding	Educational attainment	Yes	No unknown values
hhbestinkg	Disposable household income (10% groups)	No	
hhgestinkg	Standardized disposable household income (10% groups)	Yes	
hhlaagink	Deviation of low income limit	No	
hhsocink	Deviation Social Minimum	No	
hhwelvg	Prosperity of the household (10% groups)	No	
hhrijbewijsau	Number of driving licenses in household	No	
hhrijbewijsmo	Number of motorcycle licenses in household	No	
hhrijbewijsbr	Number of moped licenses in households	No	
oprijbewijsau	Person owns car driver license	Yes	No unknown values
oprijbewijsmo	Person owns motorcycle driver license	No	
oprijbewijsbr	Person owns moped driver license	No	
hhauto	Number of passenger cars in household	Yes	No unknown values
hhautol	Number of passenger cars leased or in the name of a company in household	No	
opauto	Number of cars in person's name	No	
brandstofpa1	First fuel type of newest passenger car in household	No	
xbrandstofpa1	Second fuel type of newest passenger car in household	No	
brandstofepa1	Type of electric car newest passenger car in household	No	
bouwjaarpa1	Year of manufacture newest passenger car in household	No	
kbouwjaarpa1	Year (aggregated) of manufacture newest passenger car in household	No	
kgewichtpa1	Weight class newest passenger car in household	No	
tenaampa1	Registration of newest passenger car in household	No	
brandstofpa2	First fuel type of second newest passenger car in household	No	
xbrandstofpa2	Second fuel type of second newest passenger car in household	No	

brandstofepa2	Type of electric car second newest passenger car in household	No	
bouwjaarpa2	Year of manufacture second newest passenger car in household	No	
kbouwjaarpa2	Year (aggregated) of manufacture second newest passenger car in household	No	
kgewichtpa2	Weight class second newest passenger car in household	No	
tenaampa2	Registration of second newest passenger car in household	No	
brandstofpal	First fuel type of leased car or company car in household	No	
xbrandstofpal	Second fuel type of leased car or company passenger car in household	No	
brandstofepal	Type of electric car of leased car or company passenger car in household	No	
bouwjaarpal	Year of manufacture leased car or company passenger car in household	No	
kbouwjaarpal	Year (aggregated) of manufacture leased car or company passenger car in household	No	
kgewichtpal	Weight class leased car or company passenger car in household	No	
hhmotor	Number of engines in household	No	
opmotor	Number of engines in person's name	No	
hhbrom	Number of bromfiets in household (moped with max speed between 25 and 45 km/h)	No	
opbrom	Number of bromfiets in person's name	No	
hhsnor	Number of snorfiets in household (moped with max speed under 25 km/h)	No	
opsnor	Number of snorfiets in person's name	No	
hhefiets	Electric bicycle in household	No	
hhbezitvm	Household means of transport	No	
opbezitvm	Person means of transport	Yes	No unknown values
fqnefiets	Frequency use non-electric bicycle	Yes	No unknown values
fqefiets	Frequency Use electric bicycle	Yes	No unknown values
fqbtm	Frequency use bus	No	
fqtrein	Frequency use train	No	
fqautob	Frequency use car as driver	No	
fqautop	Frequency Use car as a passenger	No	
fqbrsnor	Frequency of use brom- and/or snorfiets	No	
ovstkaart	Possession of student ov-chipkaart	Yes	No unknown values
jaar	Reporting year	Yes	
maand	Reporting month	No	
week	Reporting week	No	
dag	Reporting day	No	

