AGENT-BASED TRANSPORT DEMAND MODELLING FOR THE SOUTH AFRICAN COMMUTER ENVIRONMENT

JANET VAN DER MERWE

A dissertation submitted in partial fulfilment of the requirements for the degree

MASTER OF ENGINEERING (INDUSTRIAL ENGINEERING)

in the

FACULTY OF ENGINEERING, BUILT ENVIRONMENT, AND INFORMATION TECHNOLOGY

UNIVERSITY OF PRETORIA

January 2011

Abstract

AGENT-BASED TRANSPORT DEMAND MODELLING FOR THE SOUTH AFRICAN COMMUTER ENVIRONMENT

by

JANET VAN DER MERWE

Supervisor	:	Prof. J.W. Joubert
Department	:	Industrial and Systems Engineering
University	:	University of Pretoria
Degree	:	Masters of Engineering (Industrial Engineering)

Past political regimes and socio-economic imbalances have led to the formation of a transport system in the Republic of South Africa (RSA) that is unique to the developing world. Affluent communities in metropolitan cities are situated close to economic activity, whereas the people in need of public transport are situated on the periphery of the cities. This demographic structure is opposite to that of developed countries and complicates both the provision of transport services and the planning process thereof.

Multi-Agent Transport Simulation (MATSim) has been identified as an Agent-Based Simulation (ABS) approach that models individual travellers as autonomous entities to create large scale traffic simulations. The initial implementation of MATSim in the RSA successfully simulated private vehicle trips between home and work in the province of Gauteng, proving that there is enough data available to create a realistic multi-agent transport model. The initial implementation can be expanded to further enhance the simulation accuracy, but this requires the incorporation of additional primary and secondary activities into the initial transport demand.

This study created a methodology to expand the initial implementation in the midst of limited data, and implemented this process for Gauteng. The first phase constructed a 10% synthetic population that represents the demographic structure of the actual population and identified various socio-demographic attributes that can influence an individual's travel behaviour. These attributes were assigned to the synthetic agents by following an approach that combines probabilistic sampling and rule-based models. The second phase used agents' individual attributes, and census, National Household Travel Survey (NHTS)

and geospatial data to transform the synthetic population into a set of daily activity plans—one for every agent. All the agents' daily plans were combined into a `plans.xml` file that was used as input to MATSim, where the individuals' activity plans were executed simultaneously to model the transport decisions and behaviour of agents.

Data deficiencies were overcome by contemplating various scenarios and comparing the macroscopic transport demand patterns thereof to the results of the initial implementation and to actual counting station statistics. This study successfully expanded the initial *home-work-home* implementation of MATSim by including additional non-work activities in the transport demand. The addition of non-work activities improved the simulation accuracy during both peak and off-peak periods, and the initial demand therefore provides an improved representation of the travel behaviour of individuals in Gauteng.

Keywords

MATSim; agent-based transport simulation; transport demand modelling; transport micro-simulation; transportation planning.

Acknowledgements

My gratitude goes out to Prof. Johan W. Joubert, my supervisor, for his guidance and support for the duration of this project. His inputs and recommendations helped to transform the idea into a reality.

I am also grateful to Business Connexion for their Geographic Information System (GIS) dataset of the Gauteng transport network, to Statistics South Africa (Stats SA) for the Census 2001 and NHTS data and to the South African National Roads Agency Limited (SANRAL) for the traffic counting station statistics.

The research project was financially supported by the University of Pretoria and the South African National Research Foundation (NRF), who funded a portion of this work under Grant FA-2007051100019.

I wish to especially thank my husband, André, for his unwavering love, motivation and patience. His enthusiasm about his work is an inspiration to me. To a long list of family, friends, colleagues and mentors—thank you for your contributions and support. Without you it could not have been.

"I can do all things through Christ which strengthens me."
Philippians 4:13

Contents

Contents iv

List of Figures vi

List of Tables vii

List of Acronyms viii

1 Introduction 1
- 1.1 Transport in South Africa . . . 1
- 1.2 Introduction to transport planning . . . 2
  - 1.2.1 Limitations of the state-of-practice models . . . 3
  - 1.2.2 Advances of micro-simulation . . . 4
  - 1.2.3 MATSim: a micro-simulation toolkit . . . 5
- 1.3 Research design . . . 6
- 1.4 Research methodology . . . 7
- 1.5 Document structure . . . 7

