Utrecht University

EXPANSE

MSc thesis in Artificial Intelligence
Utrecht University - Faculty of Geosciences

Evaluation and comparison of calibration techniques for urban mobility behaviour ABM

First supervisor:
Prof. dr. Mehdi Dastani

Second supervisor:
Dr. Simon Scheider

Daily supervisor:
PhD Candidate Tabea Sonnenschein

Author:
Marco Pellegrino
6997619
m.pellegrino2@students.uu.nl

22/03/2022
Words: 16936

Abstract

In the field of agent-based modelling (ABM), new studies focus on simulating urban mobility human behaviours to analyse behaviour-environment interactions. Models need to be calibrated to represent human behaviour correctly. The literature identifies challenges in calibration since there are several approaches to calibration, but primarily dependent on the model purpose.

This study addresses the lack of understanding of what makes a calibration method suitable specifically for mobility behaviour ABM. A data-driven approach is used, calibrating the proof-of-concept model of the EXPANSE project. An Amsterdam case study was selected with the ODiN dataset.

An appropriate experimental calibration framework is presented, analysing model characteristics delivering an objective function and hierarchical level in the field of human mobility.

The work also compared cutting-edge parameters search optimisation algorithms. Results hinted at surrogate model-based methodologies. Their performances outperform the other solutions on both fitness and computational load aspects. In addition, a metrics-based framework of how to compare calibration techniques for human mobility ABM is presented.

Finally, the investigation also proved that a parameter dimensionality reduction method based on grouping does not bring any benefit compared to calibrating the entire parameter set.

This project is just the first step in this field. The results are valuable, not only to the transport choice field only but generic to the main idea of mode choice. The established framework will be a profitable tool for all the researchers to be used in urban developments for a sustainable and healthier future.

Acknowledgements

The thesis before you is the final result of a research project for the MSc in Artificial Intelligence. Eleven months of working from home and at the university libraries, through a long pandemic and difficult moments. Nonetheless, this challenging experience allowed me to discover an interdisciplinary field new to me and become passionate about it.

I would first like to thank my thesis supervisor Prof. dr. Mehdi Dastani and Dr. Simon Scheider for their feedback and suggestions. The several meetings and their expertise were invaluable to structure the project and formulate the research questions, keeping me focused on the correct aspects.

Secondly, but not less, I want to thank my daily supervisor PhD candidate Tabea Sonnenschein for guiding me through this journey. Her constant availability and expertise brought me light not only to unclarities but also to constructive discussions. It was a pleasure to cooperate and share different background knowledge, keeping me motivated during the several struggles.

I also want to thank my friend, neighbour and AI colleague Daniel Manns for the never-ending explanations of AI topics. His continuous availability and support were keys to getting out of the arduous times. In addition, I thank my friends and colleagues with which I spent days working at the library.

I cannot forget to thank my parents for allowing me to attend this master's and follow my passions. Although the distance divides us, their support was crucial to alleviate me. A particular thought goes to my father for his difficult health period.

These were the people that played a primary role in this project. Nevertheless, every single person who gave me a smile and support, even for a second in these eleven months, deserves a special thank you.

Table of contents

1 Background: Exposome Science and the EXPANSE project

The last century was marked by an exponential growth in urbanisation (UN Department of Economic & Social Affairs, 2018), determined by an increase in population and the amount of industrialization. As a consequence of uncontrolled urbanisation, side effects such as environmental degradation, worsening of water quality and air pollution are frequent (Uttara et al., 2012). Among all the side-effects, cardio-metabolic pulmonary diseases (CMPDs) are stated by the World Health Organization (2014) to cause approximately one-third of all the global premature deaths.

Traditional genetic models always tried to tackle the issue, but they are limited to describe only 10% of the diseases. As a solution, Exposome Science was created (Wild, 2005). It focuses on the effects of behaviour-environment interactions in terms of diseases that diminish our life expectancy, such as the mentioned CMPDs. Hence, the aim is to maximize one's health in a modern urban environment.

These studies are becoming fundamental to policymakers in urban development and interventions (Corburn, 2007). Hence, taking into account not only the health scale but also the social side can return valuable solutions for all the involved stakeholders. In fact, as Sonnenschein et al. (2021) underline, the complexity of Urban Exposome can be observed at three dimensions: simulating and predicting human behaviour-environment interactions (Batty, 2013), the exposures generated from the latter (Tonne et al., 2017) and the social reaction of interventions (van Ham et al., 2012). However, the authors point out as such a combined method that includes all these perspectives is still lacking.

Artificial intelligence can play a key role in filling this knowledge gap. Agent-based modelling (ABM) has the ability to simulate and predict new urban scenarios, with the benefit of not implementing them beforehand (Crooks et al., 2018). In particular, it can be used to analyse how changes, for example in the transport system of a city, influence the environment, people's behaviour and their exposure interaction (Maggi & Vallino, 2016).

The work of this thesis is part of a bigger project, namely EXPANSE (https://expanseproject.eu/). It is structured in seven Work Packages, from "measuring the External Exposome" to "incorporate user stakeholders perspectives and develop the Exposome Hub". This thesis takes place in work package 4 "evaluate intervention strategies to prevent CMPD" (Cardio-Metabolic Pulmonary Diseases), in particular in the sub-work package 4.3 "agent-based modelling of urban interventions to prevent CMPD". The aim is to model urban health interventions, combined with Social Cost Benefit Analysis (SCBA), to support policymakers in their decisions.

2 Literature review

2.1 ABM and choice modeling

Agent-based modelling (ABM) is a cutting-edge simulation method in which individuals can make autonomous decisions. It is characterised by agents that can interact which each other and with the surrounding environment (Crooks et al., 2018). Till the last decade, human-nature systems were used to be studied independently, often constraining agents influences (An, 2012). By contrast, ABM permits to take into account interaction, generating a highly realistic simulation. ABM is capable to model individual decision making of agents and their contextual social organizations (Bousquet & Le Page, 2004), but still maintaining the interactions on a large scale (Gimblett, 2002). Examples of this technology have been successfully applied to a wider range of complex situations, such as supply chains or consumer behaviour (Macal and North, 2010; An, 2012).

ABM is highly related to the "choice modelling" field of research. The latter aims at formalising the decision process of agents in a particular scenario (Samuelson, 1948). Choice modelling is relevant because it allows both the comparison of the influence of policies on agent behaviour and the understanding of such a complex system of theories of human decision making (O'Sullivan, 2004; Gilbert and Troitzsch, 2005). ABM perfectly suits these tasks because it permits to study the non-linearity of aggregated behaviours regarding individual ones (Maggi & Vallino, 2016). In fact, ABM has the ability to encapsulate decision making, allowing for dynamic behaviour, adaptation, and even learning (Heckbert, 2010).

The first models were originally developed in econometrics or sociometric to analyse and predict human decisions (Nicholson & Snyder, 2005). In the later years, by integrating notions from behavioural psychological theories, ABM became able to study the effect of individual decision making on larger contexts and more complex applications. For example, de Mooij et al. (2021) developed a Belief-Desire-Intention (BDI) ABM to simulate the impact of enforced norms on the spread of COVID-19 in the US state of Virginia. In particular, agent choices are influenced by beliefs, objectives, trust in government, and the norms imposed by the government. Comparing a set of norms, it was possible to evaluate their effect to support future decision makings. As seen, several aspects can influence the final behaviour of an agent and consequentially the overall system behaviour.

In order to incorporate and calibrate findings from behavioural theories, empirical data such as surveys is used to provide distribution of characteristics, beliefs and preferences of individual or household behaviour (Baynes & Heckbert, 2009). Observed data for choice modelling can be achieved via stated preferences or revealed preferences. The former estimates behaviours from individual preferences, collected in previous experimental setups (Louviere et al., 2000). The second one, instead, allows inferring individual preferences from their behaviours (Chambers & Echenique, 2016). It relies on the economic-based assumption that agents take decisions to maximise their utility (Samuelson, 1938). Revealed preferences are usually adopted in modelling because of more accessible empirical data availability. Indeed, the majority of existing datasets describe human behaviour, instead of individual preferences.

2.2 Human mobility behaviour

One of the main ABM challenges is to model the human mobility behaviour (Maggi & Vallino, 2016), which is the subject of this work. Human mobility behaviour is mostly defined with the problem of the transport mode choice (Maggi & Vallino, 2016). In other words, the aim is to simulate which mean of transport a specific individual uses for a particular trip of the day. As the authors explain, human mobility behaviour has a twofold importance. It allows us to simulate human behaviour, but also to compare scenarios and behaviours to be further used as bases for policymakers in urban sustainable planning (Lord et al., 2015). In particular, most of the present research aims at showing how to limit private cars' use, while promoting sustainable modes of transport.

Unfortunately, "urban transport systems have all the characteristics of complex systems" (Chapizanis et al., 2021), such as non-linearity and heterogeneous individuals. The main origin of non-linearity is our adaptive behaviour, namely "social contagion". As defined in Lindzey and Aronson (1985), "the spread of affect or behaviour from one crowd participant to another; one person serves as the stimulus for the imitative actions of another". We have clear examples of how this behaviour is present in human mobility. For instance, the more people see other individuals biking or using a train, the more they are prone to use the same observed mean of transport. Looking at infrastructure changes, the impact will be dynamic and in the long term, because people will take time to adapt and "spread" the new commuting mode.

Secondly, non-linearity in human mobility behaviour is also influenced by the environment. For instance, as illustrated by Sonnenschein et al. (2021), the construction of a bike infrastructure can have a different modal choice impact based on the already existing modal choice.

Finally, human behaviour is heterogenous between different subgroups of the population. There is a high number of heterogeneous individuals, with different goals and needs, together with a dense degree of interaction between the agents and the environment (Chapizanis et al., 2021). More importantly, heterogeneity is also spatially characterised. As Sonnenschein et al. (2021) note, an urban intervention can have a different impact if, for instance, implemented in a city where multiple agents and activities are involved, compared to a less densely populated area.

2.3 ABM and GIS to simulate the human mobility behaviour

GIS is a conceptualized framework that captures, analyzes and displays spatial and geographic referenced information (Maliene et al., 2011).

ABM and GIS have the strength, together, to address the previous human mobility behaviour challenges, namely decision making, non-linear social contagion and heterogeneity both in behaviour and spatial dimensions.

Firstly, as previously presented, ABM is composed of individuals that interact with each other and the surroundings as humans do in a city. By defining a set of behavioural rules, goals and needs, ABM addresses the problem of social contagion.

Secondly, because mobility is highly influenced by the spatial dimension, the GIS support is gaining more interest in the last years (Maggi & Vallino, 2016). As Brown et al. (2005) explain, ABM are excellent at achieving temporal representations of agents and processes, while GIS can fill the gap of spatial data representations (fields, objects and functions). Previous research used regular lattice structures (Wu, 1998), which do not correctly model complex scenarios such as cities (Maggi & Vallino, 2016). With GIS, it is possible to add geographical layers such as a housing layer, a road network layer, or a population layer (Najlis & North, 2004).

Finally, ABM and GIS can simulate a heterogeneous population. ABM alone can represent the different individual behaviours and the agents' interactions. GIS adds the possibility to locate the various agents in the environment, based on observed census data containing spatial information. This is where most of the precedent models failed, assuming homogeneous populations as a simplification.

2.4 Previous studies of simulating human mobility behaviour

First attempts of simulating human mobility have been already carried out with partial success. For example, Balać and Hörl (2021) simulated San Francisco intermodal shared mobility in the MATSim platform (Horni et al., 2016). They aimed at analysing the effect of docked bike-sharing services as a last-mile solution in trips. Both human and engine-powered vehicles, docked/station and dockless/free-floating services, and multiple operators are represented. With the latter, it is possible to state that intermodality can significantly reduce travel time in comparison to current commuting times.

Another study is brought by Inturri et al. (2019) on the case study of the Italian city of Ragusa. Their aim is related to DRST (Demand Responsive Shared Transport) services, which analyses the trade-off between efficiency, service quality and sustainability. Overall, their focus is more on vehicle transport planning. The ABM, developed in NetLogo (Wilensky, 1999), was coupled to a GIS extension to represent the actual road network and map the distribution of population at the census zone level. In this case, the behavioural choice depends on two variables, namely the walking distance from the origin zone to the closest stop and from the destination stop to the final destination, and the waiting time. Different scenarios were simulated by changing fleet composition and vehicle dispatching strategy. Their comparison allows stating the service quality and performance considerably change in accordance to the quantity and capacity of vehicles.

However, both presented studies only score a probability distribution over travel plans in terms of times, but not over activity goals as human desires. Hence, we can state it is more of a travel optimisation simulation than a fully behavioural model, where information about individuals is taken into account to influence the decision.

It is already of clear evidence that human mobility simulations through ABM are a powerful and robust tool for policymakers in urban planning. Because of that, the field of study is expanding from the transport flow maximization (Lu et al., 2018) to the ecological point of view of emissions and urban planning (Maggi & Vallino, 2016).

Plakolb et al. (2019) calculated the NO_x emissions produced by private motorised vehicles, by simulating the mobility behaviour of citizens. Their approach transforms a spatial map to a net of nodes and edges, on which agents choose how to commute. Again, we are not facing a fully behavioural model, but just assigning probabilities over travel trips. Their model is applied in the case study of Salzburg (Austria) and they investigated a set of scenarios of different changes to the traffic system.

Another study analysed the impacts of bike-sharing on individuals' usage of other transport modes (Lu et al., 2018). Their ABM considers both the use of bike-sharing services and other transport modes in the case study of the city of Taipei. As result, the simulation made it possible to understand the development of a bike-sharing system, through infrastructure extensions and incentives. Consequentially, they estimated

the environmental impacts of the simulated scenarios, in terms of SO_x, NO_x, CO, and GHG emissions. In the most minimized environmental impacts scenario, 22/year premature deaths can be avoided compared with the least minimized case. This model is also a first attempt to embed bounded rationality into the mode choice. Aspiration level, stress threshold and activation level are parameters to design the behaviour.

It is already clear by analysing the previous studies that fully behavioural agent-based models are still new to the field of simulating human mobility behaviours, especially when taking in the Exposome Science. The lack of available data and the complexity to incorporate behavioural rules brought researchers to reduce the number of parameters that affect the model. Nonetheless, recent studies such as Lu et al. (2018) or the framework proposed by Sonnenschein et al. (2021) are bringing light on how to effectively incorporate more in-depth human behaviour in urban studies, indicating the direction for future research. One of its main challenges regards calibration, which is presented in the following sections. To my knowledge, the presented studies propose different approaches to calibration, mostly incorporating only information about the travel system and not considering all the aspects of social behaviour. Uncertainty and lack of details are evident, supporting the idea that calibration, especially in the mobility research field, still need to be explored.

2.5 Model calibration

When simulating, the aim is to get the best representation possible of the reality (Janssen & Heuberger, 1995). Researchers define behaviour rules to be embedded in ABM, based on both theory and empirical evidence, but axioms are insufficiently precise to describe how much each model parameter has to influence the final decision (Crooks et al., 2018). "Calibration aims at adjusting the model parameters" (O'Sullivan, 2004) so that the ABM behaviour capture the scenario behaviour as best as possible. In other words, the main goal is to find the parameter value set that maximizes a fitness function or rather returns the highest similarity to the observed behaviour. ABM has been used for a long time, but just in the last decades, the awareness for calibration has risen.

As Railsback (2001) explains, calibration is mostly composed of two main steps. Firstly, it is necessary to formalise the desired problem, selecting the model parameters that should be calibrated. For instance, some variables could be set as constant by pre-existing theories or assumptions.

Secondly, a mathematical description of model fitness is required, taking into account the final model goal. As affirmed by Crooks et al. (2018), the latter objective function is necessary to represent the similarity between the simulated outcome and the real value from the dataset. The most used statistic measures are R^2 and the Standardized Root Mean Square Error (SRMSE) (Knudsen and Fotheringham, 1986; Crooks et al., 2018). For instance, SRMSE should be approximately equal to the standard deviation of the measurement noise. Researchers often adopt them just because of their simple implementation. There is a plethora of quantitative techniques and Janssen and Heuberger (1995) have contributed with an extensive literature review. Looking at spatial metrics, Crooks et al. (2018) report examples such as Nearest Neighbor Index and Ripley's K, which both returns the degree of clustering in the point dataset. Each objective function has its advantage and disadvantage, and its selection must be carefully related to the parameter characteristics.

2.6 The calibration problem

Calibration is also one of the biggest ABM partially unsolved problems. Prior research has identified two main obstacles: (1) the curse of high dimensionality (Beven, 2002) and (2) hierarchical structures (Crooks et al., 2018). In the next sections, I extensively analyse the two issues.

2.6.1 The curse of high dimensionality

The curse of high dimensionality refers to the fact that agent-based models are often characterised by an infeasible high number of parameters to calibrate. The difficulty arises from a lack of data to use as calibration and computational load limitations (Schulze et al., 2017). Most of the models "suffer from a lack of uniqueness in parameter estimation due to the fact that their assumptions and processes tend to outweigh the data available for a complete assessment of their goodness of fit" (Crooks et al., 2008). As the authors argue, this is especially the case of human mobility behaviour, where there is a lack of detailed empirical

data. On the computational side, calibration usually works by exploring all possible parameter sets and then compare their fitness, often performed through Markov chain Monte Carlo (MCMC) simulations. Schulze et al. (2017) reports how the latter is computationally infeasible for complex models, in particular for human mobility behaviour models (Maggi & Vallino, 2016).

To tackle this problem, a proper parameter selection and prioritisation should be performed beforehand, based on the model usage (Railsback, 2001; Beven, 2002). This is non-trivial, especially where there is a lack of previous knowledge on the subject. Keeping the focus on human mobility behaviour, prior parameter selection is novel to this field. Nonetheless, a new framework has been recently developed to guide researchers (Sonnenschein et al., 2021).