weekdag	Reporting day of the week	Yes	Only weekdays
feestdag	Reporting day is a holiday	No	
weggeweest	Person left home in the Netherlands	Yes	
redennw	Reason to not leave home	No	
redennwz	Not left: duration of illness	No	
redennww	Not left: type of weather	No	
redennwb	Not left: reason for stay abroad	No	
aantvpl	Number of regular trips in the Netherlands	No	
aantovvpl	Number of regular public transport trips in the Netherlands	No	
aantsvpl	Number of serial trips without professional with truck in the Netherlands	No	
efiets	Type of used electric bike	No	
autoeig	Used car in own name	No	
autohhl	Used car in the name of a member of the household	No	
autolwg	Used leased car from employer	No	
autolpl	Used leased car from private lease	No	
autobed	Used car in the name of a company	No	
autodorg	Used shared car from an organization	No	
autodpart	Used shared car from a private online platform	No	
autodbek	Used shared car with friends/aquaintances	No	
autoleen	Used loan car or borrowed car	No	
autohuur	Used rental car	No	
autoand	Used another type of car	No	
byzdag	Particularities on reporting day	No	
byzadr	Particularity: Other addresses	No	
byzvvm	Particularity: other means of transport	No	
byztyd	Particularity: other times	No	
byzduur	Particularity: other travel time	No	
byzroute	Particularity: other route	No	
byzreden	Reason other travel pattern	No	
reisduurop	Total travel time regular trips in the Netherlands (in minutes)	No	
afstandop	Total distance traveled on regular trips on the Netherlands (in hectometers)	No	
afstandsop	Total distance traveled on serial trips in the Netherlands (in hectometers)	No	
verpl	New trip	Yes	
verplid	Unique ID for every trip	Yes	
verplnr	Trip number	No	
toer	Departure point of trip is arrival point (tour)	No	
aantrit	Number of trip legs	No	
doel	Destination / purpose	Yes	No professional trips (i.e. driving a truck)
motiefv	Motive	No	

kmotiefv	Motive class	No	
meerwink	Several stores visited	No	
aardwerk	Nature work	No	
vertloc	Trip departure location	No	
vertgeb	Departure area	No	
vertpc	Postal code Departure point	Yes	No unknown values
vertpcbl	Postal code Departure point abroad	No	
vertgem	Departure municipality	No	
vertprov	Departure province	No	
vertcorop	Corop area Departure point	No	
vertmra	Amsterdam metropolitan region departure point	No	
vertmrdh	Rotterdam The Hague metropolitan region departure point	No	
vertspl	Parkstad Limburg city region departure point	No	
vertutr	Utrecht Province departure point	No	
aankgeb	Arrival area	No	
aankpc	Postcode Arrival point	Yes	No unknown values
aankpcbl	Postcode Arrival point abroad	No	
aankgem	Arrival municipality	No	
aankprov	Arrival province	No	
aankcorop	COROP area Arrival point	No	
aankmra	Amsterdam metropolitan region arrival point	No	
aankmrdh	Rotterdam The Hague metropolitan region arrival point	No	
aankspl	Parkstad Limburg city region arrival point	No	
aankutr	Utrecht Province arrival point	No	
pcg	Dutch post code border crossing	No	
gemg	Dutch municipality of border crossing	No	
pcblg	Foreign zip code border crossing	No	
afstv	Trip distance in the Netherlands (in hectometers)	No	
kafstv	Trip distance class in the Netherlands	No	
hvm	Main transport mode trip	Yes	
hvmrol	Role in the main transport mode	Yes	
khvm	Main transport class trip	No	
vertuur	Departure time trip	Yes	Only trips between 5 am and 10 pm
vertmin	Departure minute trip	No	
kverttijd	Departure time class	No	
aankuur	Arrival time trip	Yes	Only trips between 5 am and 10 pm
aankmin	Arrival minute trip	No	
reisduur	Travel time in the Netherlands (in minutes)	No	
kreisduur	Travel time class in the Netherlands	No	
actduur	Activity duration (in minutes)	No	
kind6	Child (ren) younger than 6	No	