2 Transport demand literature 9
- 2.1 Travel behaviour . . . 9
- 2.2 Initial requirements of MATSim . . . 9
  - 2.2.1 Population modelling . . . 10
  - 2.2.2 Initial demand modelling . . . 11
- 2.3 Conclusion . . . 13

3 Population modelling 15
- 3.1 Methodology . . . 15
  - 3.1.1 Data . . . 16
  - 3.1.2 Geographic projection . . . 17
  - 3.1.3 Population modelling procedure . . . 18
- 3.2 Population validation . . . 19
  - 3.2.1 Repeatability . . . 19
  - 3.2.2 Accuracy . . . 21

3.3 Conclusion 24

4 Initial demand modelling **26**
4.1 Methodology 26
4.1.1 Data 27
4.1.2 Activity choice procedure 28
4.1.3 Activity chaining procedure 28
4.1.4 Activity location procedure 31
4.1.5 Mode choice modelling procedure 31
4.1.6 Trip starting time procedure 32
4.1.7 Activity duration procedure 32
4.1.8 Complete individual daily plans 33
4.2 Demand validation 33
4.2.1 Repeatability 35
4.2.2 Accuracy 38
4.3 Conclusion 48

5 Discussion **50**
5.1 Brief summary of findings 50
5.2 Research agenda 51
5.2.1 Predicting transport behaviour 51
5.2.2 Multiple plans per agent 52
5.2.3 Activity chain distribution 52
5.2.4 Activity location 52
5.2.5 Mode of transport 52
5.2.6 Trip starting time 53
5.2.7 Activity duration 53

Bibliography **54**

A Extract of `plans.xml` file **57**

List of Figures

1.1 Four Step Model (FSM) 2
1.2 MATSim controller (adapted from MATSim Development Group (2008)) . . 5

2.1 Choroplethic mapping versus dasymetric mapping 11
2.2 Fully agent-based approach (Adapted from Balmer (2007)) 14

3.1 Overview of the population modelling procedure 18

4.1 SANRAL counting stations in Gauteng selected for this study 29
4.2 Distribution of work trip starting times in Gauteng, as reported by NHTS . 32
4.3 Influence of opening and closing times of secondary activities on the simulated departure and arrival times 34
4.4 Influence of initial trip starting times for non-work activities on simulation results 35
4.5 Comparison of the utility score evolution of the four scenarios of allocating non-work activity trip starting times 36
4.6 Percentage difference between the departure, arrival and *en route* vehicle counts for the five simulation runs 37
4.7 Hourly counting station errors 41
4.8 Comparison of simulation counts against actual counting station statistics in the morning peak period 43
4.9 Comparison of simulation counts against actual counting stations in the afternoon peak period 44
4.10 *Work* and *Education* trip duration comparison between the four trip starting time scenarios 46
4.11 *Work* and *Education* trip duration comparison between the four trip starting time scenarios and NHTS data 47
4.12 Income distribution of NHTS sample 49

List of Tables

3.1 Sample from a synthetic population for Gauteng 20
3.2 MAD of multiple synthetic populations' individual attributes 21
3.3 T-test results of multiple synthetic populations' individual attributes as measured against the census data 22
3.4 Comparison in age distribution between multiple synthetic populations and census data 23
3.5 Comparison in transport modes between the synthetic populations and census data 24

4.1 Activity chain distributions for Switzerland and the RSA 30
4.2 Statistics of average executed utility score comparison of five different `plans.xml` files 36
4.3 Comparison between the trip durations from the simulation results of five different `plans.xml` files 38
4.4 Error statistics for simulated versus actual traffic count comparisons 40

List of Acronyms

ABA	Activity-Based Approach
AcDG	Activity-based Demand Generation
AgDG	Agent-based Demand Generation
ABS	Agent-Based Simulation
BCEA	Basic Conditions of Employment Act, No. 75 of 1997
BCX	Business Connexion
CDF	Cumulative Distribution Function
DoT	Department of Transport
DTA	Dynamic Traffic Assignment
ESRI	Environmental Systems Research Institute
FSM	Four Step Model
GA	Genetic Algorithm
GAP	Geospatial Analysis Platform
GCS	Geographic Coordinate System
GIS	Geographic Information System
IPF	Iterative Proportional Fitting
MAD	Mean Absolute Deviation
MATSim	Multi-Agent Transport Simulation
NHTS	National Household Travel Survey
NRF	National Research Foundation
OD	Origin-Destination

PDS	Population Density Surface
PES	Post-Enumeration Survey
RSA	Republic of South Africa
SANRAL	South African National Roads Agency Limited
SP	Sub-Place
Stats SA	Statistics South Africa
TAZ	Travel Analysis Zone
UTM	Universal Transverse Mercator
XML	Extensible Markup Language

Chapter 1

Introduction

The trend of motorisation and suburbanisation is caused by overpopulated city centres, and it has fundamentally changed the travel environment of metropolitan residents. An increase in the spatial distribution between residential and work locations necessitates an increase in the number of daily trips made by an individual. Susilo and Kitamura (2008) mention that, in addition to motorisation and suburbanisation, the demographic and socio-economic composition of metropolitan cities have changed substantially.