Another way to address the curse of high dimensionality is sensitivity analysis (Schulze et al., 2017). The latter is a statistical approach, which returns the sensitivity of parameters. In other words, it makes it possible to know if a parameter has a significant influence on the model output. As consequence, only variables that bring a consistent change in the output are explored by the calibration method. There are many sensitivity analysis approaches, such as the ones presented by ten Broeke et al. (2016): one-factor-at-a-time (OFAT), which keeps one parameter fixed and varies the remainings; global sensitivity analysis, which identifies interaction effects by sampling the model output over a wide range of parameter values; and regression-based. Within the sensitivity analysis, the Hornberger-Spear-Young method (Hornberger & Spear, 1983) is often used in environmental systems due to its nonparametric requirements, without prior assumptions about variation or covariation of different parameter values (Beven, 2002). Hence, it only evaluates sets of parameter values in terms of their performance.

Finally, as Beven (2002) and Crooks et al. (2018) illustrate, every parameter values combinations leads to a simulation output. The Cartesian product of all possible parameter values of all the parameters, applied to the fitness function, create the so-called "surface space". Overall, calibration aims at finding the best parameter values for the latter set that maximizes the objective function. This can be done manually, but the procedure becomes unfeasible with the increase of parameters. Even high-performance computers struggle with exploring the entire parameter space, in the form of Markov Model simulations. Because of that, optimisation methods aim at finding a faster way for parameter space exploration. While the previous parameter selection is mostly specific to the goal of the model, optimisations methods are usually theoretically applicable to all ABMs. However, applicability does not necessarily mean optimality.

An extensive literature review about which optimisation is suitable for a particular ABM application does not exist yet. Many methods have been developed and all have advantages and disadvantages that should be taken into consideration when selecting one method instead of one other. For example, some techniques aim at lower computational load, while others at higher accuracy. There is not a single best method for all ABM purposes, so a trade-off always occurs. Focusing on human mobility behaviour, previously presented studies such as Lu et al. (2018) refers to use "exhaustive algorithms and heuristic algorithms to find the best parameter combination". However, none focused on selecting a particular calibration method instead of another based on actual evidence.

In section 2.7 I extensively present relevant literature related to state-of-the-art optimisation methods that can be useful in the field of mobility behaviour ABM.

2.6.2 Hierarchical structures

ABM often operates on multiple hierarchical levels. In other words, the output behaviour is generated by the interaction of sub-behaviours at a lower level. For example, is possible to model the traffic volume at a general city level, but also focusing on neighbourhoods, houses or even raster cells Sonnenschein et al. (2021). Each might lead to discrepant optimal parameters. Hence, the overall behaviour calibration might misrepresent the correct patterns at a finer spatial scale. One might compare one type of output with its real observed data but might miss evaluating behaviours at different levels. It is evident how hierarchical structures are part of ABM with GIS, especially when modelling the human mobility behaviour where multiple layers and spatial levels are present. Viceversa, starting from a focused level might not lead to the generally desired behaviour. Because of that, it is important to look at different levels of isolation (Beven, 2002). Focusing on ABM coupled with GIS and more in particular on mobility behaviour, the study and the approach to

hierarchical structures is still lacking (Sonnenschein et al., 2021).

One first way to correctly calibrate hierarchical structure is Pattern-Oriented Modeling (POM) (Grimm et al., 2005). Usually, bottom-up models formulate and compile details about individuals are a lower scale and observe the emergence of top-level properties. The authors illustrate that this approach leads to complexity and uncertainty, while POM tries to tackle these issues. The argumentation for such complex systems is that a single pattern observed at a specific scale and hierarchical level is not sufficient to reduce uncertainty in model structure and parameters. Their solution, instead, relies on using multiple patterns seen in the real world to guide the structure design, in order words, to select what real-world processes to include in the model. The difference from guiding the design by research questions only is that patterns might force us to include state variables and processes that are only indirectly linked to the ultimate purpose of the model and are not part of our initial conceptual model. Indeed, patterns occur at different spatial and temporal scales and different hierarchical levels. An important note on the POM approach is that collecting micro-level and interaction data clearly improves calibration (Filatova et al., 2013). However, Crooks et al. (2018) argue that it can be complicated to quantify the difference between the modelled and observed spatial patterns, even if data are available. Objective functions can sometimes be trivial to be formulated or comparison might even be subjective. Issues are especially human-related since "data collection is a very expensive task and almost always it is impossible to generate long time series for individual or group behaviour" (Troitzsch, 2004).

Secondly, it is clear that hierarchical structures require taking into account different spatial scales. This has been achieved by combining multiple objective functions. Hence, taking into account different objective functions at different levels, from large scales to finer representations, can return a clearer and more general perspective of the model fitness (Beven, 2002).

2.7 Previous studies on optimisation techniques

As previously seen, the majority of the work for researchers is to design an ABM with selected parameters that, with proper objective functions, can return a valuable parameter set. However, another field of research focuses on the more general aspect of optimising the search for parameter values, especially to address the curse of high dimensionality (Baeyens et al., 2016). Theoretically, optimisation methods should all converge to the same parameter set, but at a certain cost. Researchers tackled this problem with the help of disparate approaches and artificial intelligence plays a big role in this resolution. The majority of techniques falls into the category of evolutionary algorithms, starting from an initial state and trying to reach the global optimum by improving at each step the fitness score (Beven, 2002). Most of the methods have been even originally developed for other goals, but have later also found their applicability in ABM. In this section, I present the most important methods from the literature review, together with state-of-the-art implementations developed in the last decade.

2.7.1 Hill-climbing

The first worth mentioning method is hill-climbing, which belongs to the family of local search (Beven, 2002). Indeed, it starts from an arbitrary point of the surface space and it tries to move to a better solution in its neighbourhood. The latter is achieved by making small incremental changes in the parameter set.

The algorithm has been applied to a variety of scenarios, ABM included. As an advantage, its development simplicity is remarkable. By contrast, being a local search procedure requires knowledge of the gradient of the surface, so that the algorithm knows in which direction to climb. Convex scenarios let hill-climbing not distinguish a local optimum from the global optimum.

Also, hill-climbing performs well on smooth surfaces: it is not the case of human mobility behaviours because of its high nonlinear characterisation. However, as a double-check, hill-climbing can be run for a few simulations with different starting points: if the final results correspond we can assume the global optima has been found.

2.7.2 Genetic Algorithms (GA)

Another technique family is genetic algorithms, abbreviated to "GA" (Crooks et al., 2018). Its name is due to the characteristic of imitating a population of individuals representing different parameter sets.

The initial set is randomly chosen as starting point. Then, it is let iteratively evolved in successive generations, aiming at improving its objective function at each iteration. The process continues until a global optimum fitness is reached.

Implementations differ in the operations used to evolve the population at each iteration, such as selection, cross-over and mutation. Railsback (2001) stresses the importance of correctly define the fitness measure in this method family since the individuals' adaptation and selection is strictly based on the objective function. Hence, misrepresentations prevent the success of individuals and end up in an unrealistic model.

2.7.3 Simulated Annealing (SA)

One more method family is simulated annealing, often abbreviated to "SA" (Kirkpatrick et al., 1983). The name comes from its similarity to the water particles behaviour when the liquid is cooled. Analogously to hill-climbing, SA starts from a randomly distributed set of parameters in the parameter space to find a global optimum state, with respect to the performance measure of the optimization problem. The difference here is that there is a rule to accept new parameter sets as optimal. In detail, if the performance measure of the newly generated set is worse, the acceptance rule may still accept it. The probability is based on an exponential function of the difference of performance measure between the new and the current set. Continuing with the algorithm execution, the probability gets reduced similarly to water cooling. The acceptance rule allows avoiding local optima.

Kaveh (2017) reports that SA, acting with small temperature variations and if the solution search can reach an equilibrium for each temperature, the algorithm can attain the global optimal solution. In addition, since SA works with a metropolis process, the latter helps to avoid local optimums. However, slow temperature variations increase the computational load, since more variations are in total computed. Hence, a non-trivial trade-off between temperature function and computational load must be chosen in advance.

2.7.4 Particle Swarm Optimisation (PSO)

As the name suggests, the algorithm family is inspired by the behaviour of birds flying in flocks. Again, similarly to the previous methods, PSO starts from a random set of particles and it tries to improve the latter by looking at the objective function (Kaveh, 2017). The difference is that here the space exploration is made through simple mathematical expressions that model some interparticle communications. As Kaveh (2017) explains, the latter functions "suggest the movement of each particle toward its best-experienced position and the swarm's best position so far, along with some random perturbations".

The reviewer also reports that PSO is averagely faster than previous mentioned evolutionary algorithms. However, the overall quality of the founded solution is lower, especially when increasing the number of generations. Indeed, PSO can fall into local optima as the hill-climbing method. Recently, Zambrano-Bigiarini and Rojas (2013) proposed a PSO implementation specifically developed for ABM calibration.

2.7.5 Machine learning surrogate model

One particularly interesting implementation that caught the attention of researchers has been developed by Lamperti et al. (2018). Their idea relies upon replacing the original ABM with a computationally cheap model as a proxy by combining supervised machine learning and intelligent sampling. Hence, if the approximation error is negligible, then the surrogate model is a good replacement of the original ABM for the exploration space, calibration and sensitivity analysis.

Firstly, a large pool or parameter combinations is selected. Within the latter, a smaller random subset is picked with no repetitions. The latter is used in the original ABM, to label parameter sets into a binary or real-valued user-defined calibration criterion. Afterwards, the surrogate machine learning model is trained on the labelled combinations.

Subsequently, the model is similarly trained with a new small-size subset selected from the original sample. The process is repeated for a fixed time or till a determined threshold is reached.

Their work is based on the Kriging approach by Salle and Yıldızoğlu (2014). The latter differs for using a Nearly Orthogonal Latin Hypercube sampling. Instead, Lamperti et al. (2018) use a quasi-random Sobolo sampling which still performs well when the response surface is completely unknown and contains non-smooth regions, as common in mobility behaviour ABM.

Their case studies were applied to B&H asset pricing and islands models. Results show high accuracy in the proxy and a drastic reduction of the computation time. In particular, benefits arise in presence of highly complex models. However, future research still needs to be completed on other case studies, with even more complex models.

2.7.6 Bayesian optimisation

In the last years, Bayesian optimisation has been widely applied in machine learning scenarios, where a large number of hyperparameters needs to be tuned (Agnihotri & Batra, 2020). As the authors explain, "Bayesian optimization can be used to optimize any black-box function". Similarly to Lamperti et al. (2018), a surrogate model is learned. The code idea is to sample and therefore train according to an acquisition function by optimizing it. With the latter, there is a balance between exploring uncertain regions against their exploitation. This approach is called "case-based inference".

There are several benefits of using Bayesian optimisation. Compared to all the previous mentioned frequentist methods, even the surrogate model of Lamperti et al. (2018), Bayesian optimisation does not necessity the pre-selection of meta-models. In other words, it does not suffer from the issue of choosing the starting point of the search, which influences the overall space exploration. The latter also has another positive effect, which is not being a local optima search. In addition, Bayesian inference allows the incorporation of prior information, permitting a reduction of uncertainty. Nonetheless, the definition of the prior is still an open issue for many fields.

However, in the current year, a novel Bayesian optimisation implementation calibration stood out from the literature review of ABM calibration. Shiono (2021) proposes a likelihood-free Bayesian approach, based on the BayesFlow inference of Radev et al. (2020).

The method is called "amortized inference", compared to the previous "case-based inference". Radev et al. (2020) explain that estimation is divided into a likely expensive initial training phase, followed by a performing inference phase. With this technique, the training effort amortizes over repeated evaluations.

The training phase learns an approximate posterior, rather than a likelihood, by using a conditional invertible neural network (cINN), in particular, bidirectional long short-term memory (bi-LSTM) as a summary network. The authors claim that BayesFlow aims at drastically reducing the computational cost, even in complex models. In addition, scalability is also mentioned as one of the biggest strengths, requiring no assumption of the shape of the surface space. In theory, this calibration method could be applied to all types of ABM.

Results of Radev et al. (2020), authors of the original BayesFlow, already showed high accuracies even in presence of chaotic models (Ricker population and Levy-Flight). Shiono (2021) also presents an evaluation of BayesFlow against two MCMC methods on a macroeconomic ABM, returning the former as the winner in terms of RMSE and R^2 scores.

The author mentions that a shortcoming of using a neural network approach is that a better and more specific hyperparameters tuning can unlock even better performances, although the default settings already lead to excellent results. The combination of machine learning and Bayesian optimisation seems to be promising, as it has been recently applied also in the field of image classification (Maroñas et al., 2020).

2.7.7 Other algorithm implementations

Literature review of optimisation methods becomes extremely large if considering all the implementations developed in the past. For example, algorithms such as stochastic gradient descent (SGD) are relevant in the deep learning area and might be applied to ABM similarly to the surrogate model of Lamperti et al.

(2018) (van der Hoog, 2019). In particular, they solve the known issue of learning the optimisation gradient by using derivatives.

Another worth mentioning approach is the Direct Search algorithm of Baeyens et al. (2016). The latter minimizes an arbitrary real-valued function by adopting a new function transformation and three simplex-based operations. In particular, the former improves global exploration, while the second evolution mechanism assures the termination and global convergence to a stationary parameter set. As an advantage, the author claims a simple implementation and applicability to high-dimensions models, since computation time linearly increases with the dimension of the problem.

More specific traffic calibration tools have also been developed over the years, such as Cadyst (Flötteröd et al., 2012) and Odyst (Flötteröd, 2017). The former for instance was used to calibrate a traffic model of the Greater Zurich region in Switzerland. Although there is explicit mention about behavioural mobility, the mode choice taken into consideration is just restricted to a few trip information, without considering life aspects of the individuals. For instance, the mode choice of the Zurich case study was based only on the travel type, such as home, work or education.

2.7.8 Previous comparisons

Evaluations and comparisons of calibration optimisation techniques have just been carried out partially. In detail, none of the previous studies was related to human mobility behaviour ABM. Also, comparisons included just a small subset of the presented techniques, mostly the well-known hill-climbers and GAs.

For instance, Stonedahl and Wilensky (2011) proposes the BehaviorSearch tool, an extension to the MatSim platform (Horni et al., 2016). They analysed hill-climbing against GA and a basic random search algorithm. They used two models of bird flocking behaviour and they noticed that the GA outperforms the others two. Their intuition relies upon the fact that GA is population-based, so it can explore multiple regions of the space simultaneously like a global search. The latter consequently permits it to come up with creative or unexpected solutions to solve a problem. Nonetheless, their conclusions are just suppositions coming from case studies. Hence, it is still partially unclear to them why the GA won.

As also analysed by Prügel-Bennett (2004), the performance difference between GA and hill-climber mostly depends on the problem. In his case study, he presents a new toy optimisation problem as an example in which GA, especially with crossovers, outperform hill-climbers.

3 Research questions

It is evident from the literature review that several calibration techniques have been developed over the last century. This has especially been emphasised in the last decade with the rising of calibration awareness linked to the use of more complex models. However, it is clear that there is a gap in understanding the steps to properly calibrate a full human mobility behaviour ABM, from parameter selection to search optimisation. Because of this knowledge gap, the following research questions are formulated:

RQ1. *Which calibration methods are suitable for ABM of mobility behaviour?* Focusing on human mobility behaviour ABM, what are the methods and tools that take part in a suitable calibration?

Since the RQ requires to determine a mapping between behavioural models and calibration methods, the related subquestions are included:

- RQ1A. What are the characteristics of behavioural models?
- RQ1B. What are the characteristics of calibration methods?

RQ2. *What is an appropriate experimental setting for simulations when calibrating?*

- RQ2A *Which hierarchical levels should mobility behaviour be calibrated on?* Because of the presence of hierarchical structures in mobility behaviour ABM, what are the spatial levels on which calibration should focus and consider?
- RQ2B *What objective function should be used?*
- RQ2C *Is it possible, and if yes, how, to reduce the high dimensionality of parameter sets?*

RQ3. *What is an appropriate experimental setting for comparing calibration methods?*

- RQ3A *What hyperparameters are suitable for calibration techniques on ABM of mobility behaviour?*
- RQ3B *How to compare the fitness of different calibration methods?* Since an evaluation of the calibration method is being carried on, how can we evaluate each technique in different terms of performance?

RQ4. *In the case study of the Amsterdam ABM, which calibration method is preferred?*

4 Methodology

A data-driven approach is used to study calibration approaches to human mobility behaviour in ABM and how they compare. In section 4.1 I describe the ABM developed by the team and used for this case study. Section 4.2 and 4.3 explain how I apply calibration to such model and the selected calibration data. Finally, in section 4.4 I describe which optimisation algorithms I use, their hyperparameters, the setup for their performances comparison and the selected metrics.

4.1 Amsterdam behavioural agent-based model

The presented agent-based model used in this thesis is the result of preliminary work by the team. Hence, the model represents only a proof-of-concept, and it will be improved in the future years when applied to real simulations for the EXPANSE studies. The current goal is not to predict and analyse impacts on CMPD yet, but to focus on calibration instead. Nevertheless, this thesis aims to define a proper framework and set of procedures that will be then applied in a future working model.

The EXPANSE project is currently working on a set of cities, and Amsterdam was selected because of the large availability of resources. In particular, the designated case study area is composed of the municipalities of Amsterdam, Diemen and Ouder-Amstel (figure 1). The territory is described by postcodes (PC4 level) retrieved by the national statistics agency (CBS, 2019a). Dutch postcodes are composed of 4 numerical digits, representing the so-called PC4 level. More detailed layers, such as PC6, exist and return a very small area. However, most open data available is still based on PC4, so the presented model is also based on this standard.