volgwerk	Sequence of work trips	No
saantadr	Number of visited addresses Serial trips	No
sdezplts	All addresses in the same city	No
splaats1	City name 1	No
splaats2	City name 2	No
splaats3	City name 4	No
splaats4	City name 4	No
splaats5	City name 5	No
afsts	Distance serial trips in the Netherlands (in hectometers)	No
afstsbl	Distance serial trips abroad (in hectometers)	No
svvm1	First transport mode Serial trips	No
svvm2	Second transport mode Serial trips	No
svvm3	Third transport mode Serial trips	No
svvm4	Fourth transport mode Serial trips	No
sbeguur	Start hour series	No
sbegmin	Start minute series	No
seinduur	End hour series	No
seindmin	End minute series	No
corrverpl	Correction trips to legs	No
gehblver	Completely foreign trip	No
rit	New leg	No
ritid	Unique ID for every leg	No
ritnr	Leg number	No
afstr	Leg distance in the Netherlands (in hectometers)	No
afstrbl	Leg distance abroad (in hectometers)	No
kafstr	Leg distance class in the Netherlands	No
rvm	Leg transport mode	No
rvmrol	Role in the transport mode	No
raantin	Number of passengers in car	No
krvm	Leg transportation mode class	No
rvertuur	Leg departure hour	No
rvertmin	Leg departure minute	No
raankuur	Leg arrival hour	No
raankmin	Leg arrival minute	No
rreisduur	Leg travel time in the Netherlands (in minutes)	No
rreisduurbl	Leg travel time abroad (in minutes)	No
rvertstat	Train leg departure station	No
raankstat	Train leg arrival station	No
rtsamen	Group size Travel group train	No
rcorrsnelh	Correction due to speed	No
rvliegver	Airplane ride removed	No
factorh	Weighing factor household	No
factorp	Weighing factor person	No
factorv	Weighing factor trip	No

B

Sources for table fields in ODiN processing

Table B.1: Assumptions, choices, and definitions used for processing ODiN survey data into ActivitySim tables.

Table	Field	Source
Households	household_id	Same as person_id.
Households	home_zone_id	Mapped from PC4 to MRDH zone by using overlapping sections and choosing that with most residents.
Households	income	Corresponds with the median income of the income decile from which the respondent belongs, as obtained from CBS data.
Households	persons	Directly from ODiN (hhpers).
Households	hht	Mapped to ActivitySim requirements (obtained from Association of Metropolitan Planning Organizations Research Foundation (n.d.-a)) based on household residents and gender of the respondent.
Households	auto_ownership	Directly from ODiN (hhauto).
Households	urbanized	Directly from ODiN (sted).
Households	unemployed	1 if respondent is 18 years old or older and unemployed, otherwise 0.
Households	workers	All adults in household are workers, unless the respondent is known to be unemployed, in which case it is all adults - 1.
Persons	person_id	Directly from ODiN, but can be renumbered for convenience.
Persons	household_id	Same as person_id.

Persons	age	Directly from ODiN (leeftijd).
Persons	sex	Directly from ODiN (geslacht).
Persons	pemploy	Mapped to ActivitySim requirements (obtained from Association of Metropolitan Planning Organizations Research Foundation (n.d.-a)) based on age, and employment status. People who do not perform paid work but still perform work trips are assumed to work part time.
Persons	PNUM	ODiN only has one respondent per household, so this is always 1.
Persons	pstudent	Based on age and possession of student public transport card (Association of Metropolitan Planning Organizations Research Foundation, n.d.-a).
Persons	ptype	Based on pemploy, pstudent, and age (Association of Metropolitan Planning Organizations Research Foundation, n.d.-a).
Persons	education	Directly from ODiN (opleiding). Children under 12 years of age are assumed to still be in primary school, while older children have finished it.
Persons	driving_license	Directly from ODiN (oprijbewijsau).
Persons	roots_person	Directly from ODiN (herkomst)
Persons	has_car	Directly from ODiN (opbezitvm)
Persons	has_bike	Based on frequency of use of bike.
Persons	has_ebike	Based on frequency of use of ebike.
Persons	student_pt	Directly from ODiN (ovstkaart).
Persons	urbanized	Directly from ODiN (sted).
Persons	school_zone_id	Obtained by deducing from trip purposes and destinations, and then mapped to MRDH zones by using overlaps and choosing the area with the most students.
Persons	workplace_zone_id	Obtained by deducing from trip purposes and destinations, and then mapped to MRDH zones by using overlaps and choosing the area with the most jobs.
Persons	home_zone_id	Mapped from PC4 to MRDH zone by using overlapping sections and choosing that with most residents.
Persons	num_joint_tours	Equals zero for all persons, very little information included in the survey.
Override persons	cdap_activity	M if the person did mandatory trips, N if only non mandatory trips, H if person stayed home.
Override persons	mandatory_tour_frequency	Counts of school and work tours, mapped to strings read by ActivitySim.
Override persons	_escort	Counts of escort tours.
Override persons	_shopping	Counts of shopping tours.
Override persons	_othmaint	Counts of othmaint tours.