Motorisation continuously increases the number of vehicles on the roads. But vehicles alone do not produce traffic. Every vehicle has a driver with a travel agenda and schedule that conflicts with that of other drivers (Rieser, 2004). The aggregated conflict of a number of individuals' travel schedules produces traffic. Illenberger et al. (2007) state that transportation is no longer only about extending the infrastructure, but about efficiently using the existing transport networks.

1.1 Transport in South Africa

The South African transport system has unique characteristics in the international transport environment. Land-use and the population distribution in the Republic of South Africa (RSA) contradict that of first world countries (Fletterman, 2008). In first world countries, the people in need of public transport—the low income segment—are located near employment opportunities, which are most often close to city centres. In the RSA, the low income segment of the population is situated on the periphery of metropolitan cities due to racial segregation. The racial segregation is a result of the Apartheid era and complicates the provision of transport services.

Public transport in the RSA captures only 50% of the passenger transport market—the low income segment. The other 50% of the passenger transport market either do not have access to the public transport network, or it is not a viable alternative if compared to private transport (Fletterman, 2008). The low use of public transport leads to over-utilisation of the road network in the constantly growing metropolitan cities. The disparate commuting circumstances result in complex interactions between its constituent units.

1.2 Introduction to transport planning

The history of travel demand modelling in the RSA has been dominated by the Four Step Model (FSM) (Diedericks and Joubert, 2006). The FSM has been associated with the extensive development of transport systems by focusing on the construction of new infrastructure facilities, as discussed by Davidson et al. (2007). The model in Figure 1.1 deals with transport network complexities by formulating the transport process as four consecutive steps, namely *trip generation*, *trip distribution*, *modal split* and *route assignment* (McNally, 2000). The *trip generation* step determines the number of outgoing and