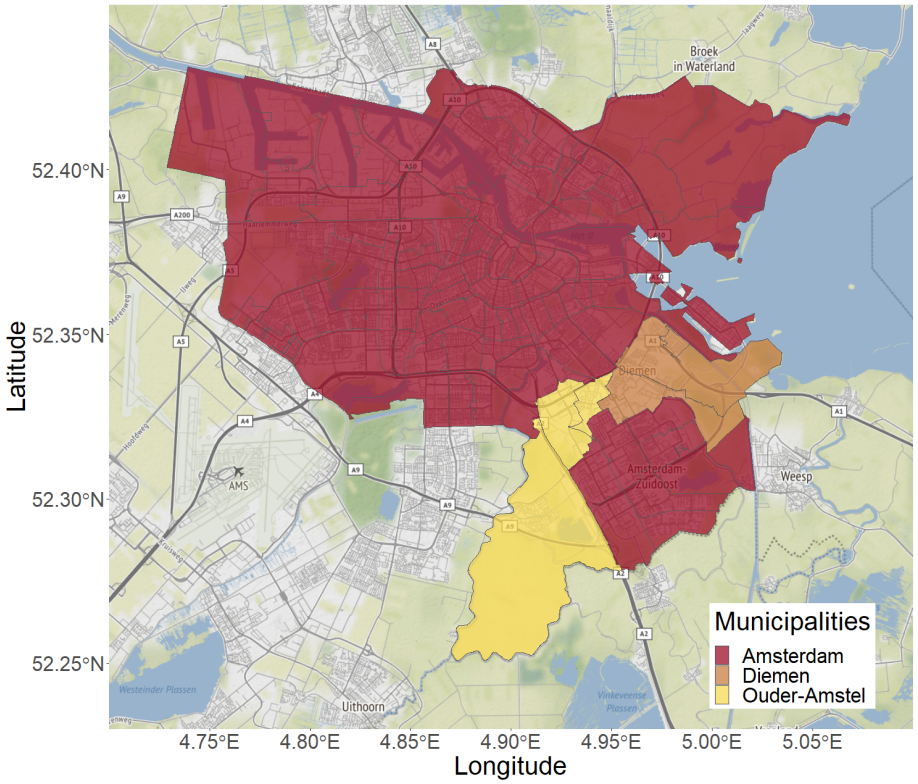


Figure 1: Municipalities included in this research. Amsterdam, Diemen and Ouder-Amstel are the subjects of this case study.

Among the various software for ABM implementation, the open-source GAMA (Taillandier et al., 2019)

was selected. It is an open-source modelling and simulation development environment for building spatially explicit agent-based simulations. The last stable release GAMA 1.8.1 is used, based on the very old Java 1.8 version. Because of that, simulations performances are influenced by such an old implementation. However, a new version (1.8.2) should be released soon, based on the more recent Java 8, and better performances can be achieved in future studies. GAMA is still chosen because of its medium-large scale capabilities. Another reason is its built-in capability to deal with spatial data, such as spatial networks. Finally, an active online community is a positive factor to consider, especially for future developments.

The ABM aims to represent individuals' mobility behaviour. In particular, it concerns mobility decisions, such as walking, biking and driving. The EXPANSE goal is to predict effects of environmental interventions on a city's population in terms of behavior. In order to estimate the impact of these variables that can be influenced by interventions, such as the case of urban planning variables, the model has to be explicit and explainable.

The selection of factors to include in a model is a non-trivial point of model design. The latter task was done based on literature and previous work that investigated the relevance of such variables for behavior. It is not the focus of this thesis, which instead targets calibration.

In brief, for this proof-of-concept ABM, a data-driven approach by analysing the available attributes included in the calibration dataset will hep the researchers with the selection. In future studies, the aim it to use an automatised knowledge extraction method of review studies, based on BERT (Devlin et al., 2018) and automated ontology engineering, to inform the attribute selection of behavioural models and do structural model validation.

The used model is composed of two main components: the agents (subsection 4.1.1) and the environment where the agents' displacements occur (subsection 4.1.2).

4.1.1 Agents attributes

For this study, as better explained in section 4.2.1, agent attributes are configured with the representative Dutch travel survey ODiN (CBS, RWS-WVL, 2020). The attribute availability from the latter dataset influenced the model design. In particular, each agent contains the following information:

Attribute	**Description**
Group age	Categorical information to group the age: • "minor": age<=8 • "youngadult": 9 <= age <= 17 • "adult" 18 <= age <= 49 • "senior": 50 <= age <=69 • "elderly": age>=70
Residential postcode	Postcode (level PC4) of where the agent's house is located.
Current position	Current location of the agent expressed in postcode (PC4 level).
Has child	Binary value that expresses if the agent has a child. 0 means no child, while 1 at least one child.
Absolved education	It expresses the agent's education level. Possible values: • 0: nothing or primary schools • 0.5: low or secondary vocational schools • 1: university
Body mass index (BMI) group	BMI is an international measure that describes whether the individual has a healthy weight in relation to height. Here the information is classified into two main groups: • "overweight": BMI is "obese" or "moderate overweight" • "normal": otherwise

Car in household	Binary value expresses if the agent's household possesses at least one car. 0 means no car, while 1 at least one car.
Has driving license	Binary value expresses if the agent possesses a driving license. 0 means no license, while 1 yes.
Carpooling	Binary value expresses if the agent has made at least a trip with carpooling during the simulation day. 0 means not at all, while 1 yes.
Habit using bike	It expresses the frequency of the agent using bikes (both electric and non-electric). Possible values: • 0: never or almost never • 0.25: a few times a year • 0.5: a few times a month • 0.75: a few times a week • 1: daily or almost daily
Habit car driver	It expresses the frequency of the agent using the car as a driver. Possible values: • 0: does not apply; agent under 17 years old • 0.2: never or almost never • 0.4: a few times a year • 0.6: a few times a month • 0.8: a few times a week • 1: daily or almost daily
Habit car passenger	It expresses the frequency of the agent using the car as a passenger. Possible values: • 0: never or almost never • 0.25: a few times a year • 0.5: a few times a month • 0.75: a few times a week • 1: daily or almost daily
Leisure	It expresses the average frequency of displacements completed with a motive of leisure.
Household income	Standardised household disposable income, divided into 10% groups. As CBS reports (CBS, 2020a), "the disposable income is adjusted for differences in household size and composition. This is also referred to as purchasing power". It indicates the household's average income, returning an indication of economic well-being expressed with a float value between 0 and 1.
Income group	Categorical information calculated on the previous household income attribute: • "low": household income <=2 • "middle": 3 <= household income <= 5 • "high": household income <= 10

Table 1: Attributes of the agent population. The table is divided in sub-blocks based on the attributes nature for better understanding. For example, the age information (age, group age) are separate from the spatial ones (residential postcode, current position).

Location-based attributes, namely the residential postcode and the current position, are expressed in a string postcode format. In particular, because of the calibration dataset used (see section 4.2.1), the spatial level available is PC4. In order to spatially represent agents, the postcodes are firstly transformed into geographical coordinates by CBS data (CBS, 2019b) and then into coordinates points by taking the centroid. In future simulations, the aim is to use neighbourhoods codes for smaller areas and higher accuracy.

4.1.2 Environment

The model's environment is crucial, primarily due to its influence of the modal choice, as the previous literature review highlighted. GAMA includes a graphical view, but it will not be discussed in this thesis since it does not play a role in calibration. More importantly, the environment is divided into raster grid cells with a resolution of 200x200 meters. In order to represent the environment, various kinds of spatial information were transformed to statistical summaries for each cell, namely:

Attribute	**Attribute**
Population density	The spatial distribution of population in 2020 with country total adjusted to match the corresponding UNPD estimate.
Retail density	Venues density of arts entertainment, food, nightlife and shops.
Green coverage	Density of green spaces, including official parks and smaller urban green areas.
Public transport density	Density of how many stops of public means of transports, such as trains, buses and undergrounds.
Street connectivity (road intersection density)	Density of how many streets intersections are present in the cell.
Pedestrian accidents	Number of accident reports from the police that happened in the cell from the 1st January 2003 to the 31st December 2020.
Highway length	Total length of highways in the cell.
Distance to Central Business District (CBD)	Distance of the cell centroid to the closest commercial and business center, such as commercial space and offices.

Table 2: Environmental factors implemented in the ABM.

In addition, the model contains weather information in the form of rain precipitations for each simulation day. Since the study area is not particularly large, the information is not spatially referenced.

4.1.3 Agents behaviour

When it comes to modelling human mobility behaviour, many factors can be considered. Indeed, as highlighted by the literature, formalising human behaviour is one of the major issues. Here, the goal of the behaviour is the decision-making process of modal choice. In other words, the selection of a means of transport, given each a displacement to do. Hence, this study does not model the decision-making behaviour of the daily life of an agent, such as generating an activity schedule and tasks, because those are not part of the modal choice mobility focus. In a future evolution of this EXPANSE ABM, agents' schedules will be generated based on heterogeneous statistics retrieved from Harmonised European Time Use Surveys (HETUS) (Eurostat, 2022). The latter consists of national surveys conducted in EU countries to quantify how much time people spend on various activities.

For this research, instead, a more simplified approach is taken by generating an agent population that corresponds to the ODiN survey in terms of amount and activity schedule as declared by the participants, as better explained in section (4.2.3). The activity schedule is transformed into a location schedule in the form of a postcode PC4 list, of 10 minutes steps, for each agent. In addition, the day of the performed displacements is included.

When it comes to means of transport, only walking, biking and car driving have been implemented in the ABM because of limited time.

Behaviour implementation. Figure 2 illustrates the overall behaviour of an agent during one simulation. Each simulation represents one day, from 00:00 to 23:59.

In detail, every 10 minutes, the agent compares its current location with the location schedule at the correspondent time. If there is no equivalence, a location change has to happen, and the displacement manager is triggered. The model works only with locations within the case study area as a simplification.

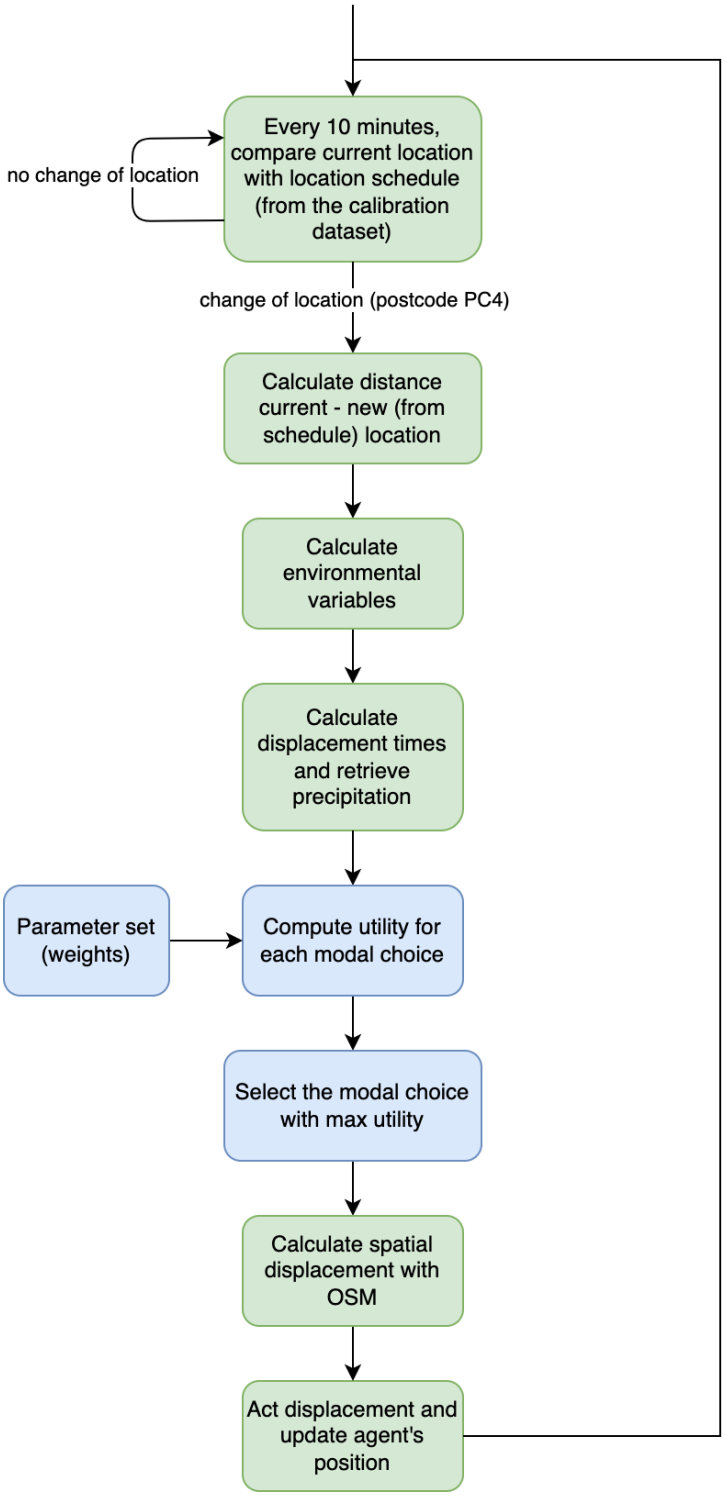


Figure 2: Agent behaviour during a simulation of the ABM. Green tasks highlight preparation tasks such as data retrieving or spatial updates, while blue steps are the core parts of modal choice.

One reason is that the intent is to model mobility behaviour in the Amsterdam case study area only. Another consideration is simplicity. Indeed, implementing the spatial information of all the postcodes in The Netherlands, or even abroad, would exponentially increase the size of the polygonal data used to represent such information. For this proof-of-concept model, such large scale information is not necessary. Firstly, an estimate of the displacement is calculated by computing the euclidean distance in meters from the origin to the destination point. To achieve this, postcodes are linked to polygonal data formats as explained for the location (see section 4.2.1) and then transformed into points by taking the centroid as an approximation.

The ABM starts by retrieving the average for each environmental factor (section 4.1.2) of the cells intersected by the line that connects the origin and destination points. Also, the rain precipitation is retrieved by matching the displacement day with the weather dataset (section 4.2.1).

Subsequently, the travel time is estimated for each modal choice. The latter is estimated with the formula:

$$\text{displacement time [modal choice]} = \frac{\text{displacement distance}}{\text{speed [modal choice]}}$$

where the modal choice speed is a constant value depending on the modal choice and based on previous literature. In detail:

- Walk speed = 1.4 m/s, which corresponds to 5km/h. Retrieved from Browning et al. (2006).
- Bike speed = 3.33 m/s, which corresponds to 12km/h. Retrieved from BicycleDutch (2018).
- Car speed = 11.11 m/s, which corresponds to 40km/h. No average speed data source was found by the team. Hence, this value assumption is based on the fact that most major streets have a 50km/h speed limit, but traffic lights reduce the average speed.

As a model simplification, the latter values are fixed and homogeneous for the entire population. However, future model designs could implement a differentiation based on heterogeneous agent characteristics, such as age or BMI group. In addition, dynamic values depending on the street type can be implemented in the following model versions.

4.1.4 Modal choice decision

Once the displacement is determined and attributes computed, the agent moves to the core modal choice decision-making phase.

The ABM uses a linear utility maximisation framework by aggregating behavioural and environmental attributes that are previously identified to be significant for each modal choice. BDI has been initially established as a possible solution, but assuming an unknown behavioural process can potentially introduce errors in a model. Hence, a more basic linear utility maximisation allows fewer assumptions about the decision making processes. In particular, for this thesis, having a more computationally efficient model is significant for calibration.

In detail, a utility value is calculated for each modal choice because each considers different factors due to its nature. Each utility aggregates behavioural and environmental factors by summing them and averaging them by the number of factors. Each factor is, in turn, composed of the product of the attribute values previously calculated and the weight subjected to calibration. Weights are divided by modal choice, so the values found, for example, for walking or biking, differ from the car ones. As a specification, all weights have two decimals of digits precision, and they range from -1.00 to 1.00 because factors can positively or negatively influence the utility. Formally, we can describe the utility function and modal choice selection as:

$$\text{utility}_j = \frac{\sum_{i=1}^{k} \text{weight}_{i,j} \ * \ \text{attribute}_{i,j}}{k}, \forall j \in [\text{walk, bike, car}]$$
$$, k = \text{total number of factors}$$

$$\text{modal choice} = max(\text{utility})$$

As previously stated, a data-driven approach by analysing the available attributes included in the calibration dataset helped the selection of attributes to include non only in the model but also to consider each modal choice. This was achieved by multiple regression models where the targeted mean of transport was predicted (independent variable) with all the selected attributes. McFadden pseudo R^2 approximates the amount of variation that was explained to the ABM. Scores of 0.35 for walking, 0.33 for biking and 0.44 for driving were found. As a consequence, table 3 describes the factors that take part in the utility for each means of transport.

Factor	**Modal choice**		
	Walk	Bike	Car
Travel time	✓	✓	✓
Travel distance	✓	✓	✓
Household income	✓	✓	✓
Absolved education		✓	✓
Leisure	✓		
Agent has child	✓		✓
Habit using bike	✓	✓	
Habit car driver			✓
Habit car passenger			✓
Carpooling			✓
Agent has car license			✓
Car in household			✓
Precipitation	✓	✓	✓
Population density	✓		
Retail density	✓	✓	
Green coverage	✓	✓	
Public transport density	✓		
Road intersection density	✓	✓	
Pedestrian accidents	✓	✓	
Highway length			✓
Distance to CBD		✓	✓

Table 3: Factors that are considered for each modal choice. The decision arises from a previous literature meta-study where the attribute significance is studied in relation to the modal choice.

It is clear from the literature that models need to be able to take into account heterogeneous groups of agents. Certain behavioural factors can be different among diverse social groups. For example, it is intuitive to believe that the willingness to travel distance differs between age or weight. A young person might be more active than an elderly and, consequently, more willing to travel by bike for long distances. Because of this characteristic, one important aspect of this work is the implementation of heterogeneous weights, in this case, of the travel distance.

The latter factor has two weights: one correlated with the age group and one with the BMI one. As the before-mentioned example, travel distance plays a different role in their model choice. Hence, 5 travel distance weights are linked to the age groups, and 2 are linked to the BMI group. Each agent uses only the weight corresponding to its attribute during a simulation. In detail, these are the possible weights for travel distance. They only apply for the walk and bike modal choice because it is believed that age and BMI groups do not affect the driving distance choice. Weights are denoted by the suffix "`w_`". Since the travel distance weight is also dependent on the modal choice, the following weight notation includes the related `modal_choice` filter that can assume 'walk', 'bike', 'car' values. For example, `w_distance['walk', BMI_group]` represents the weight for the modal choice walk based on the the BMI group, while `w_distance['bike', age_group]` for biking and age group.

- `w_distance[modal_choice, age_group=='minor']`

- `w_distance[modal_choice, age_group=='youngadult']`
- `w_distance[modal_choice, age_group=='adult']`
- `w_distance[modal_choice, age_group=='senior']`
- `w_distance[modal_choice, age_group=='elderly']`
- `w_distance[modal_choice, BMI_group=='normal']`
- `w_distance[modal_choice, BMI_group=='overweight']`

For example, a senior, normal weight, low-income individual will use the weights `w_distance[age_group=='senior']` and `w_distance[BMI_group=='normal']`.