Override persons	_othdiscr	Counts of othdiscr tours.
Override persons	_eatout	Counts of eatout tours.
Override persons	_social	Counts of social tours.
Override persons	non_mandatory_tour_frequency	Based on _escort, _shopping, _othmaint, _othdiscr, _eatout, and social, coded according to ActivitySim definitions.
Trips	trip_id	Directly from ODiN, but can be renumbered for convenience.
Trips	person_id	Corresponds with persons table.
Trips	household_id	Corresponds with persons and households table.
Trips	tour_id	Corresponds with tours table.
Trips	outbound	True if the trip is before arriving to main tour destination, false otherwise.
Trips	purpose	Directly from ODiN (doel), ”touren/wandelen”, and ”ander doel” are mapped randomly to either eatout or social.
Trips	destination	Mapped to MRDH zones by using overlaps and choosing the area with the most relevant zone attribute depending on purpose.
Trips	origin	Mapped to MRDH zones by using overlaps and choosing the area with the most relevant zone attribute depending on purpose.
Trips	depart	Directly from ODiN (vertuur)
Trips	trip_mode	Directly from ODiN (hvm). Speedpedelec, bromfiets, snorfiets, skates and scooters mapped as ebike, handicapped vehicles mapped as walk.
Override trips	trip_num	Numbering of trips per (inbound or outbound) part of the corresponding tour.
Tours	tour_id	Canonical tour IDs as defined by ActivitySim, based on person_id, tour_type, and tour number. Obtained from source code. Tours are assumed to be home-based, subtours are work-based.
Tours	person_id	Corresponds with persons table.
Tours	household_id	Corresponds with households and persons tables.
Tours	tour_type	Assumed to be the first of the following list to appear as trip purpose in the tour: work, university, school, escort, shopping, othmaint, eatout, social, othdiscr, business, home. For subtours, the tour type is ”at work”.
Tours	tour_category	Mandatory if tour purpose is school or work, at work if tour is a subtour, and non mandatory otherwise.
Tours	destination	Destination for the trip corresponding to the main tour purpose.
Tours	origin	Departure point of first trip in tour.
Tours	start	Start time of first trip in tour.

Tours	end	End time of first trip in tour.
Tours	tour_mode	Mode of first trip in tour.
Tours	parent_tour_id	If subtour, then the ID of the parent tour, otherwise blank.
Override tours	stop_frequency	Counts of stops per (inbound and outbound) part of the tour, as a string readable by ActivitySim.
Override tours	tdd	Trip time of departure an duration, in hours, mapped to alternatives readable by ActivitySim.
Override tours	atwork_subtour_frequency	Count of subtours per subtour purpose, as a string readable by ActivitySim. Non work tours are blank.
Override tours	composition	Composition of joint tours. Since there are no joint tours, this is blank for all.

C

Public repository

For transparency and reproducibility purposes the scripts wrote during the elaboration of this research to prepare the data, the source code modifications, and the configuration files used for the model runs performed in this research have been uploaded to an online repository and made available to the public. To see the repository please go here or copy the url in your browser: https://github.com/davidmatheus002/activity_based_modeling.git

Bibliography

Alonso-González, M., Liu, T., Cats, O., Van Oort, N., & Hoogendoorn, S. (2018). The potential of demand-responsive transport as a complement to public transport: An assessment framework and an empirical evaluation. *Transportation Research Record*, *2672*(8), 879–889. https://doi.org/10.1177/0361198118790842

Arentze, T., & Timmermans, H. (2004a). Multistate supernetwork approach to modelling multi-activity, multimodal trip chains. *International Journal of Geographical Information Science*, *18*(7), 631–651. https://doi.org/10.1080/13658810410001701978

Arentze, T., Hofman, F., van Mourik, H., Timmermans, H., & Wets, G. (2000). Using decision tree induction systems for modeling space-time behavior. *Geographical Analysis*, *32*(4), 330–350. https://doi.org/10.1111/j.1538-4632.2000.tb00431.x

Arentze, T., Pelizaro, C., & Timmermans, H. (2005). Implementation of a model of dynamic activity-travel rescheduling decisions: An agent-based micro-simulation framework.