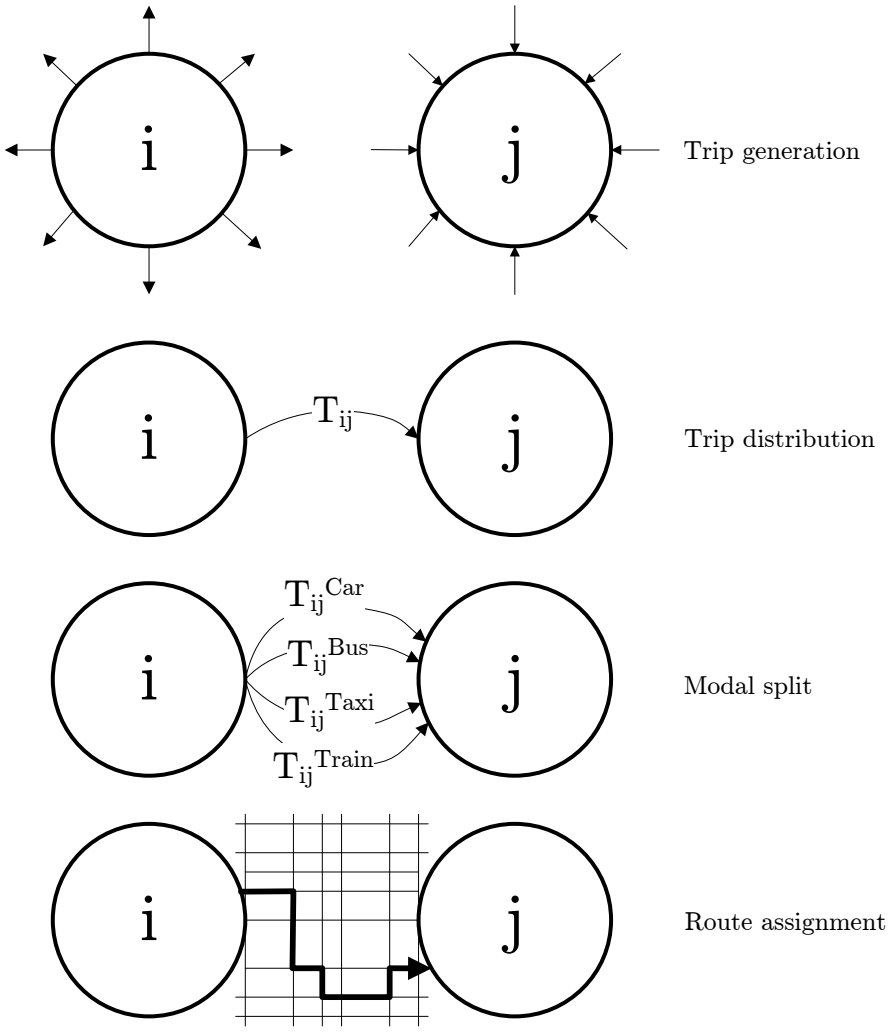


Figure 1.1: Four Step Model (FSM)

incoming trips for every zone. The *trip distribution* step connects the origins and destinations to form trips and produces an Origin-Destination (OD) matrix. In the *modal split* step, the different modes of transport are considered, resulting in an OD matrix per mode of transport. During the last step, *route assignment*, routes are allocated to every trip that is simulated. An equilibration of demand and performance is reached in this step, resulting in link volumes and travel times that can be compared to actual traffic statistics.

1.2.1 Limitations of the state-of-practice models

Kitamura (1988) states that the FSM examines each step in isolation and uses a static and steady-state route assignment step. In the FSM, demand on a transport network is generated from aggregated historic data in terms of the trips between various zones in the simulated region (McNally, 2000). The characteristics of the transport network that serves the demand also predicts the traffic flow and modal distribution. The FSM does not result in a realistic representation of people's decision making processes and offers, at best, an imperfect setting for behavioural analysis of travel.

The traditional traffic demand generation models, such as VISUM, are based on the OD matrices created by the FSM (Balmer et al., 2005). VISUM differentiates between subgroups in the population, such as juniors, working adults, non-working adults and seniors, by the respective distribution of activity patterns. A synthetic population is generated from census data and an activity chain[1] is assigned to every subgroup of the population. Activity locations are assigned using the gravity model, and a logit model is used to assign the mode choices. The results are aggregated into hourly zonal OD matrices based on time-specific transition probabilities between the various activity types.

The state-of-practice models are realistically and critically investigated by Davidson et al. (2007), indicating its limitations. The two major deficiencies of the models are:

Internal inconsistencies There are numerous internal inconsistencies across different outcomes of the models. One of the inconsistencies is the unavoidable and uncontrolled discrepancies between the amount of home-based and non-home-based trips produced by, and attracted to, each zone. Another inconsistency is the imbalanced modal split for outbound and inbound trips of a specific zone.

Base year replication problems The models are unable to replicate the base year transport statistics for the study area without major adjustments to the model parameters. Even with major adjustments made to trip generation rates, mode-specific constants and trip tables, the model may well be irrelevant for another base year.

To overcome the deficiencies of the aggregated state-of-practice models, two steps are required (Balmer, 2007). The process of aggregating traffic data limits the predictive power of the state-of-practice models by reducing the amount of available traffic information (Charypar et al., 2006). The first missing step involves the disaggregation of individual travellers, where an individual's daily decisions are connected to his demographic data. Individuals' decisions are modelled more realistically when the travellers are disaggregated.

Merchant and Nemhauser (1978) stress the importance of time-dependent demand, stating that there is little evidence that the constant-demand assumption of the state-of-practice models provides a reasonable approximation to the dynamic travel behaviour that occurs in peak traffic. The second missing step of the state-of-practice models changes the

[1] All the activities of one person in chronological order for one day (Rieser, 2004).

current time-independent model to a time-dependent model (Balmer, 2007). The time-dependent model does not assume that all vehicles are in a steady state, and it therefore considers peak traffic, shockwaves and trip duration.

Diedericks and Joubert (2006) state that the state-of-practice models, such as the FSM, can not reflect the emerging interactions within the complex South African transport system. To overcome the shortcomings of the state-of-practice models, Fourie (2009) identifies and investigates micro-simulation as an approach to model the disparate transport system.

1.2.2 Advances of micro-simulation

An ideal transport planning method should consider all aspects of traffic without losing important information by aggregating data. It is impossible to completely understand complex transport systems, but the system can be improved by understanding the behaviour and decision making processes of the constituent units.