Both attributes and weights are scaled to the range [0,1] so that also the final utilities are in the same range and allow the maximization selection. To achieve this, environmental attributes (see sections 4.1.2) are normalised with formula:

$$a\_scaled = \frac{a_i - min(a)}{max(a) - min(a)}, \forall a_i \in a$$
$$, \forall a \in \{\text{environmental attributes}\}$$

In detail, a_i is each value of the attribute a.

Travel time and travel distance cannot be scaled with the aforementioned formula because since the values are calculated at each simulation step, it is not possible to know a priori which are their maximum and minimum values. Hence, the use of the hyperbolic tangent function Tanh is here chosen as an approximation for the travel time and distance attributes. The latter function always returns values between 0 and 1 by converging very high numbers to 1. It is true that differences in very high values become imperceptible using the Tanh function. However, but it is likely that very large distances have a similar impact on belief, while it is more important to differentiate smaller distances. Hence, this approach is a good approximation and it should not significantly distorce the modal choice behaviour.

The choice of normalising attributes instead of using the real values arises from the necessity of having all the attributes with the same impact on the final utility. For example, without normalising, the travel distance attribute usually contains a considerable high value, disproportionate to the other attributes. In that case, the utility will be mostly influenced by that attribute, shadowing the impact of the other attributes. This modelling choice is relevant in such a field where the future final goal is to analyse the change of behavioural weights when the model is subjected to different infrastructure changes.

Travel time and distance attributes are also inverted because a higher value must negatively influence the utility to maximise. Weights, instead, are determined by the calibration process.

As previously explained for the travel distance weight, all weights that exists for more than one modal choice are distinguished with a list-programming-like notation. For example `w_time['walk']` is the weight for the travel time for the modal choice walk, while `w_time['bike']` is for biking. Instead, there are weights such as `w_population_density_walk` that do not have any multiple value because, as table 3 suggests, they only exists for one modal choice. In addition, since the travel distance weight is heterogeneous, both the age group weight and the BMI one are multiplied together in the utility formula, to describe a unique travel distance weight.

Here are the full utility functions for the three implemented means of transport:

$$
\begin{aligned}
\texttt{walk\_utility} = (& \\
& (1 - tanh(\texttt{time['walk']})) * \texttt{w\_time['walk']} + \\
& (1 - tanh(\texttt{distance})) * \texttt{w\_distance['walk', age\_group]} * \texttt{w\_distance['walk', BMI\_group]} + \\
& \texttt{population\_density} * \texttt{w\_population\_density\_walk} + \\
& \texttt{retail\_density} * \texttt{w\_retail\_density['walk']} + \\
& \texttt{green\_coverage} * \texttt{w\_green\_coverage['walk']} + \\
& \texttt{public\_transport\_density} * \texttt{w\_public\_transport\_density\_walk} + \\
& \texttt{road\_intersection\_density} * \texttt{w\_road\_intersection\_density['walk']} + \\
& \texttt{pedestrian\_accidents} * \texttt{w\_pedestrian\_accidents['walk']} + \\
& \texttt{leisure} * \texttt{w\_leisure} + \\
& \texttt{has\_child} * \texttt{w\_has\_child['walk']} + \\
& \texttt{habit\_use\_bike} * \texttt{w\_habit\_use\_bike['walk']} + \\
& \texttt{precipitation} * \texttt{w\_precipitation['walk']} + \\
& \texttt{household\_income} * \texttt{w\_household\_income['walk']} \\
&)/13
\end{aligned}
$$

$$
\begin{aligned}
\texttt{bike\_utility} = (& \\
& (1 - tanh(\texttt{time['bike']})) * \texttt{w\_time['bike']} + \\
& (1 - tanh(\texttt{distance})) * \texttt{w\_distance['bike', age\_group]} * \texttt{w\_distance['bike', BMI\_group]} + \\
& \texttt{retail\_density} * \texttt{w\_retail\_density['bike']} + \\
& \texttt{green\_coverage} * \texttt{w\_green\_coverage['bike']} + \\
& \texttt{road\_intersection\_density} * \texttt{w\_road\_intersection\_density['bike']} \\
& \texttt{pedestrian\_accidents} * \texttt{w\_pedestrian\_accidents[bike]} + \\
& \texttt{distance\_CBD} * \texttt{w\_distance\_CBD['bike']} + \\
& \texttt{absolved\_education} * \texttt{w\_absolved\_education['bike']} + \\
& \texttt{habit\_use\_bike} * \texttt{w\_habit\_use\_bike['bike']} + \\
& \texttt{precipitation} * \texttt{w\_precipitation['bike']} + \\
& \texttt{household\_income} * \texttt{w\_household\_income['bike']} \\
&)/11
\end{aligned}
$$

$$
\begin{aligned}
\texttt{car\_utility} = (& \\
& (1 - tanh(\texttt{time['car']})) * \texttt{w\_time['car']} + \\
& (1 - tanh(\texttt{distance})) * \texttt{w\_distance\_car} + \\
& \texttt{highway\_length} * \texttt{w\_highway\_length\_car} + \\
& \texttt{distance\_CBD} * \texttt{w\_distance\_CBD['car']} + \\
& \texttt{car\_household} * \texttt{w\_car\_household} + \\
& \texttt{car\_license} * \texttt{w\_car\_license} + \\
& \texttt{has\_child} * \texttt{w\_has\_child['car']} + \\
& \texttt{absolved\_education} * \texttt{w\_absolved\_education['car']} + \\
& \texttt{habit\_car\_passenger} * \texttt{w\_habit\_car\_passenger} + \\
& \texttt{habit\_car\_driver} * \texttt{w\_habit\_car\_driver} + \\
& \texttt{carpooling} * \texttt{w\_carpooling} + \\
& \texttt{precipitation} * \texttt{w\_precipitation['car']} + \\
& \texttt{household\_income} * \texttt{w\_household\_income['car']} \\
&)/13
\end{aligned}
$$

Taking into account the heterogeneity weights division, there are 19 weights for walk and car separately, while 13 for car. Hence, there are 51 individual weights in total to calibrate.

Usually, the modelling design meets the calibration feasibility requirements to find a compromise. However, the amount of weights to be calibrated here is rather extensive because in such a model where the future final goal is to analyse the parameter changes when subjected to urban interventions, it is not possible, for example, to fix some weights to reduce the parameter space. There is no previous literature about possible fixed weight values because the discovery of behavioural weights is based on a data-driven approach. In addition, as previously mentioned, the attributes have been selected in a previous significance analysis; hence each should play a consistent role, and its removal might compromise the final fitness quality.

After computing the utilities, the modal choice is determined by picking the highest utility.

To conclude the module, the displacement is retrieved by fetching an OSM local server. The latter implementation is based on the GitHub repository https://github.com/Project-OSRM/osrm-backend. It returns the spatial coordinates of the points that are part of the move, based on origin and destination points, but also the modal choice. With this, GAMA can graphically represent the displacement. The agent position is also updated with the new location and ready to check a change after 10 minutes.

4.2 Calibration of the agent-based model

4.2.1 Data sources

For this case study of the city of Amsterdam, Diemen and Ouder-Amstel, the ODiN 2019 dataset is used to configure and calibrate the agent population (CBS, RWS-WVL, 2020). The latter is composed of individual statements on where they travelled in The Netherlands on a specific day, for what purpose, by what means of transport and how long it took to get there. The dataset also contains additional personal characteristics about the individual. It is a cross-sectional dataset, which means that variations in travel behaviour on an individual level cannot be determined.

Since ODiN does not incorporate any information about the individual BMI, the latter attribute was synthetically reconstructed. Firstly, a distribution of the BMI index over the stratified population is retrieved from the national statistics CBS (CBS, 2020b). Then, a second CBS dataset (CBS, 2019c) permits distributing the BMI index over the PC4 residential addresses.

Combining ODiN 2019 and the BMI information made it possible to construct an agent population containing all the ABM individual attributes (see section 4.1.1). Also, each agent contains a list of trips

taken place on the survey day, in the form of origin/destination postcodes (PC4), times and modal choice.

For the environmental attributes (section 4.1.2), the following different data sources are used:

- **Population density**. Retrieved from WorldPop (2020)
- **Retail density**. Retrieved from Foursquare (2022)
- **Green coverage**. Retrieved from OSM via Geofabrik (https://www.geofabrik.de/) as suggested in Novack et al. (2018).
- **Public transport density**. Retrieved from of Amsterdam (2021).
- **Street connectivity** (road intersection density). Retrieved from OSM and processed via Mapcruzin (https://mapcruzin.com/).
- **Pedestrian accidents**. Retrieved from Rijkswaterstaat (2022).
- **Highway length**. Retrieved from of Amsterdam (2017).
- **CBD** (Distance to Central Business District). Calculated as the distance from the point with the largest retail density attribute.

Finally, daily weather information in the form of precipitation is retrieved from the national weather portal KNMI (2022).

4.2.2 Data pre-processing

ODiN is a national-wide collection; however, since it is location referenced, it is possible to focus on the case-study area. Because of the latter, the following series of filtering and data cleaning was necessary. Initially, the original national dataset contained 53380 participants.

Firstly, participants that live outside of the study area are removed. Hence, based on the official national CBS reference (CBS, 2019a), I select individuals with residential postcodes in the case study area (Amsterdam, Diemen and Ouder-Amstel). After this initial filtering, the population results in 2133 individuals.

Secondly, agents that are missing information on their characteristics are removed. Individual attributes play a role in the modal choice, and it would be incorrect to calibrate an agent with incomplete and distorted information. With this pre-processing, the population size was reduced to 1919 individuals.

Afterwards, individuals that have a missing origin/destination postcode are removed. The same filtering applies to omitted origin/destination times. Again, incomplete data can lead to behavioural misrepresentation. As a result of such a pre-processing, an agent population of 1911 individuals is generated. Because of computational constraints, a fixed random subset of 1000 agents is then selected for the simulations. Hence, every simulation is run with the same set of agents.

A few adjustments had to be made for the correct operation of the model. The used ABM focuses on the case study behaviour only. Because of the latter, displacements that are partially outside the previously mentioned postcodes are not considered. In the case of an individual that goes outside of the area but returns to another postcode, the new position is updated without generating a displacement. Also, short displacements that happen in the same postcode area are not considered. Indeed, there is insufficient information to simulate a trip because the starting and endpoint would coincide. Finally, displacements with not implemented modal choices are discarded.

To generate the travel schedule for each agent, as explained in the next section 4.2.3, a list of postcode locations visited every 10 minutes by the agent is retrieved from the carried out displacements. Also, the correspondent modal choice is present in a related list for each travel.

For better understanding, a random agent contains the following processed ABM entries:

- **Group age**: "adult"
- **Residential postcode**: 1091

- **Has child**: 0
- **Absolved education**: 1.0
- **Car in household**: 0
- **Has driving license**: 1
- **Carpooling**: 0
- **Habit using bike**: 1.0
- **Habit car driver**: 0.6
- **Habit car passenger**: 0.5
- **Leisure**: 0.2
- **Household income**: 1.0
- **Income group**: "high"
- **BMI group**: "normal"
- **Postcode schedule**: [1096, .., 1091, ..., 1096, ..., 1091]
- **Modal choices**: [bike, walk, bike]
- **Day of commuting**: 24/11/2019
- **Precipitation**: 0.0

4.2.3 Calibration implementation

Figure 3 illustrates the overall tasks that are involved in calibrating the model weights. Firstly, the exploration algorithm randomly selects the initial parameter set. The way this is implemented depends on the unique algorithm. Section 4.4 presents a collection of evaluated approaches, but the same overall scheme can be applied to every algorithm.

Next, A simulation of 1000 agents is initiated. The quantity of agents depends on the available computational resources. A higher number of agents represents better the case study population, so researchers should try to maximise this quantity as much as possible. In the simulations, agents compute their displacements following the given ODiN 10 minutes locations schedule. The ABM generates modal choices for each displacement and compares the latter predictions with the ODiN observed values. Finally, an objective function (section 4.2.4) is computed on the latter comparison and helps determine if the simulated parameter set is optimal or which one should be explored next.

4.2.4 Objective function

RQ2A inquires about the level on which the objective function should focus. In this model and more in general in human mobility behavioural models, the aim is to calibrate the individual behaviour. Hence, the objective function must compare how well the model can predict modal choices. Compared to other studies such as Plakolb et al. (2019), the aim is not to calibrate the average behaviour focusing on the transportation flow.

Literature review proposes objective functions such as SRMSE or R^2. These metrics work only with continuous outputs. However, in human mobility calibration, the model tries to classify displacements into modal choice classes, so the output is categorical. The process should be seen as a machine learning classification problem. Intuitively, accuracy can be a possible solution. It returns the number of true positives over the entire quantity of predictions. Nonetheless, the mobility problem represents a scenario with

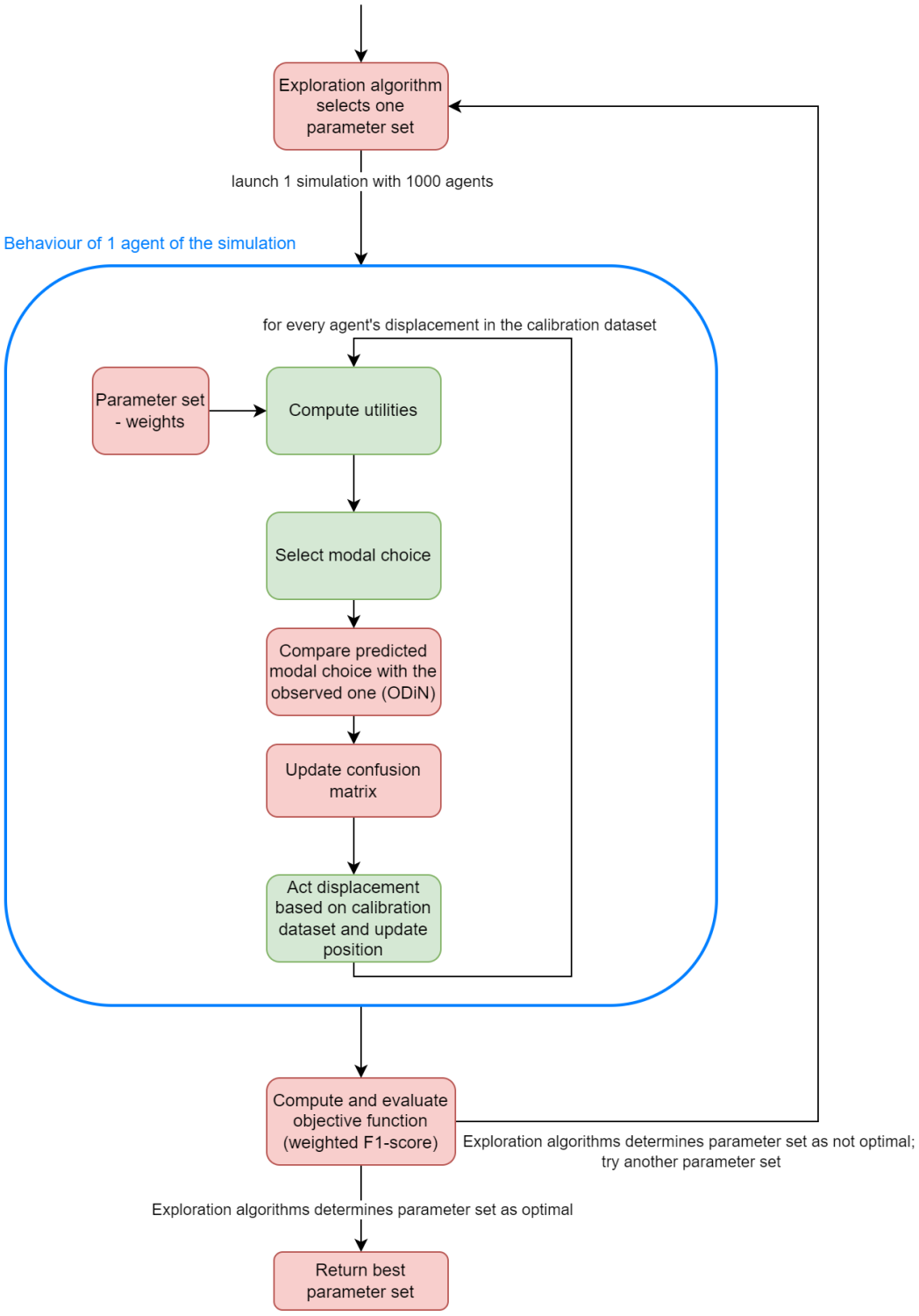


Figure 3: Functioning overview of the calibration process. Red steps highlight calibration tasks, while green ones related to the ABM itself. The blue border delimits the are of one simulation with the fixed 1000 agents ODiN 2019 population.

imbalanced classes. Indeed, after analysis, the processed calibration ODiN dataset, 1328 of the displacements appear to happen by bike, while only 308 by foot and 515 by car. Accuracy could then return a high result because the model might predict the most common means of transport (bike), but still, it cannot consider less-used modal choices.

To tackle RQ2B, the weighted-averaged F1-score appears to be a proper objective function since it takes into account unbalanced classes by combining recall and precision. As figure 3 illustrates, after each modal choice prediction a confusion matrix is updated. F1 scores are calculated for each modal choice at the end of the day simulation. The scores are combined into the weighted-averaged F1 score by weighting in proportions of the observed modal choice classes of the calibration dataset:

$$\begin{aligned}\text{Weighted-averaged F1 score } =&(F1_{walk} * \% \text{ walk observed displacements}) + \\ &(F1_{bike} * \% \text{ bike observed displacements}) + \\ &(F1_{car} * \% \text{ car observed displacements})\end{aligned}$$

Finally, the exploration algorithm decides if and how to select the next parameter space to simulate based on such weighted-averaged F1 score.

4.3 Approaching the curse of high dimensionality

As explained in section 4.1.4 this case study ABM contains 51 parameters, each having values ranging from -1.00 to 1.00 with steps of 0.01, so 201 potential values for each parameter. Because of this, there are 201^{51} possible parameter set combinations. Considering that one simulation takes around 2 minutes to be completed, it is clear that a basic approach that tries to run every parameter set is not computationally feasible. As the literature previously reviewed, the most common and suitable approaches to reduce the exponential parameter set dimensionality are sensitivity analysis and parameter selection.