Arentze, T., & Timmermans, H. (2004b). A learning-based transportation oriented simulation system [Publisher: Pergamon]. *Transportation Research Part B: Methodological*, *38*(7), 613–633. https://doi.org/10.1016/j.trb.2002.10.001

Arentze, T., & Timmermans, H. (2007). Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Research Part B: Methodological*, *41*(7), 772–783. https://doi.org/10.1016/j.trb.2007.01.001

Association of Metropolitan Planning Organizations Research Foundation. (n.d.-a). *Examples — ActivitySim 1.0.4 documentation*. Retrieved July 25, 2022, from https://activitysim.github.io/activitysim/examples.html?highlight=hht

Association of Metropolitan Planning Organizations Research Foundation. (n.d.-b). *Software development — ActivitySim 1.0.4 documentation*. Retrieved January 28, 2022, from https://activitysim.github.io/activitysim/development.html

Aydin, N., Seker, S., & Özkan, B. (2022). Planning location of mobility hub for sustainable urban mobility. *Sustainable Cities and Society*, *81*, 103843. https://doi.org/10.1016/j.scs.2022.103843

Bao, Q., Kochan, B., Shen, Y., Creemers, L., Bellemans, T., Janssens, D., & Wets, G. (2018). Applying FEATHERS for travel demand analysis: Model considerations. *Applied Sciences*, *8*, 211. https://doi.org/10.3390/app8020211

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. MIT Press.

Berjisian, E., & Habibian, M. (2019). Developing a pedestrian destination choice model using the stratified importance sampling method. *Journal of Transport Geography*, *77*, 39–47. https://doi.org/10.1016/j.jtrangeo.2019.04.009

Bierlaire, M. (2018, December 19). *PandasBiogeme: A short introduction* (TRANSP-OR 181219). Ecole Polytechnique Fédérale de Lausanne. https://transp-or.epfl.ch/documents/technicalReports/Bier18.pdf

Bruzzone, F., Scorrano, M., & Nocera, S. (2020). The combination of e-bike-sharing and demand-responsive transport systems in rural areas: A case study of velenje. *Research in Transportation Business & Management*, 100570. https://doi.org/10.1016/j.rtbm.2020.100570

Castiglione, J., Bradley, M., & Gliebe, J. (2015). *Activity-based travel demand models: A primer* (S2-C46-RR-1). Transportation Research Board of the National Academies. Washington, D.C. https://doi.org/10.17226/22357

Centraal Bureau voor de Statistiek. (2019). *StatLine - Inkomen van huishoudens; inkomensklassen, huishoudenskenmerken*. Retrieved May 27, 2022, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83932NED/table?ts=1650379016879

Centraal Bureau voor de Statistiek. (2021). *Dutch national travel survey* [Statistics netherlands] [Last Modified: 13-09-2021T14:41:17]. Retrieved December 17, 2021, from https://www.cbs.nl/en-gb/onze-diensten/methods/surveys/korte-onderzoeksbeschrijvingen/dutch-national-travel-survey

Council of the EU. (2021, December 9). *Transport, telecommunications and energy council (transport), 9 december 2021*. Retrieved April 27, 2022, from https://www.consilium.europa.eu/en/meetings/tte/2021/12/09/

Daisy, N., Liu, L., & Millward, H. (2020). Trip chaining propensity and tour mode choice of out-of-home workers: Evidence from a mid-sized canadian city. *Transportation*, *47*(2), 763–792. https://doi.org/10.1007/s11116-018-9915-2

de Jong, G., & Kroes, E. (2008, April). Verbeterpunten in het landelijk model systeem. Retrieved February 25, 2022, from https://significance.nl/wp-content/uploads/2019/03/2008-GDJ-Verbeterpunten-in-het-Landelijk-Model-Systeem.pdf