The FSM models transport demand with individual, independent trips as the primary unit of analysis. Travel demand models are progressing from the FSM towards more behaviourally realistic micro-simulation models such as the Activity-Based Approach (ABA) and Agent-Based Simulation (ABS) (Vovsha et al., 2005). Micro-simulation represents transport demand on an individual level, recognising that a large number of individual trips result in the observed macroscopic demand patterns (Davidson et al., 2007).

The ABA uses tours—a number of consecutive dependent trips—as the primary unit of analysis in transport demand models (Vovsha et al., 2005). Where the ABA focuses on tours, ABS continues the disaggregation by modelling individual travellers as autonomous entities with unique activity schedules (Balmer et al., 2005). In ABS, every entity has a set of predefined behaviour alternatives that can be influenced by other entities and by the surrounding physical environment (Balmer, 2007). The decision making processes of every entity in an ABS can be modelled explicitly, but information can also be aggregated to obtain the overall statistics of the transport system (Balmer et al., 2005; Davidson et al., 2007).

Micro-simulation offers a number of advantages over state-of-practice transport planning model, which increases the popularity of these models in transport simulation, analysis and forecasting (Balmer, 2007; Balmer et al., 2005; Ciari et al., 2007; Rieser, 2004). The decision making processes of every individual can be modelled explicitly, changing travel related decisions from the fractional probability used at aggregated level, to discrete decisions made by individuals (Davidson et al., 2007). Not only can information of every traveller and his decision making processes be collected with micro-simulation, but information can also be aggregated to obtain overall statistics. Unlike the state-of-practice models, ABS is capable of modelling the complex transport system in the RSA, and predict the effect of governing policies (Davidson et al., 2007; Fourie, 2009).

1.2.3 Multi-Agent Transport Simulation (MATSim): a micro-simulation toolkit

MATSim is a transport related ABS that provides a variety of tools and approaches to implement large-scale transport simulations (Balmer et al., 2008). According to Charypar et al. (2006), travel demand is generated as an aggregated by-product when each person is represented as an individual agent whose daily activity plan is executed in the simulation. Figure 1.2 depicts the various steps involved in MATSim, as discussed by Balmer (2007).

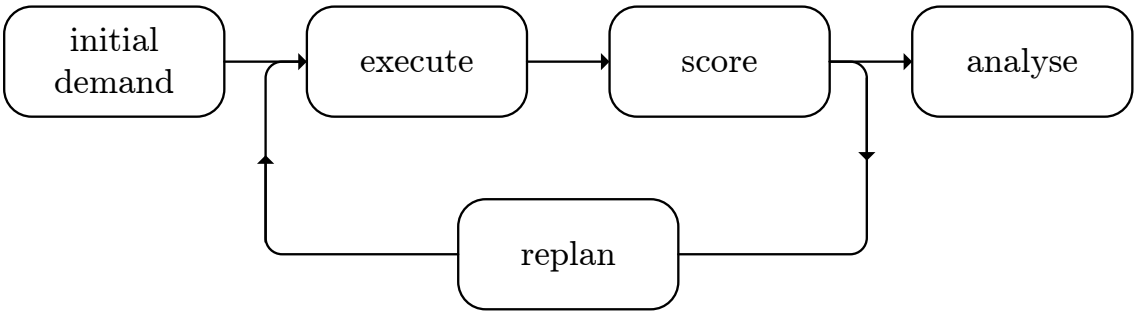


Figure 1.2: MATSim controller (adapted from MATSim Development Group (2008))

Initial demand A synthetic population of individuals, known as agents, is created. These agents live in a virtual world that is based on the socio-economic attributes of the study area. Every agent holds several attributes such as gender, age, employment status, driver's license ownership and car availability, and has a daily activity plan that represents the agent's activity patterns, activity times, mode choices and route choices throughout the day.

Execute The mobility simulation executes all the agents' daily activity plans simultaneously. Congestion is caused on road links as a result of coinciding activities, locations and times in a number of agents' plans.

Score After the execution, every agent rates the performance of his daily activity plan according to the utility function in Equation (1.1),

$$U_{total} = \sum_{i=1}^{n} U_{perf,i} + \sum_{i=1}^{n} U_{late,i} + \sum_{i=1}^{n} U_{travel,i} \tag{1.1}$$

where n is the number of activities or trips in the plan, $U_{perf,i}$ is the positive utility earned when performing activity i, $U_{late,i}$ is the negative utility earned for arriving late for activity i and $U_{travel,i}$ is the negative utility earned for travelling to activity i.