Sensitivity analysis requires running an extensive series of simulations, where a subset of parameters is kept fixed to analyse the impact of the remainings. Although the result permits focusing on a significant parameter set, sensitivity analysis is also highly computationally expensive. In this case study, the approach would not work because, indeed, the number of required simulations would also be unfeasible.

As the literature review highlighted, there are no previous studies on human mobility. Models usually consider different factors, so the chances of finding a framework applicable to the case study ABM are null. Instead, the parameter selection method focuses on a weights subset based on the model goal.

4.3.1 Grouping parameter sets into subsets

In this thesis, to tackle RQ2C, an approach that combines the ideas of sensitivity analysis and parameter selection is tested. The method aims at dividing parameter sets into smaller subsets by analysing their correlation. Both strongly negative and strongly positive correlations are used to delimit groups. Indeed, independent parameter groups should theoretically not influence each other. In addition, the grouping of parameters is also done to capture the interaction effects among the group parameters. Since variables are correlated within a group, it is possible to assume that they interact with each other, and as a consequence, they need to be calibrated together, not separately. If no correlation is found, grouping is still possible rationally by organising weights based on the factor nature. For example, environmental factors are expected to be grouped.

Unlike sensitivity analysis, the investigation is not done by running simulations but by studying the correlation of the agents' variables. Hence, this provides a computationally cheap method because the environmental and population calibration datasets already provide values. As a disadvantage, this method implies running the exploration algorithm as many times as the number of variable groups. Nonetheless, such a guided approach could lead to better outcomes in terms of fitness quality and remain computationally feasible.

For this study, a joined dataframe is used by combining the list of agents' displacements with the environmental variables of the original travel cell. A correlation matrix will be generated using Pearson correlation and p-values. In detail, the variables of Table 3 are analysed through a correlation matrix.

As explained in the next section, calibration exploration space algorithms are applied to both the entire parameter space and using this grouping methodology. The calibration comparison metrics defined in section 4.4.6 are used also to determine if this approach to high dimensionality is valuable.

4.3.2 Subsets order

As a consequence of reducing parameter sets into subsets, an order in calibration must be introduced. Intuitively, groups with a higher impact should be calibrated first while keeping the other fixed. To address this issue, a logistic regression model is built on the previous dataframe of all variables of all groups. The outcome model is a binary attribute for the modal choice, indicating the observed modal choice of the displacement. Hence, a regression model is built for each modal choice.

Each coefficient is firstly converted to its absolute value since the aim is to consider the impact strength, not its direction. Next, an average among the three modal choice coefficients is computed. Subsequently, the average of the previous average is calculated for each group emerged by the correlation. With this average coefficient per group, an order is drawn by starting with the group with the higher average impact.

4.3.3 Fixed subsets initial values

Another consequence of introducing subsets is the choice of values for parameter sets that remain fixed during a simulation. Unlike the order of the subset, this is instead a trivial problem. In addition, different parameter sets in a simulation lead to a different outcome, so the choice is of relevance.

If the weights represented absolute values, their initial fixed value could be chosen from the literature review. In this ABM, weights represent the proportional impact on the correlated variable, and no previous study can help with the value selection. Two solutions could be used: 0 as a fixed value for all weights or a random set of values. The first approach appears to make more sense for two reasons. Firstly, assigning the same value to all weights gives an equal impact, which is precisely the ideal situation for a weight that still needs to be calibrated. Secondly, 0 is the mean value for the weights value range [-1.00, 1-00], ensuring a neutral effect. Hence, 0 is used in this study as a fixed initial value for weights in the subgroups that still need to be calibrated in future simulation steps.

4.4 Evaluation of exploration techniques

For this thesis, hill-climbing, simulated annealing, genetic algorithms, the machine learning surrogate model by Lamperti et al. (2018) and BayesFlow by Radev et al. (2020) are evaluated and compared. GAMA already provides an implementation for the first three algorithms. Instead, the other two approaches are provided by the researchers in the form of Python code, and the software integration with GAMA was part of this work. In particular, the communication between the simulation and the external Python implementation was implemented. The latter incorporation requires directly modifying the GAMA software and cannot be distributed in external plugins.

As an experiment, each algorithm is run following the division into subsets explained in section 4.3. Each algorithm is also run on the entire parameter set as a comparison. With this, it is possible to complete RQ2C, knowing if the subsets division also leads to good qualitative outcomes. Section 4.4.6 defines the chosen comparison metric. In addition, since hill-climbing, simulated annealing and genetic algorithm are exploration space algorithms, the weighted-averaged F1 scores of the run parameter sets are plotted. Such analysis permits inspecting the algorithms' exploration behaviours in such a complex parameter space and insights on the surface space shape on a local level.

Most algorithms require settings that directly translate into the required number of ABM simulations, such as training set sizes. Because of time and resources constraints, this value has to be limited in this study, although higher values can clearly lead to better performance. Nevertheless, all algorithms will be subject to the same limitations for an equal evaluation. Also, the presented situation resembles the scenario

for most researchers without High-Performance Computing (HPC). All simulations are run on a machine equipped with a 2.6GHz quad-core Intel "Core i7" 6700HQ CPU and 16GB of RAM for this study. In the following subsections, the hyper-parameters used in each algorithm are described.

4.4.1 Hill climbing

Compared to the other algorithms, hill climbing presents as the simplest in terms of tasks and required hyperparameters. The only required decision is the number of maximum iterations allowed. However, the latter quantity does not directly translate into a maximum number of ABM simulations since the algorithm explores a certain, unknown a priori, number of neighbouring parameter sets. For this study, 50 maximum iterations are chosen.

Hill climbing is not expected to perform well on such a non-linear space full of irregularities. However, it might still result in acceptable outcomes if, by chance, it explores good parameter sets. Also, it is known that the method ends up in local optima because of its quest based on gradients. Hence, the high presence of local optima should result in a relatively short computing time.

The literature review suggests that the algorithm is run several times, and then the best resulting outcome is selected. In this thesis, only three different initialisation are performed because of limited resources.

4.4.2 Simulated annealing

Recalling the literature review, the algorithm is based on hill-climbing, but it adds a probability of accepting a worse parameter set to explore some direction even if they are in principle do not have a positive indication.

The following four hyperparameters are required:

- **Initial temperature**. The literature review suggests, in general, to start with a high value, to initially accept more non-optimal solutions. 2000 is high enough for this study to allow such exploration, and it remains computationally feasible.
- **Final temperature**. Usually, a very small value, 10 is chosen.
- **Temperature decrease**. The percentage value indicates how fast the temperature decreases, resulting in being less open to accepting non-optimal solutions. Values around 75% are often used, so 0.75 is chosen as a value.
- **Number of iterations per temperature**. The number directly translates into a number of ABM simulations per temperature level. Again, because of computational limits, this value is limited to 2.

To summarise, researchers should mostly tune the initial temperature and the number of iterations per temperature to define the total number of simulations. A higher value can lead to a broader exploration and, hopefully, better outcomes.

4.4.3 Genetic algorithm

This method is known among researchers for its trivial hyperparameter tuning. Many studies approached the search via trial and error explorations. Because of the limited time and impossibility of exploring many hyperparameters, a piece of general advice which have been used in many GA implementations is followed (De Jong, 1975).

- **Population dimension**. The area of search should be big enough to represent the scenario, but a higher value leads to an increase of the computational load (Hassanat et al., 2019). Because of known limited resources, the value 10 is chosen.
- **Crossover probability**. 0.7
- **Mutation probability**. 0.1

- **Maximum generations**. 10
- **Stochastic selection**. The roulette selection is chosen because parents are picked based on their fitness. In addition, this method is often used when fitnesses do not differ much, such as in this scenario (Lipowski & Lipowska, 2012).

4.4.4 Machine learning surrogate model

Lamperti et al. (2018)'s idea is to train a surrogate model with a decision-tree-based algorithm that uses the gradient boosting framework XGBoost (T. Chen & Guestrin, 2016). Code implementation is provided at the page https://github.com/amirsani/online_surrogate_modeling. Sobol sequence implementation, differently from the authors, comes from Walz (2021) because the original version is limited to 40 parameters, while the latter can reach up to 1111 dimensions.

Figure 4 illustrates the application of their methodology into this case study. The algorithm starts by letting the researcher set the hyper-parameters, such as set sizes and the budget, which is the total of test set items. Subsequently, the procedure draws a large pool of parameter sets to restrict the problem to a more computationally feasible one. The latter is selected with quasi-random Sobol sampling (Morokoff & Caflisch, 1994), one of the major strengths of this approach as highlighted in the literature review. The ABM runs the training set of parameter sets, and the surrogate XGBoost model learns the resulting objective functions. Next, as a test set, a random out-of-training set is drawn from the pool set. As Lamperti et al. (2018) suggest, the test set size should be equal to the logarithm of the budget (Ross et al., 2011). The ABM is run on these parameter sets, fed to the XGBoost model to predict the objective functions and evaluated in terms of MSE. The latter evaluation metric is suggested by Lamperti et al. (2018) when dealing with such continuous space. MSE and budget are used as stopping criteria. If the latter conditions are not met, the algorithm starts again by sampling another training set. Otherwise, it predicts objective functions over the large pool set and returns the best parameter set.

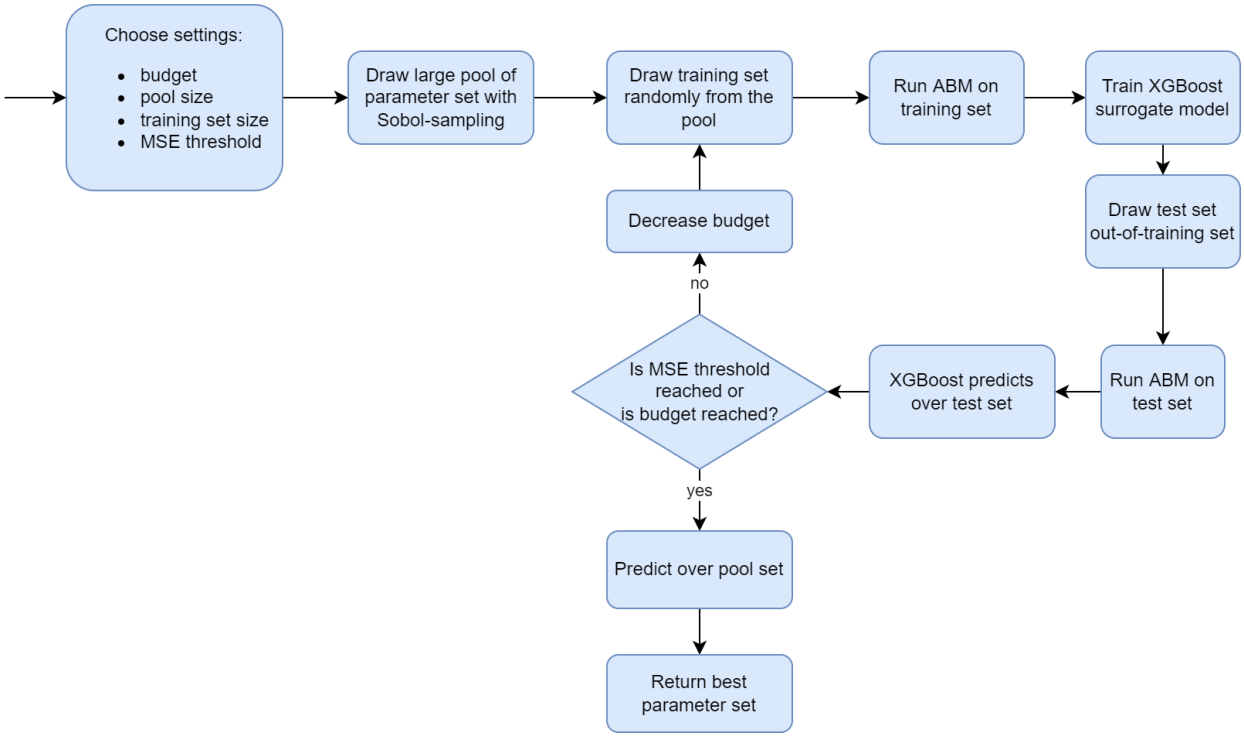


Figure 4: Implementation of the surrogate modelling approach by Lamperti et al. (2018) in this thesis.

The learning set sizes and acceptancy thresholds are crucial here. These are the considerations that should be taken into account and the values I choose for this study:

- **Pool set size**. The overall set is drawn with a quasi-random Sobol sampling technique to make the parameter space search more approachable but still representative. Since these parameter combinations will not be generated and simulated by the ABM, there is no problem setting this setting to a high value. For this case study, 100000 is chosen as a value.
- **Training set size**. Finding a good compromise between representativeness and feasibility is trivial when dealing with a complex and computationally expensive ABM. Indeed, this quantity directly translates to a number of ABM simulations that need to be run to calculate the respective objective function, then learned as a true label from the XGBoost surrogate model. The decision is even more trivial because if the MSE threshold is not reached in the first iteration, the same quantity of simulations has to be run again. For this case study, a training size of 20 is chosen because of time limitations. In the case of running further training iterations, the overall time would still be acceptable.
- **Budget**. This represents the number of total test items to evaluate the model through MSE. Again, this directly translates into a number of simulations required for the ABM. Because of time constraints, the value is limited to 20.
- **MSE threshold**. This value describes the second early-stop condition. The amount strictly depends on the objective function values and how accurate prediction is expected. The aim is to minimise MSE, getting close to 0. Since the goal is to reach a precise weighted-averaged F1 score, 0.1 as the MSE threshold is chosen.

4.4.5 Bayesian optimisation

BayesFlow (Radev et al., 2020) implementation is provided at the page https://github.com/stefanradev93/BayesFlow. Figure 5 illustrates the incorporation into this thesis. The overall idea is to train the cINN given a training dataset of simulations. The authors provide both online and offline training methods. In the former, data points are generated at every training iteration, while in the latter, the training phase goes through a pre-existing dataset. Since the authors do not specify how to sample points, offline training is here applied to a dataset generated with the quasi-random Sobol sampling method from Lamperti et al. (2018). In other words, the training points are sampled from a uniform prior distribution. Once the summary network is converged, it predicts a posterior distribution of parameter sets, given a distribution of objective functions. In this study, the latter is a discrete (steps of 0.001) uniform distribution between 0 and 1 of weighted-averaged F1 scores. Finally, it is possible to extract the best parameter set with the maximum likelihood.

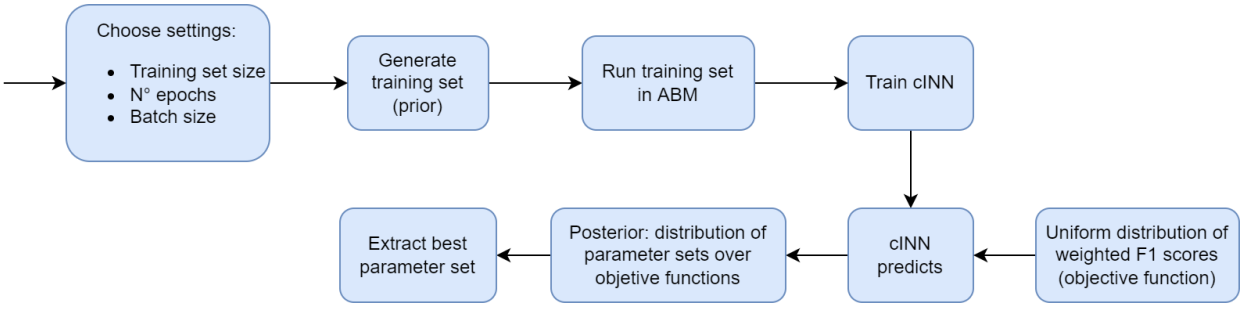


Figure 5: Implementation of the likelihood-free Bayesian approach by Radev et al. (2020) in this thesis.

A big advantage of using this approach is that it does not require many hyperparameters to be tuned. In detail:

- **Train set size**. Similarly to Lamperti et al. (2018), both solutions involve training a network, and it is clear how the quantity of entry points determines the network prediction ability. Radev et al. (2020) and Shiono (2021) highlight how this quantity depends on each singular case study. Again, the size of the training sets remains a trivial point because it translates into a required number of ABM simulations. Based on time constraints, the value 50 is chosen.
- **Number of epochs**. The value depends on the convergence of the summary network. 30 is chosen as a value.
- **Batch size**. It consists of the number of simulations used together in a training iteration. In general machine learning cases, the size might be set to a high value when many points need to be trained. Here, the training set does not consist of many data points; hence the batch size is set to 4.

4.4.6 Comparison metrics

The described comparison is the object of RQ3B. For such evaluation, I define the following metrics:

- **Weighted-averaged F1 score** (ABM objective function). It is used to give a general score on the model classification ability.
- **Individual F1 scores** per modal choice classes. It is used to dive into the details of which classes are best predicted. Combining the results with the McFadden pseudo R^2 score of the preliminary work, it is also possible to assess the power and correctness of the ABM.
- **Computational time**, expressed in in minutes. It is the total measured calibration execution time.
- **Number of ABM simulations**. The number is often directly proportional to the computational time, but some algorithms also incorporate one-time training phases. In addition, future works will probably have different simulation duration due to different models or more powerful machines.

Hence, the thesis aims to give an overview of how the different techniques perform on the Amsterdam case study, from a perspective of how well the model predictions are and the required resources. Finally, to select the best optimisation algorithm for this case study (RQ4), a trade-off of evaluation measures will be necessary, taking into account all various aspects.

5 Results

5.1 Division in sub-parameter sets

Firstly, the division in sub-parameter sets is performed. The correlation matrix of figure 6 highlights how there is a strong and significant correlation between the most of the environmental factors (all excluded the precipitation):

- Pedestrian accidents
- Retail density
- Distance to CBD
- Highway length
- Road Intersection density
- Population density
- Green cover. density
- Public transport density

Both strongly positive and strongly negative correlations are relevant for the grouping because they both highlight a influence on each other. For example, the negative correlation of distance to CBS and the highway total length means that the further from a business center, the smaller the total highway length. This result is in line with expectations and also returns an indicator of data quality. Indeed, a higher total highway length is likely in the outer part of the city, far away from CBDs, because of more rural areas.