Diana, M. (2012). Measuring the satisfaction of multimodal travelers for local transit services in different urban contexts. *Transportation Research Part A: Policy and Practice*, *46*(1), 1–11. https://doi.org/10.1016/j.tra.2011.09.018

Diana, M., Quadrifoglio, L., & Pronello, C. (2007). Emissions of demand responsive services as an alternative to conventional transit systems. *Transportation Research Part D: Transport and Environment*, *12*(3), 183–188. https://doi.org/10.1016/j.trd.2007.01.009

Dianat, L., Habib, K., & Miller, E. (2020). Modeling and forecasting daily non-work/school activity patterns in an activity-based model using skeleton schedule constraints. *Transportation Research Part A: Policy and Practice*, *133*, 337–352. https://doi.org/10.1016/j.tra.2020.01.017

Eluru, N., & Choudhury, C. (2019). Impact of shared and autonomous vehicles on travel behavior. *Transportation*, *46*(6), 1971–1974. https://doi.org/10.1007/s11116-019-10063-1

Esztergár-Kiss, D., & Lopez Lizarraga, J. (2021). Exploring user requirements and service features of e-micromobility in five european cities. *Case Studies on Transport Policy*, *9*(4), 1531–1541. https://doi.org/10.1016/j.cstp.2021.08.003

ETH Zürich. (2016, August 10). *The multi-agent transport simulation MATSim* (ETH Zürich, A. Horni, K. Nagel, TU Berlin, & K. W. Axhausen, Eds.). Ubiquity Press. https://doi.org/10.5334/baw

European Commission. (2016, July 20). A european strategy for low-emission mobility. Retrieved December 2, 2021, from https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52016DC0501

European Commission. (2019). The european green deal. Retrieved May 17, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1576150542719&uri=COM%3A2019%3A640%3AFIN

European Commission. (2020). Sustainable and smart mobility strategy – putting european transport on track for the future. Retrieved May 17, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0789

European Commission. (2022). *Sustainable cities and communities* [International partnerships - european commission]. Retrieved May 17, 2022, from https://ec.europa.eu/international-partnerships/sdg/sustainable-cities-and-communities_en

Fu, X., & Lam, W. H. K. (2014). A network equilibrium approach for modelling activity-travel pattern scheduling problems in multi-modal transit networks with uncertainty. *Transportation*, *41*(1), 37–55. https://doi.org/10.1007/s11116-013-9470-9

Gallo, M., & Marinelli, M. (2020). Sustainable mobility: A review of possible actions and policies. *Sustainability*, *12*(18), 7499. https://doi.org/10.3390/su12187499

Goudappel. (2022). *Traffic modelling software OmniTRANS expert*. Retrieved May 18, 2022, from https://www.goudappel.nl/en/expertise/data-and-it-solutions/traffic-modelling-software-OmniTRANS-Expert

Hafezi, H., Liu, L., & Millward, H. (2019). A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation*, *46*. https://doi.org/10.1007/s11116-017-9840-9

Han, Y., Chen, C., Peng, Z.-R., & Mozumder, P. (2021). Evaluating impacts of coastal flooding on the transportation system using an activity-based travel demand model: A case study in miami-dade county, FL. *Transportation*. https://doi.org/10.1007/s11116-021-10172-w

Hasnine, M. S., & Habib, K. N. (2018). What about the dynamics in daily travel mode choices? a dynamic discrete choice approach for tour-based mode choice modelling. *Transport Policy*, *71*, 70–80. https://doi.org/10.1016/j.tranpol.2018.07.011

Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for paris and île-de-france based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, *130*, 103291. https://doi.org/10.1016/j.trc.2021.103291

Infrastructure Victoria. (2017, December 14). *Melbourne activity based model video - infrastructure victoria*. Retrieved January 17, 2022, from https://www.youtube.com/watch?v=_P9dtgLP4h8

Katoshevski-Cavari, R., Arentze, T. A., & Timmermans, H. J. P. (2011). Sustainable city-plan based on planning algorithm, planners' heuristics and transportation aspects. *Procedia - Social and Behavioral Sciences*, *20*, 131–139. https://doi.org/10.1016/j.sbspro.2011.08.018

Kitamura, R., Kitamura,
bibinitperiod R., & Fujii, S. (1997). Two computational process models of activity-travel behavior. *Theoretical Foundations of Travel Choice Modeling*, 251–279.