The utility function improves or deteriorates the agent's score according to the actions of the agent during a specific day. The highest achievable utility score for a plan is derived when an agent arrives on time for every activity, spends maximum time performing the activity and spends minimum time travelling towards the activity.

Replan Agents learn more about their environment as time passes and as they share road infrastructure with other agents, and it is possible for them to change certain aspects of their daily activity plans to attempt achieving a higher overall utility score. The changes can include different departure times to activities, as well as different routes to activities.

After a number of iterations it will become increasingly difficult for agents to further improve their utility scores, allowing the model to settle into a "relaxed state". The relaxed state represents the typical state of the study area, and it can be used to determine the expected traffic load on road links. Road pricing policies can also be evaluated against the relaxed state, and their effect on the transport system can be quantified.

1.3 Research design

The Optimisation Group of the University of Pretoria has embarked on establishing a South African implementation of MATSim, the first results being obtained by Fourie (2009). The initial implementation simulated private vehicle trips between home and work in Gauteng with an average traffic flow error of approximately 40% during peak times. This research project will refine the process of generating initial transport demand for Gauteng by extending the simulated activities to incorporate additional primary and secondary activities and comparing the results to both that of the initial implementation as well as to actual traffic counts.

The generation of initial demand can be divided into two phases: population modelling and initial demand modelling. During the first phase, a synthetic population that maintains the demographic structure of the actual population will be constructed for Gauteng. A random realisation of the most recent 2001 census data will be used to generate synthetic agents, each with a home location in a specific Sub-Place (SP). Census and National Household Travel Survey (NHTS) data will be used to allocate attributes to every agent, which might influence the agent's travel behaviour.

The second phase involves the initial demand modelling of Gauteng, where each agent in the synthetic population is assigned a complete daily activity plan. Firstly, every agent will be assigned a primary activity location, whether the agent is working, attending school or attending other activities. A daily activity plan will be derived around the primary activity, and will contain at least the following: the activities that are performed by the agent during a specific day, the order in which they are performed, the location of every activity and the mode of transport used between activities. NHTS data will be used for this phase, since it provides statistics for the various activity types and modes of transport. The results of this phase are used as input into MATSim to simulate the private vehicle trips in Gauteng.

The two phases of this study can be combined into the following research statement:

Modelling agent-based transport demand for Gauteng.

The statement leads to the following questions:

- *How can a synthetic population be created that maintains the demographic structure and attributes of the actual population?*
- *How can the travel behaviour of individuals be converted into daily activity plans for every agent?*
- *Do we improve on the initial implementation by simulating actual traffic with improved accuracy?*

1.4 Research methodology

This study will construct a synthetic population for Gauteng as a random realisation of 10% of the 2001 census data. During this process it will be ensured that the synthetic population is a geographic representation of the actual population. The population statistics of every SP in Gauteng, as well as NHTS statistics, will be used to allocate a number of attributes to every agent in the synthetic population.

The synthetic population generation process is executed in *ArcGIS*, and repeated a number of times to ensure consistent and reliable results. The distribution of individual attributes throughout Gauteng will be validated against the overall population statistics of Gauteng.

Several scenarios are contemplated to overcome data deficiencies and initial transport demand is generated for each of the scenarios. The census and NHTS data will be used to allocate a primary activity location to every agent, after which a daily activity plan is constructed for every agent in the synthetic population. The activities in which every synthetic agent participates is determined and a mode of transport is allocated to every agent. An agent's daily activity plan consists of the sequential primary and secondary activities performed by the agent; the location of the activities; the starting time and duration of the activities; and the mode of transport between the activities. The daily activity plans that use private vehicles will be used as input into MATSim and the results thereof validated against the relevant NHTS and 2001 census data.

1.5 Document structure

In Chapter 2 we review literature on generating transport demand, not only in the broader body of knowledge, but also more specifically in the current version of MATSim. Various approaches to generate the initial demand for MATSim are discussed and evaluated based on their advantages, disadvantages and data requirements.

A synthetic population is constructed as a random realisation of census data in Chapter 3. Various socio-demographic attributes are identified that can influence an individual's transport behaviour and instances of these attributes are assigned to every agent in the synthetic population.

Based on the individual attributes, various daily activity plans are constructed for every agent in Chapter 4. The various plans cater for different scenarios of unavailable or incomplete data. For every scenario, the activity plans of all the agents are simultaneously executed and the results compared to actual traffic counts to determine how accurately the simulation represents reality.