Using the correlation information, the set of 21 factors is split into two subgroups as in table 4: non-environmental (13 factors) and environmental weights (8 factors). This result is consistent with expectations since it makes sense that factors of the same nature are correlated.

Secondly, a logistic regression model is computed. The same table 4 illustrates the resulting coefficients. An average value is firstly determined between the three modal choices coefficients. Next, an average value (group. avg.) is computed, averaging the latter outcome between the factors of the same group. Results suggest how the non-environmental group (average coefficient 1.17) has to be calibrated first, and the environmental one subsequently (average coefficient 1.03). As table 5 clarifies, this division translates into 38 weights for the first group and 13 for the second during the simulations.

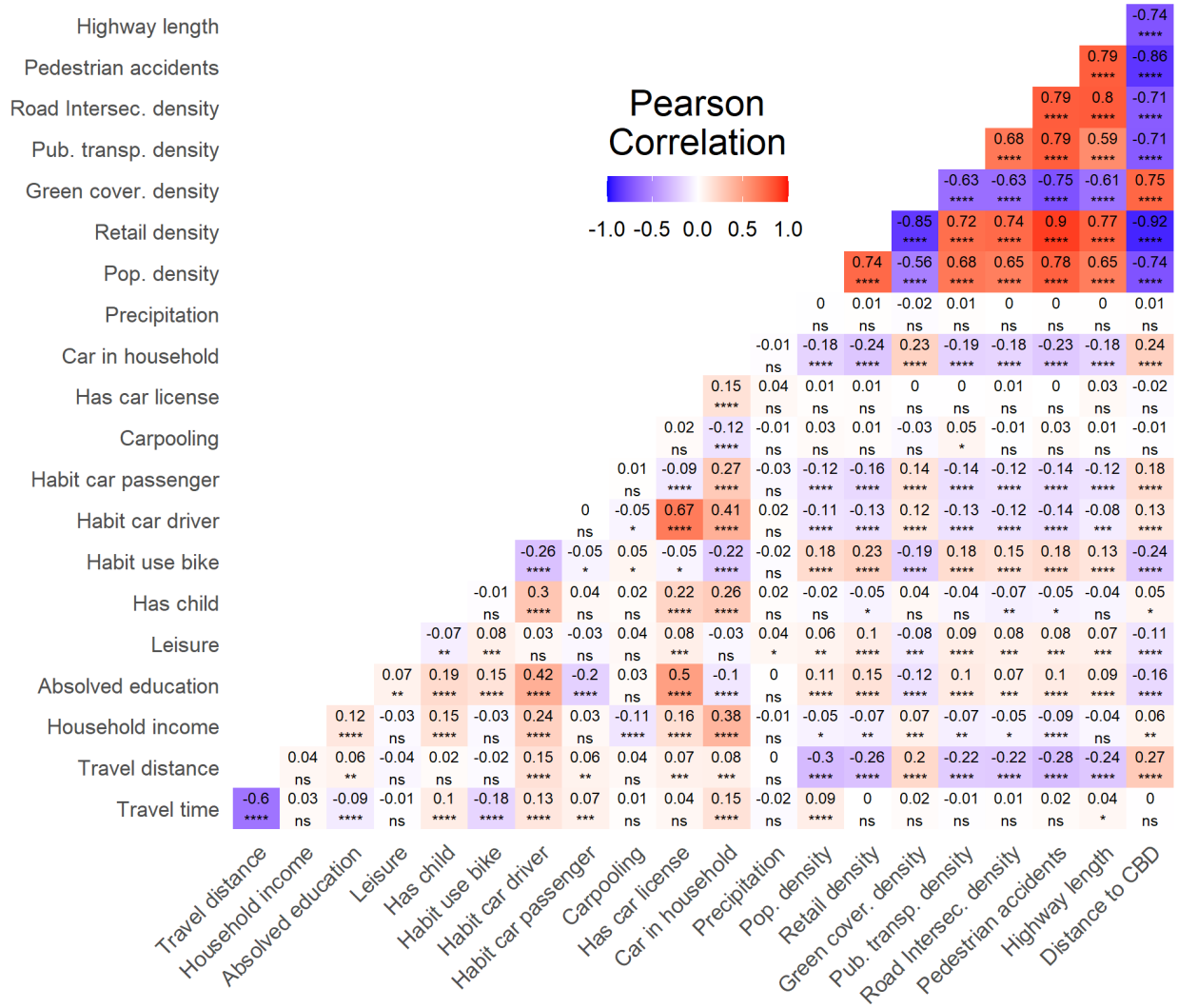


Figure 6: Correlation matrix of factors that play a role in the agent's utility. P-value significance is denoted with asterisks. '****' corresponds to p-values lower than 0.0001, '***' between 0.0001 and 0.001, '**' between 0.001 and 0.01, while '*' higher than 0.05. It is evident how the environmental factors in the top right corner are highly and significantly correlated among each other.

Factors	**Coef. walk**	**Coef. bike**	**Coef. car**	**Avg. coef.**	**Group**	**Group avg.**
Travel distance	2.93	-4.56	-4.31	3.93	non-env.	1.17
Travel time (based on modal choice)	0.00	0.00	0.00	0.00	non-env.	1.17
Habit car driver	-0.84	-2.18	3.61	2.21	non-env.	1.17
Has car license	-0.06	0.21	-0.39	0.22	non-env.	1.17
Habit use bike	-2.45	4.92	-2.52	3.30	non-env.	1.17
Carpooling	-1.00	-1.36	2.72	1.69	non-env.	1.17
Habit car passenger	-0.41	-0.96	1.70	1.02	non-env.	1.17
Precipitation	-0.87	0.89	-0.84	0.87	non-env.	1.17
Absolved education	0.46	0.67	-1.38	0.84	non-env.	1.17
Leisure	0.73	-0.44	-0.17	0.45	non-env.	1.17
Car in household	0.11	-0.34	0.81	0.42	non-env.	1.17
Has child	-0.49	0.12	0.14	0.25	non-env.	1.17
Household income	0.10	-0.03	0.08	0.07	non-env.	1.17
Pedestrian accidents	3.82	-2.49	-0.29	2.20	env.	1.03
Retail density	-2.24	1.82	-1.27	1.77	env.	1.03
Distance to CBD	0.17	-2.13	1.96	1.42	env.	1.03
Highway length	1.71	-1.64	0.12	1.16	env.	1.03
Road Intersec. density	-1.18	0.77	-0.13	0.69	env.	1.03
Population density	-0.18	-0.51	1.20	0.63	env.	1.03
Green cover. density	0.17	-0.28	0.17	0.21	env.	1.03
Public transp. density	0.19	-0.03	-0.14	0.12	env.	1.03

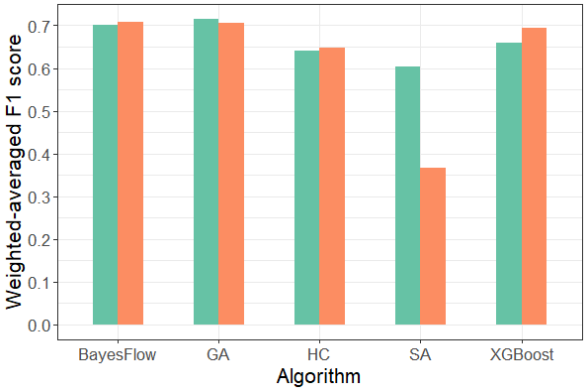
Table 4: Standardised coefficients for each modal choice resulting from the binary logistic regression. Averages among the three modal choices (avg. coef.) are calculated. Based on the two groups defined by the correlation matrix (figure 6), group averages are calculated and delineate a group order.

Non-environmental (1st group)	Environmental (2nd group)
w_time['walk']	w_population_density_walk
w_time['bike']	w_retail_density[walk]
w_time['car']	w_green_coverage[walk]
w_distance['walk', age_group=='minor']	w_public_transport_density_walk
w_distance['walk', age_group=='teenager']	w_road_intersection_density['walk']
w_distance['walk', age_group=='youngadult']	w_pedestrian_accidents['walk']
w_distance['walk', age_group=='adult']	w_retail_density['bike']
w_distance['walk', age_group=='senior']	w_green_coverage['bike']
w_distance['walk', age_group=='elderly']	w_road_intersection_density['bike']
w_distance['walk', BMI_group=='normal']	w_pedestrian_accidents['bike']
w_distance['walk', BMI_group=='overweight']	w_distance_CBD['bike']
w_distance['bike', age_group=='minor']	w_highway_length_car
w_distance['bike', age_group=='teenager']	w_distance_CBD['car']
w_distance['bike', age_group=='youngadult']	
w_distance['bike', age_group=='adult']	
w_distance['bike', age_group=='senior']	
w_distance['bike', age_group=='elderly']	
w_distance['bike', BMI_group=='normal']	
w_distance['bike', BMI_group=='overweight']	
w_distance_car	
w_precipitation['walk']	
w_precipitation['bike']	
w_precipitation['car']	
w_leisure	
w_child['walk']	
w_habit_use_bike['walk']	
w_household_income['walk']	
w_education['bike']	
w_habit_use_bike['bike']	
w_income['bike']	
w_car_household	
w_carpooling	
w_car_license	
w_habit_car_driver	
w_habit_car_passenger	
w_has_child['car']	
w_absolved_education['car']	
w_household_income['car']	

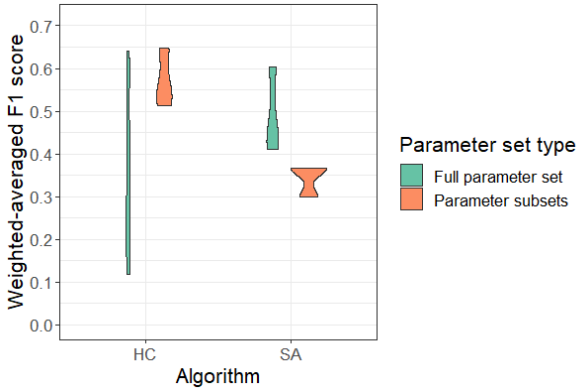
Table 5: Division in two subgroups emerged by the factors correlation matrix and ordered by the logistic regression analysis.

5.2 Evaluation and comparison of exploration algorithms performances

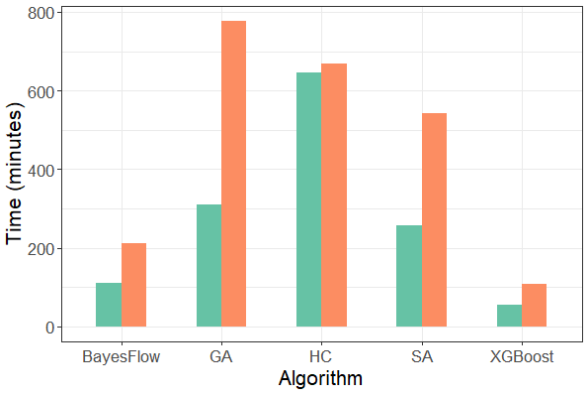
Each of the five algorithms is run to find an optimal parameter set. Local search algorithms, namely hill climbing and simulated annealing, are run three times to tackle their known local optimum problem. The other approaches are run only once. Hence, hill climbing and simulated annealing showed values for the weighted-averaged F1 score are the best-obtained score among the three initialisation. Instead, their time and number of simulations correspond to the sum of the three executions. Weighted-averaged F1 score, total time and the number of simulations are shown in figure 7. Metrics are also divided by calibrating on the entire parameter space or using the subsets division emerged by the previous results of section 5.1.



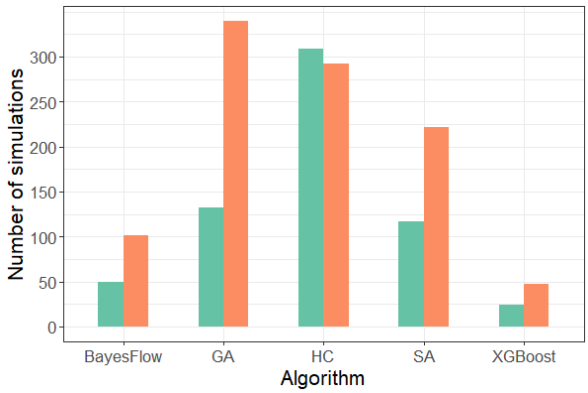
(a) Weighted-averaged F1 scores. Hill climbing and simulated annealing scores correspond to the best obtained value during their three different initialisations.



(b) Hill climbing and simulated annealing distribution of the weighted-averaged F1 scores of the three different initialisations. Hill climbing shows a large variance, returning three highly divergent values (0.32, 0.12 and 0.64). By contrast, the other algorithms have more unified scores.



(c) Total measured time comparison. Hill climbing and simulated annealing values are the sum of the total time across the three different initialisations.

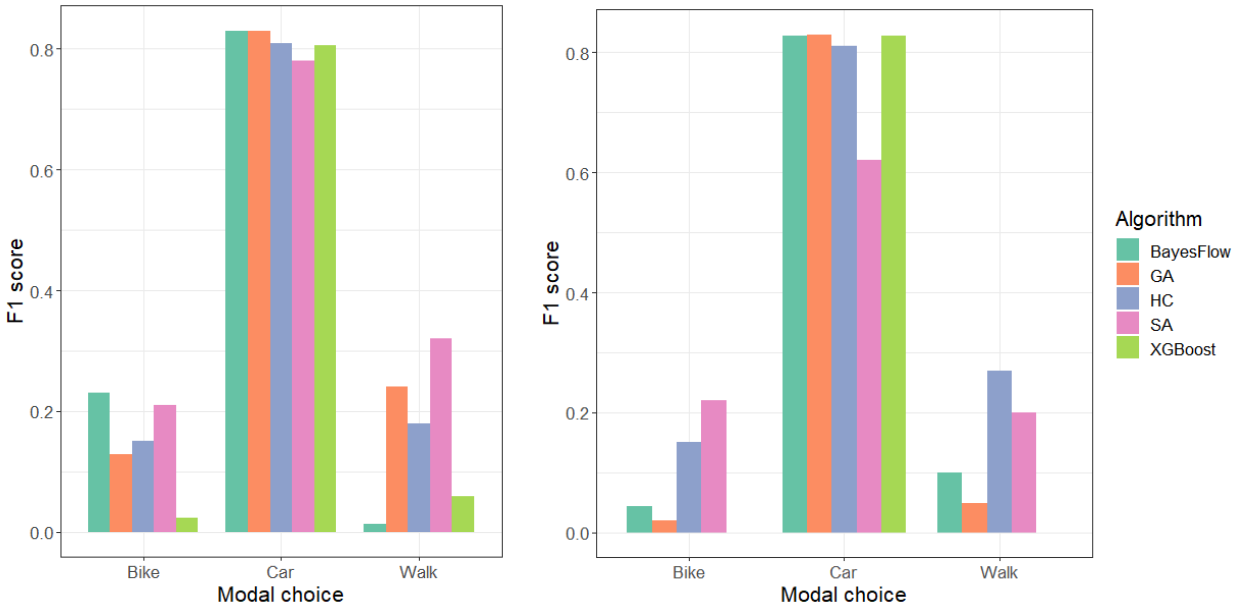


(d) Total number of ABM simulations required by each algorithm. Hill climbing and simulated annealing values are the sum of the total time across the three different initialisations.

Figure 7: Comparison of weighted-averaged F1 score, time and number of simulations. Values are grouped by the presented algorithms, divided by calibrating on the entire parameter set (red) and using the division into subsets (green). "XGBoost" is used to label the machine learning surrogate model approach of Lamperti et al. (2018), while "BayesFlow" the Bayesian surrogate model approach of Radev et al. (2020).

5.3 Individual modal choice classes

The examination of the individual F1 score per each modal choice class is depicted in figure 8.



(a) Calibration on the entire parameter space.

(b) Calibration using division in groups.

Figure 8: Comparison of F1 scores over the three individual modal choice classes. Results are divided by calibrating on the entire parameter set (left) and using the division into subsets (right). Again, hill climbing and simulated annealing results correspond to the best-obtained parameter set among the three initialisation.

5.4 Correlation of found optimal parameter sets with regression coefficients

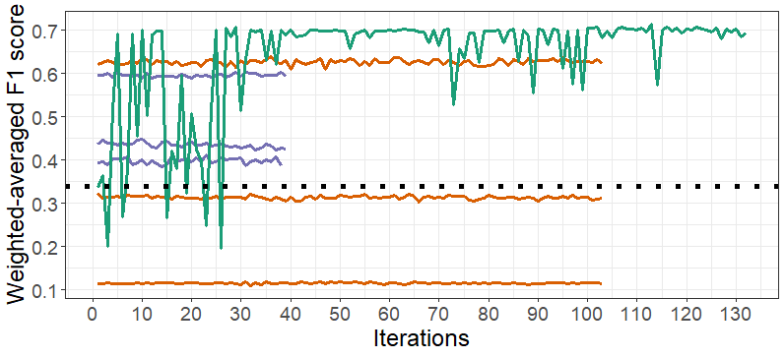
Table 6 shows the correlation between the optimal parameter set found by each algorithm and the logistic regression coefficients of table 4. Such analysis permits to inpect if the predicted parameter set is consistent with the direction given by the logistic regression analysis. Results do not highlight any correlation.

Algorithm	Parameter set type	Pearson correlation score	P-value
Hill climbing	Full parameter set	0.01	0.962
	Parameter subsets	0.30	0.084
Simulated annealing	Full parameter set	0.23	0.182
	Parameter subsets	-0.10	0.551
Genetic algorithm	Full parameter set	0.02	0.914
	Parameter subsets	0.06	0.715
Machine learning (XGBoost)	Full parameter set	-0.01	0.914
	Parameter subsets	0.06	0.711
Bayesian optimisation (BayesFlow)	Full parameter set	-0.02	0.887
	Parameter subsets	0.30	0.073

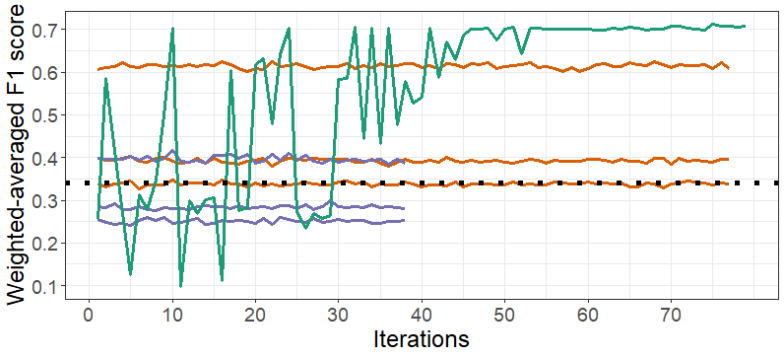
Table 6: Correlation analysis by computing a Pearson correlation score between the optimal parameter set found by each exploration algorithm and the coefficients returned by the logistic regression of table 4.