Knapen, L., Adnan, M., Kochan, B., Bellemans, T., van der Tuin, M., Zhou, H., & Snelder, M. (2021). An activity based integrated approach to model impacts of parking, hubs and new mobility concepts. *Procedia Computer Science*, *184*, 428–437. https://doi.org/10.1016/j.procs.2021.03.054

Lah, O. (2019, January 1). Chapter 7 - sustainable urban mobility in action. In O. Lah (Ed.), *Sustainable urban mobility pathways* (pp. 133–282). Elsevier. https://doi.org/10.1016/B978-0-12-814897-6.00007-7

Leite Mariante, G., Ma, T.-Y., & Van Acker, V. (2018). Modeling discretionary activity location choice using detour factors and sampling of alternatives for mixed logit models. *Journal of Transport Geography*, *72*, 151–165. https://doi.org/10.1016/j.jtrangeo.2018.09.003

Lemp, J., & Kockelman, K. (2012). Strategic sampling for large choice sets in estimation and application. *Transportation Research Part A: Policy and Practice*, *46*(3), 602–613. https://doi.org/10.1016/j.tra.2011.11.004

Liao, F. (2016). Modeling duration choice in space–time multi-state supernetworks for individual activity-travel scheduling. *Transportation Research Part C: Emerging Technologies*, *69*, 16–35. https://doi.org/10.1016/j.trc.2016.05.011

Liao, F., Arentze, T., & Timmermans, H. (2010). Supernetwork approach for multimodal and multiactivity travel planning. *Transportation Research Record*, *2175*(1), 38–46. https://doi.org/10.3141/2175-05

Liao, F., Arentze, T., & Timmermans, H. (2014). Multi-state supernetworks: Recent progress and prospects. *Journal of Traffic and Transportation Engineering (English Edition)*, *1*(1), 13–27. https://doi.org/10.1016/S2095-7564(15)30085-4

López Díaz, M., Heinrichs, M., & Beige, S. (2020). Comparison of non-fixed and fixed locations for subsistence activities in destination choice. *Procedia Computer Science*, *170*, 714–719. https://doi.org/10.1016/j.procs.2020.03.166

McFadden, D. (1977). Modelling the choice of residential location. *undefined*. Retrieved February 25, 2022, from https://www-semanticscholar-org.tudelft.idm.oclc.org/paper/Modelling-the-Choice-of-Residential-Location-McFadden/55a63c2a72325a86de9a17814fb6243c132ac19a

Melkonyan, A., Gruchmann, T., Lohmar, F., & Bleischwitz, R. (2022). Decision support for sustainable urban mobility: A case study of the rhine-ruhr area. *Sustainable Cities and Society*, *80*, 103806. https://doi.org/10.1016/j.scs.2022.103806

Metropoolregio Rotterdam Den Haag. (2021). *The power of partnership* [MRDH Metropoolregio Rotterdam Den Haag]. Retrieved May 3, 2022, from https://mrdh.nl/power-partnership

Nerella, S., & Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models [Publisher: SAGE Publications Inc]. *Transportation Research Record*, *1894*(1), 11–19. https://doi.org/10.3141/1894-02

Newman, J. (2021). *Larch documentation — v5.7.0* [Larch]. Retrieved April 28, 2022, from https://larch.newman.me/v5.7.0/intro.html

OpenStreetMap. (2021). *OpenStreetMap* [OpenStreetMap]. Retrieved December 17, 2021, from https://www.openstreetmap.org/

Ortúzar, J. d. D., & Willumsen, L. G. (2011, May 3). *Modelling transport*. John Wiley & Sons.