Chapter 5 summarises the findings of the study and reviews the contributions thereof. A future research agenda is established by discussing opportunities to further enhance the study.

Chapter 2

Transport demand literature

Transport demand is derived from a need or desire to participate in spatially distributed activities (Recker, 1995). Every individual has a daily schedule that comprises an interrelated set of activities and the need for economic performance necessitates the mobility of goods. The mobility needs of both people and goods imply traffic.

2.1 Travel behaviour

The daily travel behaviour of a household is to a large extent routine, as shown by the studies of Garling and Axhausen (2003) and Schlich and Axhausen (2003). However, individual household members occasionally make deliberate choices that influence their travel behaviour (Garling et al., 1994). There has been a number of studies to determine how such choices are made.

The study of Susilo and Kitamura (2008) examines how long-term changes in the demographic and socio-economic attributes of a population, as well as changes in the travel environment, impact commuters' travel patterns. The study shows that both private and public transport commuters increased the number of non-work visits over the years preceding the study, but they decreased the total time spent on non-work activities. Overall, both private and public transport commuters exhibit a tendency of expansion in activity engagement and travel. An important finding of the study is that, contrary to popular belief, private transport commuters do not tend to chain trips in all travel environments. Instead, private transport commuters tend to have more activity chains with a lower average number of activities per chain, while public transport commuters combine more activities into every activity chain.

2.2 Initial requirements of Multi-Agent Transport Simulation (MATSim)

MATSim provides a variety of tools and approaches to model travel demand and traffic flow (Balmer et al., 2008). Rieser et al. (2007) discuss the importance of initial require-

ments to ensure a fully functional simulation—MATSim requires a synthetic population of individuals with individual transport related attributes and a daily activity plan for every member of the population.

2.2.1 Population modelling

The purpose of population modelling is to construct a synthetic population that maintains the demographic structure of the actual population. A variety of population modelling methods exists. The methods range from trivial procedures with a low level of demographic accuracy to complex procedures with a higher levels of accuracy.

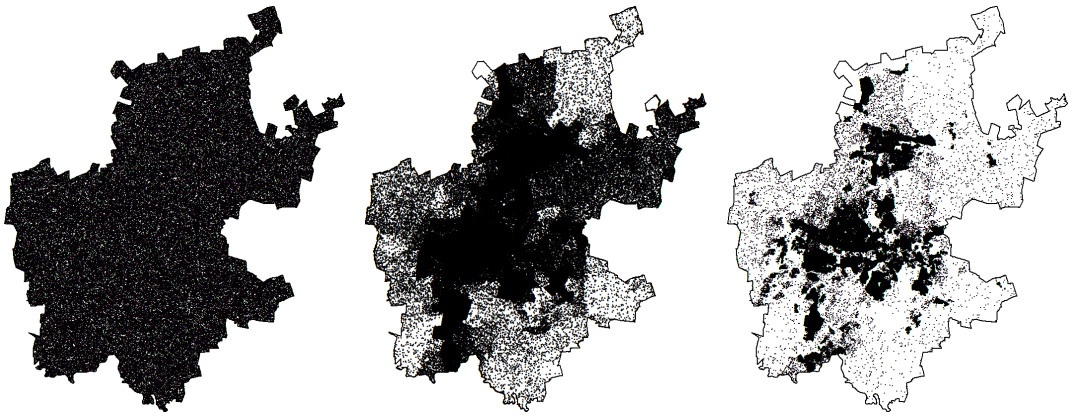
The most straightforward method of cartographically portraying a population in a study area is known as choroplethic mapping, which assumes that the population is equally distributed over the land area (Yuan et al., 1997). According to Langford et al. (2008), it is improbable that the internal population distribution of an area is indeed spatially uniform and choroplethic mapping should therefore only be used when no additional information is available.

Dasymetric mapping refines the process of choroplethic mapping by revealing the intrinsic boundaries of the population distribution and by accounting for variations in the population density (Holt et al., 2004; Moon and Farmer, 2001). Dasymetric mapping estimates the distribution of aggregated data by incorporating additional useful information. The additional information can vary from aggregated data for smaller land areas, satellite imagery that distinguishes between occupied and unoccupied land and land-use data that differentiate between residential and economic areas (Langford et al., 2008).