5.5 Comparison of local search trends with GA

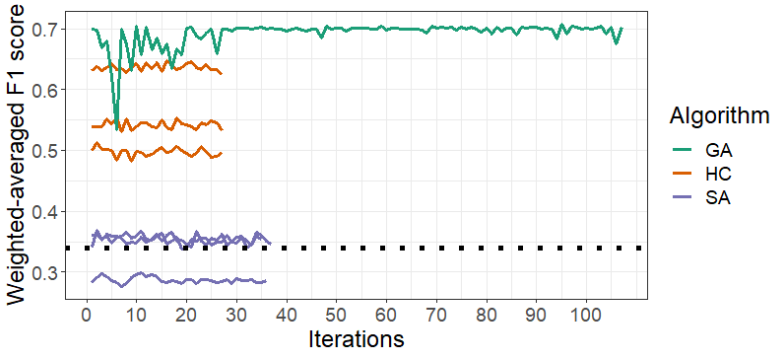
Focusing on the exploration algorithms that do not involve the learning of a surrogate model, namely hill climbing, simulated annealing and genetic algorithm, their exploration trend is plotted in figure 9. A constant trend is individuated in local search algorithms, while genetic algorithm highlights a converging behaviour to a better solution. Again, results are divided by calibrating on the entire parameter space or using the subsets division emerged by the previous results of section 5.1.



(a) Calibration on the entire parameter space.



(b) Calibration on the first subgroup.



(c) Calibration on the second subgroup.

Figure 9: Weighted-averaged F1 score exploration when simulating with hill climbing (HC), simulated annealing (SA) and genetic algorithm (GA). Each trend point corresponds to the weighted-averaged F1 score of a simulation using a parameter set explored by the algorithm. The black dashed line highlights the average of weighted-averaged F1 scores (0.34) recorded by sampling 50 parameter sets using the quasi-random Sobol technique. The latter gives a threshold to compare the goodness of algorithms with a baseline of a completely random approach.

6 Discussion

In the field of agent-based modelling, new studies focus on simulating human behaviours to analyse behaviour-environment interactions (Maggi & Vallino, 2016). Within the latter, urban mobility plays a significant role (Lu et al., 2018). Models need to be calibrated to correctly represent human behaviour since the general structure is usually insufficiently precise (Crooks et al., 2018). Calibration is one of the most significant issues in ABM due to its high-dimensionality and hierarchical structures. There are several approaches to calibration; however, they are primarily dependent on the model purpose. It is clear that how calibration should be performed depends on the model goal. Hence, there is a lack of understanding of what makes a method suitable specifically for mobility behaviour. Because of that, this study analyses what makes calibration methods suitable for human mobility ABM. In addition, how the simulation and the calibration process should be performed is investigated by using the presented EXPANSE Amsterdam case study as a data-driven approach.

6.1 RQ1: Suitable calibration methods for ABM of mobility behaviour

The RQ1 requires an analysis of behavioural models and calibration methods characteristics highlighted in the following chapters and a final mapping in paragraph 6.1.3.

6.1.1 RQ1A: Characteristics of behavioural models

Analysing the presented model and running simulations, it is possible to derive and confirm literature review characteristics of more general ABM for human mobility behaviour.

Firstly, it is clear that such models contain a vast parameter space: 201^{51} in this model implementation. Looking at the iteration plot of the local search algorithms (figure 9), it appears that neighbouring parameter sets return similar, if not almost identical, fitness values. This behaviour suggests that the surface space is homogeneous on a local scale. Nonetheless, it is not possible to conclude at a higher level. The same figure returns weighted-averaged F1 scores ranging from 0.2 to 0.7. Hence, the parameter space is still highly irregular and unpredictable.

Secondly, previous studies reported non-linear behaviour (Chapizanis et al., 2021), primarily because of human social contagion. In this study, it was not possible to measure such characteristics since no agent interaction was implemented. Future developments of the model aim at implementing such a feature.

Another critical point highlighted from the human mobility behaviour literature is the heterogeneity of individuals (Chapizanis et al. (2021); Sonnenschein et al. (2021)). The case study model implements this population differentiation, incorporating heterogeneous weights for the travel distance willingness based on groupage and BMI. The current model design is just a starting point, and heterogeneity can be expanded by including other social classes, for example, income groups, to affect affordability factors. The heterogeneity groups choice should be driven by behavioural theories, although there is the necessity of deepening the current state of the research. However, looking at the computational side, the consequence is an exponential increase in the number of weights and the overall complexity. Hence, a compromise should always be taken into account.

Behavioural models could be based on different types of decision-making functions. For this proof-of-concept, a linear function model was chosen. This approach permits building regression models and comparing their coefficients with the found parameter sets. The advantage is that the algorithms' performance differences can be attributed entirely to the methods and not the model. Results of such correlation (table 4) did not highlight any link to a theoretical optimal parameter set. Combining the latter with the weighted-averaged F1 scores of figure 7a, it is possible that no algorithm found a truly optimal parameter set. Indeed, several high scores might confuse the algorithms in the search. Also, as aforementioned, the parameter space is very vast, and model bottlenecks could be present.

Finally, the use of individual scores for each class has the advantage of a possible correlation with pre-analysis in the case of a linear model. In particular, multiple regression models were built, where the targeted mean of transport was predicted (independent variable) with all the selected attributes. McFadden pseudo-R^2 approximates the amount of variation that was explained to the ABM. Scores of 0.35 for walking,

0.33 for biking and 0.44 for driving were found. Manually comparing these results with the ones of figure 8, a correspondence can be identified, with the driving modal choice having a higher presence in both investigations. This means that the pre-model analysis is able to give a preview of the ABM behaviour.

6.1.2 RQ1B: Characteristics of calibration exploration methods

The analysed state-of-the-art exploration techniques can be differentiated in algorithms with a local search and a non-local approach.

Local search algorithms (HC and SA)

The iterations plots (figure 9) demonstrate how pursuing neighbouring parameter sets does not bring any benefit. The simulated annealing temperature characteristic was presented as an improvement over hill climbing, theoretically exploring broader areas. However, no improvement over hill climbing was obtained, probably because of such a vast space. For future studies, it could be valuable to analyse different temperature increases since the proposed hyperparameters might end up in too small temperature changes that do not permit a sufficiently wide exploration.

In addition, local searchers require substantial time (figure 7c) and several simulations (figure 7d). Especially in hill climbing, even just one neighourhood is large enough to be computationally expensive.

The literature review mentions that both algorithms usually end up in local optima (Beven, 2002), and the low-compared recorded results of figure 7a confirm the previous studies. Initialising the method three times did not bring any substantial benefit. Looking at the distribution of the different initialisations (figure 7b), results even highlight a high variance, making especially hill climbing very unstable.

Focusing on the best weighted-averaged F1 score found (figure 7a), hill climbing performs quite well, while simulated annealing returns lower values, especially when calibrating on subgroups. The latter outcome might be caused by a mediocre parameter set found for the first group (non-environmental) that affected the performances on the final group, as suggested by figure 9c. However, the result could also be due to bad luck, with poor explored neighbourhoods in all three initialisations.

Non-local search algorithms (GA, XGBoost, BayesFlow)

By contrast, non-local search algorithms are genetic algorithm, the machine learning surrogate model by Lamperti et al. (2018) and Bayesian optimisation "BayesFlow" by Radev et al. (2020). All techniques sample the surface area in a non-localised manner and get a broader picture of the space. The genetic algorithm randomly matches solutions driven by the mutation probability, while the other two approaches train a surrogate model over a sampled training set.

All algorithms return very high and similar fitness as highlighted in plot 7a. While it is not possible to draw preferences on the quality side, the machine learning surrogate and BayesFlow perform considerably better than genetic algorithm on the computational load side (figure 7c and figure 7d).

Overall, the surface space is so vast and non-linear that more simulations would be required to draw more precise conclusions. However, results already show considerable improvements of non-local searchers over local searches. The former reach 0.7 fitness points, while the local searchers appear limited to 0.6 points.

6.1.3 Mapping between behavioural models and calibration methods

Considering the in-depth analysis of the two previous chapters, the next match can be established between calibration methods applied to human mobility behaviour ABM.

Because of the non-linear and vast space, local searches algorithms are not sufficient. Several initialisations are necessary to find an acceptable fitness, although demanding unsustainable computational loads. In addition, results supported the literature review about the critical problem of ending up in a local optimum.

Non-local search approaches, instead, appear promising. The genetic algorithm still highlights long required simulation times, and the complex hyperparameters tuning still leaves challenges for further studies. By contrast, the machine learning surrogate by Lamperti et al. (2018) and the Bayesian optimisation "BayesFlow" by Radev et al. (2020) demonstrate a good balance of quality fitness measure and computational load. It is not easy to prefer one based on the obtained results only. However, this research already gives the direction to continue with these preferred methods for future work.

6.2 RQ2: Appropriate experimental calibration setting for simulations

The literature review highlighted several critical points when calibrating, particularly the presence of hierarchical levels, the difficulty of defining an objective function, and the curse of high parameter dimensionality (Crooks et al. (2018); Beven (2002); Railsback (2001)). The analysis of the employed ABM can be used as a data-driven approach to establish a more general framework of solutions for ABM of human mobility behaviour.

6.2.1 RQ2A: Hierarchical levels

Previous research highlighted the general advice of incorporating different spatial levels (Crooks et al., 2018), usually using multiple objective functions (Beven, 2002).

In this case study, the aim is to calibrate the individual behaviour; hence the level should be individual only, looking at differences in modal choice decision making of each agent. There is no necessity to incorporate different spatial levels when focusing on the modal choice only. Other model goals could be, for example, to calibrate the average mobility behaviour, perhaps in a particular area. In that case, the fitness could be captured by averaging differences in behaviours over the population in that area or at some flow key locations, such as train stations (Sonnenschein et al., 2021).

6.2.2 RQ2B: Objective function

As detailed explained in section 4.2.4 the modal choice prediction scenario translates into a classification problem with imbalanced classes. In detail, the ABM predicts the modal choice class for a given displacement, and the calibration process compares the prediction with observed calibration data. Hence, it appears logical of using a weighted-averaged F1 score, based on the observed modal choice class proportions. In addition, looking at the algorithm comparisons of figure 7a, the model results in different fitness values, returning a properly functional and usable metric that allows discrimination in positive and negative performances.

6.2.3 RQ2C: Reducing parameter dimensions

Usually, modelling design should meet computational feasibilities. However, if the aim is to analyse how urban interventions affect behavioural changes in future studies, the parameter reduction might affect the model goal. Sensitivity analysis or parameter selection is not possible either because of the unfeasible computational load in the former and the lack of previous frameworks in the latter. In this project, an approach for ABM with linear functions by grouping and ordering weights based on their impact on the modal choice is tested as detailed explained in section 4.3.

Looking at the fitness results of figure 7a, no improvement is detected in contrast to directly calibrating on the entire parameter space. It is arduous to conclude because the used model might influence the results. For example, higher weighted-averaged F1 scores of around 0.7 points could never be reached; hence both grouping and non-grouping approaches could have reached a model bottleneck.

Moving to the computational load, both times and the number of iterations are doubled when using the grouping approach, except for hill climbing. A general increase is expected because having two groups to calibrate translates into running each algorithm twice. More groups would imply an even higher computational load. In addition, it is not recommended to decrease tuning hyperparameters of exploration algorithms to achieve faster executions because the parameter space of a subgroup is still consistently large to require enough training sets or explorations. Contrarily to all other algorithms, hill climbing records almost identical execution times and iterations because having a smaller parameter set to calibrate results in fewer neighbourhood explorations.

Taking into account the iteration plots of figure 9, simulated annealing performs worse than hill climbing with the grouping approach. This behaviour is unexpected since simulated annealing should improve hill climbing thanks to its temperature strategy. The result can be due to randomness in such small different initialisations or a poor first subgroup found (figure 9b). Also, the temperature hyperparameter might be too small and make the simulating annealing behaviour converge to hill climbing.

Considering such results, it appears that the grouping methodology does not solve the curse of the high dimensionality problem for this case study and similar human mobility ABM. Also, it is intuitive to believe that grouping leads to missing information on the fixed group while calibrating. Indeed, the problem of the initial value for the fixed group (section 4.3.3) is non-trivial.

Another approach to the grouping methodology could be that given the reduction in model complexity, an increase in decimal digits of the weights can lead to higher model precision. Nonetheless, the still critically high parameter dimensionality does not suggest further investigations in this direction.

Different techniques can be used to tackle the parameter dimension in future studies. The presented correlation of figure 6 can be used as an indicator for a factor analysis. Hence, it can be possible to reduce and abstract a smaller parameter set that still expresses the same piece of information.

Finally, cutting-edge technologies such as the employment of autoencoder neural networks could drastically reduce the computational complexity of models (Rizvi et al., 2018). The idea resides in learning a network that is able to abstract the weights information and reduce it to a smaller parameter set. The ABM is calibrated on the resized set, and the values of the weights are then translated into the full parameter set.

6.3 RQ3: Appropriate experimental setting for comparing calibration methods

6.3.1 RQ3A: Suitable hyperparameters for presented calibration methods

This thesis aims to give an overall direction of what range of hyperparameters should be used by experimenting with one choice set per algorithm. It is clear that hyperparameters highly influence the performances of the exploration algorithms. However, tuning hyperparameters requires extensive procedures, often based on trial and error approaches (Hutter et al., 2015). The main strategy is to have explorations or training sets large enough to capture an overview of the surface space. However, limited time resources and the necessity of applying multiple algorithms for comparison influenced the hyperparameters choices in this study. Once a suitable exploration method is determined, future research should focus on tuning the hyperparameters.

- **Hill climbing.** The number of 50 maximum iterations is always reached in the presented experiments; hence a higher threshold should be used to avoid early stopping and allow more exploration. Nonetheless, as described in section 6.1.2, it appears reasonable to believe that even a higher threshold would not help because of such a complex surface space.

- **Simulated annealing.** Looking at the iteration plots of figure 9, it appears that the temperature hyperparameter is not sufficient to skip neighbourhoods enough to explore areas wider than hill climbing. Hence, a smaller value is recommended for further studies.

- **Genetic algorithm.** Results using a standard crossover of 0.7 and mutation of 0.1 appear promising (De Jong, 1975). Nonetheless, studies could investigate the use of larger populations and iterations to allow a wider collection of data points.

- **Machine learning surrogate** by Lamperti et al. (2018) and **BayesFlow** by Radev et al. (2020). The two algorithms are grouped together because they share the same surrogate model approach, and they both lead to similar results. Further research should focus on the use of these methodologies by sampling larger training sets.

6.3.2 RQ3B: Comparison metrics to compare calibration methods

It is clear from the literature that the main critical aspects in calibration are the fitness quality and the computational load. Hence, the comparison metrics should cover these two main aspects. The weighted-averaged F1 score and individual F1 scores are used as fitness quality measures in the carried out experiments, while the time and number of ABM simulations are used as computational load.

The chosen objective function should be the primary quality metric since it drives the calibration algorithm and expresses the ABM prediction aspect. The weighted-averaged F1 score is used in this case study

since the ABM corresponds to a classification problem with unbalanced classes. In addition, as aforementioned in chapter 6.1.1, the individual F1 scores can be a support tool for the researchers returning hints for further developments.

When calibrating such complex models, the computational metrics underline a proportion between time and number of simulations. This does not come as a surprise since the only discrepancy could happen in machine-learning algorithms where a training and inference phase might take time. However, the latter time interval appears to be irrelevant over the large ABM simulation total time. Individual surrogate model training phases never took longer than 1 minute, and this time can even be improved by using a GPU. Hence, if the researchers work with such a model where it is possible to estimate an average simulation time, the total time can be inferred by multiplying the number of simulations.

Finally, the computational load expressed by the literature review can also be seen in terms of used memory resources. In this work, it was not possible to measure this aspect. However, this aspect should not be particularly problematic for any algorithms. Exploration space algorithms do not contain any particular data structure, while more sophisticated machine-learning approaches are used to deal with big data, consistently larger than a few ABM simulation data points. When simulating, the most prominent memory bottleneck still resides in the ABM itself, with an increase in memory load due to the number of agents or details. However, this is not a problem of calibration techniques themselves, so it cannot be listed as a metric.

6.4 RQ4: Preferred calibration method for the Amsterdam case study

All things considered, the results of the presented grouping parameter reduction method do not suggest its use as it is now. Hence, it is still preferred to calibrate the entire parameter set than use the subgrouping approach. The related discussion chapter 6.2.3 highlights how the current state-of-the-art still needs improvements, such as a heavier parameter reduction via factor analysis.

Looking at the exploration algorithms, the surrogate model-based approaches have been proven to outperform the other presented techniques (see chapter 6.1.2). Also, the genetic algorithm recorded good fitness performances. Nonetheless, the high computational times and complex hyperparameters tuning shifts the overall choice towards the surrogate machine learning approach by Lamperti et al. (2018) and Bayesflow by Radev et al. (2020). As aforementioned, this study could not draw a final winner between the latter two as they equally performed. Also, their similar characteristics make it difficult to base the choice on the theoretical side.

6.5 Conclusion

This study addressed the lack of understanding of what makes a calibration method suitable specifically for mobility behaviour ABM. As the issue could not be addressed by theory only, a data-driven approach was used by calibrating the proof-of-concept model of the EXPANSE project. An Amsterdam case study was selected using the ODiN dataset.