Outwater, M., & Charlton, B. (2008). The san francisco model in practice: Validation, testing, and application [ISSN: 1073-1652 Issue: 42]. *Transportation Research Board Conference Proceedings*, *2*. Retrieved January 28, 2022, from https://trid.trb.org/view/883861

Patyal, V. S., Kumar, R., & Kushwah, S. (2021). Modeling barriers to the adoption of electric vehicles: An indian perspective. *Energy*, *237*, 121554. https://doi.org/10.1016/j.energy.2021.121554

Philip, M., Sreelathe, T., & Soosan, G. (2013). Activity based travel behavioural study and mode choice modelling, 10.

Popuri, Y., Ben-Akiva, M., & Proussaloglou, K. (2008). Time-of-day modeling in a tour-based context: Tel aviv experience [Publisher: SAGE Publications Inc]. *Transportation Research Record*, *2076*(1), 88–96. https://doi.org/10.3141/2076-10

Pozsgay, M., & Bhat, C. (2001). Destination choice modeling for home-based recreational trips: Analysis and implications for land use, transportation, and air quality planning. *Transportation Research Record*, (1777), 47–54. https://doi.org/10.3141/1777-05

Rijkswaterstaat. (2018). *Rijkswaterstaat data register*. Retrieved December 17, 2021, from https://maps.rijkswaterstaat.nl/dataregister/srv/dut/catalog.search#/metadata/e09ddb5d-1f69-414f-84ba-852340d31c0f

Sharmeen, F., & Meurs, H. (2019). The governance of demand-responsive transit systems—a multi-level perspective. In M. Finger & M. Audouin (Eds.), *The governance of smart transportation systems: Towards new organizational structures for the development of shared, automated, electric and integrated mobility* (pp. 207–227). Springer International Publishing. https://doi.org/10.1007/978-3-319-96526-0_11

Shiftan, Y., & Ben-Akiva, M. (2010). A practical policy-sensitive, activity-based, travel-demand model. *The Annals of Regional Science*, *47*(3), 517–541. https://doi.org/10.1007/s00168-010-0393-5

Stam, B., van Oort, N., van Strijp-Harms, H., van der Spek, S., & Hoogendoorn, S. (2021). Travellers' preferences towards existing and emerging means of first/last mile transport: A case study for the almere centrum railway station in the netherlands. *European Transport Research Review*, *13*(1), 56. https://doi.org/10.1186/s12544-021-00514-1

Tajaddini, A., Rose, G., Kockelman, K. M., & Vu, H. (2020, September 17). *Recent progress in activity-based travel demand modeling: Rising data and applicability* [Publication Title: Models and Technologies for Smart, Sustainable and Safe Transportation Systems]. IntechOpen. https://doi.org/10.5772/intechopen.93827

Tsoleridis, P., Choudhury, C., & Hess, S. (2022). Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities. *Journal of Choice Modelling*, *42*, 100336. https://doi.org/10.1016/j.jocm.2021.100336

United Nations. (2015, December 12). Paris agreement. Retrieved May 17, 2022, from https://unfccc.int/sites/default/files/english_paris_agreement.pdf

Vo, K., Lam, W., & Li, Z.-C. (2021). A mixed-equilibrium model of individual and household activity–travel choices in multimodal transportation networks. *Transportation Research Part C: Emerging Technologies*, *131*, 103337. https://doi.org/10.1016/j.trc.2021.103337

Vo, K. D., Lam, W. H. K., Chen, A., & Shao, H. (2020). A household optimum utility approach for modeling joint activity-travel choices in congested road networks. *Transportation Research Part B: Methodological*, *134*, 93–125. https://doi.org/10.1016/j.trb.2020.02.007

Zephyr Transport. (2020, September 14). *Introduction to ActivitySim*. Retrieved April 14, 2022, from https://www.youtube.com/watch?v=rWzQQSQHB6c

Zondag, B., & van Grol, R. (2021, October). *Toelichtende notitie LMS/NRM. thema's review* (Notitie No. 21034). Significance quantitative research. Den Haag.