This research analysed the model characteristics, demonstrating the critical points of high parameter dimensionality and computational load of the literature review. In addition, an appropriate experimental calibration framework is presented, giving a clear overview of what appropriate objective function and hierarchical level should be used in the field of human mobility.

Besides, a comparison of cutting-edge optimisation algorithms for the optimal parameter search is delivered. Hill climbing, simulated annealing, genetic algorithm, two surrogate model approaches based on machine learning and Bayesian optimisation are evaluated. Results hinted at pursuing the research with surrogate model-based methodologies. Their performances outperform the other solutions on both quality and computational load aspects. In addition, a metrics-based framework of how to compare calibration techniques for human mobility ABM is presented.

The investigation also proved that a parameter dimensionality reduction method based on grouping does not bring any benefit compared to calibrating the entire parameter set.

This project is just the first step to address the knowledge gap of calibration in the human mobility ABM field. Future studies should extend this research by analysing different techniques of parameter reduction such as factor analysis and experiment algorithms hyperparameters tuning. At the same time, the long calibration times are one of the main research bottlenecks. Hence, there is the necessity for computational improvements such as CPU parallelisation and the use of high-performance computing. Finally, the ODiN dataset contained common drawbacks of human mobility data, such as the lack of past behaviour data and high spatial resolution. For example, the presented work was limited to a high level PC4 spatial resolution. Furthermore, walkability and bikeability information would increase the model prediction ability. Mobility research has been sharply increasing in recent years, and there is hope that a more refined data collection will be performed.

The results are valuable, not only to the transport choice field only but generic to the main idea of mode choice. For example, other agents' choices are about selecting the next activity, location, destination or route. This will have considerable benefits for behavioural models, such as the EXPANSE one, as mode choice is at the base of all the components it is composed of. Additionally, the established method will be applicable to other cities as case studies. Overall, the framework will be a profitable tool for all the researchers to be used in urban developments for a sustainable and healthier future (L. Chen, 2012).

Acronyms

ABM Agent-Based Modelling.

BDI Belief-Desire-Intention.

CBD Central Business District. Commercial and business center, such as commercial space and offices.

CBS Centraal Bureau voor de Statistiek. National Dutch statistics centre.

GA Genetic Algorithm.

HC Hill Climbing.

MSE Mean Squared Error.

SA Simulated Annealing.

References

Agnihotri, A., & Batra, N. (2020). Exploring bayesian optimization. *Distill*, *5*. https://doi.org/10.23915/distill.00026

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models [Modeling Human Decisions]. *Ecological Modelling*, *229*, 25–36. https://doi.org/10.1016/j.ecolmodel.2011.07.010

Baeyens, E., Herreros, A., & Perán, J. R. (2016). A direct search algorithm for global optimization. *Algorithms*, *9*(2). https://doi.org/10.3390/a9020040

Balać, M., & Hörl, S. (2021). Simulation of intermodal shared mobility in the san francisco bay using matsim. *Arbeitsberichte Verkehrs- und Raumplanung*, *1618*. https://doi.org/10.3929/ethz-b-000481951

Batty, M. (2013). *The new science of cities*. The MIT Press. http://www.jstor.org/stable/j.ctt9qf7m6

Baynes, T., & Heckbert, S. (2009). Micro-scale simulation of the macro urban form: Opportunities for exploring urban change and adaptation., 14–24.

Beven, K. J. (2002). *Rainfall-runoff modelling: the primer*. John Wiley & Sons.

BicycleDutch. (2018). Dutch cycling figures. https://bicycledutch.wordpress.com/2018/01/02/dutch-cycling-figures/

Bousquet, F., & Le Page, C. (2004). Multi-agent simulations and ecosystem management: A review. *Ecological Modelling*, *176*(3), 313–332. https://doi.org/10.1016/j.ecolmodel.2004.01.011

Brown, D. G., Riolo, R., Robinson, D. T., North, M., & Rand, W. (2005). Spatial process and data models: Toward integration of agent-based models and gis. *Journal of Geographical Systems*, *7*(1), 25–47. https://doi.org/10.1007/s10109-005-0148-5

Browning, R. C., Baker, E. A., Herron, J. A., & Kram, R. (2006). Effects of obesity and sex on the energetic cost and preferred speed of walking [PMID: 16210434]. *Journal of Applied Physiology*, *100*(2), 390–398. https://doi.org/10.1152/japplphysiol.00767.2005

CBS. (2019a). Buurt, wijk en gemeente 2019 voor postcode huisnumme. https://www.cbs.nl/nl-nl/maatwerk/2019/42/buurt-wijk-en-gemeente-2019-voor-postcode-huisnummer

CBS. (2019b). Kerncijfers per postcode. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

CBS. (2019c). Kerncijfers wijken en buurten 2019. https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=84583NED&_theme=236

CBS. (2020a). Income distribution (standardised income). https://www.cbs.nl/en-gb/visualisations/income-distribution

CBS. (2020b). Leefstijl en (preventief) gezondheidsonderzoek; persoonskenmerken. https://www.cbs.nl/nl-nl/cijfers/detail/83021NED?q=overgewicht#Ondergewicht_59

CBS, RWS-WVL. (2020). Odin 2019. https://doi.org/10.17026/dans-xpv-mwpg

Chambers, C. P., & Echenique, F. (2016). References. *Revealed preference theory* (pp. 198–214). Cambridge University Press. https://doi.org/10.1017/CBO9781316104293.015

Chapizanis, D., Karakitsios, S., Gotti, A., & Sarigiannis, D. A. (2021). Assessing personal exposure using agent based modelling informed by sensors technology. *Environmental Research*, *192*, 110141. https://doi.org/10.1016/j.envres.2020.110141

Chen, L. (2012). Agent-based modeling in urban and architectural research: A brief literature review. *Frontiers of Architectural Research*, *1*(2), 166–177. https://doi.org/10.1016/j.foar.2012.03.003

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Corburn, J. (2007). Reconnecting with our roots: American urban planning and public health in the twenty-first century. *Urban Affairs Review*, *42*(5), 688–713. https://doi.org/10.1177/1078087406296390

Crooks, A., Castle, C., & Batty, M. (2008). Key challenges in agent-based modelling for geo-spatial simulation [GeoComputation: Modeling with spatial agents]. *Computers, Environment and Urban Systems*, *32*(6), 417–430. https://doi.org/10.1016/j.compenvurbsys.2008.09.004

Crooks, A., Heppenstall, A., & Malleson, N. (2018). 1.16 - agent-based modeling. In B. Huang (Ed.), *Comprehensive geographic information systems* (pp. 218–243). Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.09704-9

De Jong, K. A. (1975). An analysis of the behavior of a class of genetic adaptive systems.

de Mooij, J., Dell'Anna, D., Bhattacharya, P., Dastani, M., Logan, B., & Swarup, S. (2021). Quantifying the effects of norms on COVID-19 cases using an agent-based simulation. *Proceedings of the 22nd International Workshop on Multi-Agent Systems and Agent-Based Simulation, MABS@AAMAS 2021 (To Appear)*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Eurostat. (2022). Harmonised european time use surveys (HETUS). https://ec.europa.eu/eurostat/web/time-use-surveys

Filatova, T., Verburg, P., Parker, D., & Stannard, C. (2013). Spatial agent-based models for socio-ecological systems: Challenges and prospects. *Environmental modelling & software*, *45*, 1–7. https://doi.org/10.1016/j.envsoft.2013.03.017

Flötteröd, G. (2017). A search acceleration method for optimization problems with transport simulation constraints. *Transportation Research Part B: Methodological*, *98*, 239–260. https://doi.org/10.1016/j.trb.2016.12.009

Flötteröd, G., Chen, Y., & Nagel, K. (2012). Behavioral calibration and analysis of a large-scale travel microsimulation. *Networks and Spatial Economics*, *12*(4), 481–502. https://doi.org/10.1007/s11067-011-9164-9

Foursquare. (2022). Places API. https://developer.foursquare.com/docs/places-api-overview

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*.

Gimblett, H. R. (2002). Integrating geographic information systems and agent-based technologies for modelling and simulating social and ecological phenomena. *Integrating Geographic Information Systems and Agent-Based Techniques for Simulating Social and Ecological Processes*, 1–20. https://doi.org/10.1093/oso/9780195143362.001.0001

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 987–91. https://doi.org/10.1126/science.1116681

Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A., & Prasath, V. B. S. (2019). Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. *Information*, *10*(12). https://doi.org/10.3390/info10120390

Heckbert, S. (2010). *Calibration of agent-based models in ecological economics* (Doctoral dissertation). http://hdl.handle.net/11343/35794

Hornberger, G., & Spear, R. (1983). An approach to the analysis of behavior and sensitivity in environmental systems.

Horni, A., Nagel, K., & Axhausen, K. (Eds.). (2016). *Multi-agent transport simulation matsim*. Ubiquity Press. https://doi.org/10.5334/baw

Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *KI - Künstliche Intelligenz*, *29*(4), 329–337. https://doi.org/10.1007/s13218-015-0381-0

Inturri, G., Le Pira, M., Giuffrida, N., Ignaccolo, M., Pluchino, A., Rapisarda, A., & D'Angelo, R. (2019). Multi-agent simulation for planning and designing new shared mobility services [Modal shift, emission reductions and behavioral change: Transport policies and innovations to tackle climate change]. *Research in Transportation Economics*, *73*, 34–44. https://doi.org/10.1016/j.retrec.2018.11.009

Janssen, P., & Heuberger, P. (1995). Calibration of process-oriented models [Modelling Water, Carbon and Nutrient Cycles in Forests]. *Ecological Modelling*, *83*(1), 55–66. https://doi.org/10.1016/0304-3800(95)00084-9

Kaveh, A. (2017). Particle swarm optimization. *Advances in metaheuristic algorithms for optimal design of structures* (pp. 11–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-46173-1_2

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

KNMI. (2022). Knmi devepoer portal. https://developer.dataplatform.knmi.nl/get-started

Knudsen, D. C., & Fotheringham, A. S. (1986). Matrix comparison, goodness-of-fit, and spatial interaction modeling. *International Regional Science Review*, *10*(2), 127–147. https://doi.org/10.1177/016001768601000203

Lamperti, F., Roventini, A., & Sani, A. (2018). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, *90*, 366–389. https://doi.org/https://doi.org/10.1016/j.jedc.2018.03.011

Lindzey, G., & Aronson, E. (1985). Handbook of social psychology. *Psychological Medicine*, *I and II (third edition)*(2).

Lipowski, A., & Lipowska, D. (2012). Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications*, *391*(6), 2193–2196. https://doi.org/https://doi.org/10.1016/j.physa.2011.12.004

Lord, S., Fremond, M., Bilgin, R., & Gerber, P. (2015). Growth modelling and the management of urban sprawl: Questioning the performance of sustainable planning policies. *Planning Theory and Practice*, *16*, 385–406. https://doi.org/10.1080/14649357.2015.1061140

Louviere, J. J., Hensher, D., & Swait, J. (2000). *Stated choice methods*. Cambridge University Press. https://EconPapers.repec.org/RePEc:cup:cbooks:9780521788304

Lu, M., Hsu, S.-C., Chen, P.-C., & Lee, W.-Y. (2018). Improving the sustainability of integrated transportation system with bike-sharing: A spatial agent-based approach. *Sustainable Cities and Society*, *41*, 44–51. https://doi.org/10.1016/j.scs.2018.05.023

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, *4*(162). https://doi.org/10.1057/jos.2010.3

Maggi, E., & Vallino, E. (2016). Understanding urban mobility and the impact of public policies: The role of the agent-based models [Climate Change Targets and Urban Transport Policy]. *Research in Transportation Economics*, *55*, 50–59. https://doi.org/10.1016/j.retrec.2016.04.010

Maliene, V., Grigonis, V., Palevičius, V., & Griffiths, S. (2011). Geographic information system: Old principles with new capabilities. *URBAN DESIGN International*, *16*(1), 1–6. https://doi.org/10.1057/udi.2010.25

Maroñas, J., Paredes, R., & Ramos, D. (2020). Calibration of deep probabilistic models with decoupled bayesian neural networks. *Neurocomputing*, *407*, 194–205. https://doi.org/https://doi.org/10.1016/j.neucom.2020.04.103

Morokoff, W. J., & Caflisch, R. E. (1994). Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, *15*(6), 1251–1279. https://doi.org/10.1137/0915077

Najlis, R., & North, M. J. (2004). Repast for gis. *Social dynamics: Interaction, reflexivity and emergence*, 255–259.

Nicholson, W., & Snyder, C. (2005). *Microeconomic theory: Basic principles and extensions*. Mason, OH: Thomson/Southwestern.

Novack, T., Wang, Z., & Zipf, A. (2018). A system for generating customized pleasant pedestrian routes based on openstreetmap data. *Sensors*, *18*(11). https://doi.org/10.3390/s18113794

of Amsterdam, M. (2017). City street layer data. https://maps.amsterdam.nl/open_geodata/?k=275

of Amsterdam, M. (2021). Public transport layer data. https://maps.amsterdam.nl/open_geodata/?k=381

O'Sullivan, D. (2004). Complexity science and human geography. *Transactions of the Institute of British Geographers*, *29*(3), 282–295. https://doi.org/10.1111/j.0020-2754.2004.00321.x

Plakolb, S., Jäger, G., Hofer, C., & Füllsack, M. (2019). Mesoscopic urban-traffic simulation based on mobility behavior to calculate nox emissions caused by private motorized transport. *Atmosphere*, *10*. https://doi.org/10.3390/atmos10060293

Prügel-Bennett, A. (2004). When a genetic algorithm outperforms hill-climbing. *Theoretical Computer Science*, *320*(1), 135–153. https://doi.org/10.1016/j.tcs.2004.03.038

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks.

Railsback, S. F. (2001). Concepts from complex adaptive systems as a framework for individual-based modelling. *Ecological Modelling*, *139*(1), 47–62. https://doi.org/10.1016/S0304-3800(01)00228-9

Rijkswaterstaat. (2022). Verkeersongevallen - bestand geregistreerde ongevallen nederland. https://data.overheid.nl/dataset/9841-verkeersongevallen---bestand-geregistreerde-ongevallen-nederland#panel-description

Rizvi, S. Z., Abbasi, F., & Velni, J. M. (2018). Model reduction in linear parameter-varying models using autoencoder neural networks. *2018 Annual American Control Conference (ACC)*, 6415–6420. https://doi.org/10.23919/ACC.2018.8431912

Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 627–635). PMLR. https://proceedings.mlr.press/v15/ross11a.html

Salle, I., & Yıldızoğlu, M. (2014). Efficient sampling and meta-modeling for computational economic models. *Computational Economics*, *44*(4), 507–536. https://doi.org/10.1007/s10614-013-9406-7

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10.2307/2548836

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253. http://www.jstor.org/stable/2549561

Schulze, J., Müller, B., Groeneveld, J., & Grimm, V. (2017). Agent-based modelling of social-ecological systems: Achievements, challenges, and a way forward. *Journal of Artificial Societies and Social Simulation*, *20*(2), 8. https://doi.org/10.18564/jasss.3423

Shiono, T. (2021). Estimation of agent-based models using bayesian deep learning approach of bayesflow. *Journal of Economic Dynamics and Control*, *125*, 104082. https://doi.org/10.1016/j.jedc.2021.104082

Sonnenschein, T., Scheider, S., de Wit, G. A., & Vermeulen, R. (2021). Agent-based modeling of urban health interventions: Prospects, methodological considerations and challenges [working paper]. *Expanse project*.

Stonedahl, F., & Wilensky, U. (2011). Finding forms of flocking: Evolutionary search in abm parameter-spaces. In T. Bosse, A. Geller, & C. M. Jonker (Eds.), *Multi-agent-based simulation xi* (pp. 61–75). Springer Berlin Heidelberg.

Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q.-N., Marilleau, N., Caillou, P., Philippon, D., & Drogoul, A. (2019). Building, composing and experimenting complex spatial models with the gama platform. *GeoInformatica*, *23*(2), 299–322. https://doi.org/10.1007/s10707-018-00339-6

ten Broeke, G., van Voorn, G., & Ligtenberg, A. (2016). Which sensitivity analysis method should i use for my agent-based model? *Journal of Artificial Societies and Social Simulation*, *19*(1), 5. https://doi.org/10.18564/jasss.2857

Tonne, C., Basagaña, X., Chaix, B., Huynen, M., Hystad, P., Nawrot, T. S., Slama, R., Vermeulen, R., Weuve, J., & Nieuwenhuijsen, M. (2017). New frontiers for environmental epidemiology in a changing world. *Environment International*, *104*(January), 155–162. https://doi.org/10.1016/j.envint.2017.04.003

Troitzsch, K. G. (2004). Validating simulation models. *In 18th European Simulation Multi-Conference. Networked Simulations and Simulation Networks*, 98–106.

UN Department of Economic & Social Affairs. (2018). *World Urbanization Prospects* (Vol. 12). https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf

Uttara, S., Bhuvandas, N., & Aggarwal, V. (2012). Impacts of urbanisation on environment. *IJREAS*, *2*.

van der Hoog, S. (2019). Surrogate modelling in (and of) agent-based models: A prospectus. *Computational Economics*, *53*(3), 1245–1263. https://doi.org/10.1007/s10614-018-9802-0

van Ham, M., Manley, D., Bailey, N., Simpson, L., & Maclennan, D. (2012). *Neighbourhood Effects Research: New Perspectives*. Springer.

Walz, D. (2021). Concise implementation of the sobol sequence for generating low-discrepancy quasi-random numbers in up to 1111 dimensions. https://github.com/DavidWalz/sobol

Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*, *14*(8), 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456

Wilensky, U. (1999). Netlogo. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*. http://ccl.northwestern.edu/netlogo/

World Health Organization. (2014). Global Status Report On Noncommunicable Diseases 2014.

WorldPop. (2020). Netherlands - population counts. https://data.humdata.org/dataset/worldpop-population-counts-for-netherlands

Wu, F. (1998). An experiment on the generic polycentricity of urban growth in a cellular automatic city. *Environment and Planning B: Planning and Design*, *25*(5), 731–752. https://doi.org/10.1068/b250731

Zambrano-Bigiarini, M., & Rojas, R. (2013). A model-independent particle swarm optimisation software for model calibration. *Environmental Modelling & Software*, *43*, 5–25. https://doi.org/10.1016/j.envsoft.2013.01.004