An advantage of dasymetric mapping is the spatial disaggregation into sparsely and densely populated areas. Therefore, dasymetric mapping provides a finer-grained analysis of the population distribution within an area than choroplethic mapping (Langford et al., 2008). Figure 2.1 illustrates the difference between choroplethic and dasymetric mapping if applied to Gauteng. Where choroplethic mapping distributes the total population evenly over the province, dasymetric mapping incorporates the population data of numerous demarcated areas within Gauteng and distributes the number of individuals within every area evenly.

According to Moon and Farmer (2001), the geographical boundaries that are derived for a census survey are designed to fit administrative and governmental functions. These boundaries often do not coincide with actual human settlement patterns of communities, resulting in a poor representation of the population. In an approach to improve the representation of human settlement patterns, it is suggested that the patterns be represented as a continuous surface over the landscape. The resulting continuous surface is known as a Population Density Surface (PDS), which clearly shows the internal structure and discontinuities—populated and unpopulated areas—in the data.

A PDS can be created for an area by first distinguishing between occupied and unoccupied land, and then combining this information with zonal census data to create a



(a) (b) (c)

Figure 2.1: Choroplethic mapping versus dasymetric mapping: (a) choroplethic mapping, where the population is distributed evenly across Gauteng; (b) dasymetric mapping, where the population in every Geospatial Analysis Platform (GAP) mesozone in Gauteng is distributed evenly across that mesozone; (c) dasymetric mapping, where the population in every SP in Gauteng is distributed evenly across that SP.

dasymetric map. The dasymetric map can be converted to a PDS by calculating the number of people per km^2. By viewing human settlement patterns as a continuous surface, the PDS provides a fundamentally different starting point than viewing human settlement patterns in discrete zones (Moon and Farmer, 2001).

All the population modelling methods are restricted by the available census data. Yuan et al. (1997) stress that census data are mostly reported in an aggregated tabular format to ensure confidentiality and to lower the cost of geographic coding. Thus, census data are often deficient with regards to spatial referencing, complicating the process of creating a disaggregated synthetic population.

Balmer et al. (2004) focus on the importance of the behavioural responses of individual travellers. Individuals have various attributes that influence their decision making processes and their travel behaviour. Socio-demographic attributes such as home location, gender, age, employment status, driver's license ownership, car availability, income, the number of persons in the household and the primary activity location can be allocated to every individual in the synthetic population (Balmer, 2007; Ciari et al., 2007; Frick and Axhausen, 2004). These attributes, along with the individual's interaction with the environment, will determine the strategies according to which the individual reacts in specific transport situations.

2.2.2 Initial demand modelling

Individual travel behaviour needs to be modelled realistically if the actions of individuals in a transport system are to be predicted accurately (Kitamura, 1988). Travellers learn about their environment by integrating the outcomes of consecutive transport experiences and their growing knowledge of the travel environment (Arentze and Timmermans, 2005).

By improving their knowledge about the system, travellers develop strategies to cope with the dynamic system and current travel decisions (Ettema et al., 2004). Through increased experience, an individual's perception of aspects such as travel time is improved. The improved experience results in a more efficient use of road networks, and ultimately a more efficient transport demand distribution in time and space.

To ensure a realistic representation of individuals' travel behaviour, it is necessary to construct individual daily activity schedules. By generating a schedule for every individual, it is possible to monitor every traveller during a desired time period. Traffic is produced when individuals follow their daily schedules, and the schedules of several travellers coincide (Balmer, 2007). Aggregated information can be extracted from the simulation to determine, for instance, the time-dependent traffic volumes, modal split and activity chain distributions.

Balmer (2007) and Balmer et al. (2008) discuss the everyday transport decisions made by individuals. An individual decides in how many activities to participate during a specific day and what those activities will be. The sequence of the activities is established, and the individual determines the activity locations, activity starting times and activity durations. Some other decisions include the group composition of every activity, and the route and transport mode selection between every two activities. All these transport decisions can be incorporated into a complete daily activity plan for every individual (Charypar and Nagel, 2005).

The two main methods of creating the daily activity plans of individuals are Activity-based Demand Generation (AcDG) and Agent-based Demand Generation (AgDG). The method of AcDG can be used to generate initial disaggregated transport demand, as discussed by Balmer (2007). AcDG constructs daily activity plans for every member of the synthetic population, and derives transport demand from the necessary connections between consecutive activities (Charypar and Nagel, 2005). A daily activity plan that contains a compulsory activity chain is assigned to every individual, and all the individuals' trip data are aggregated over specific time periods to produce time-dependent Origin-Destination (OD) matrices. The time-dependent matrices are incompatible with the steady-state route assignment step of the Four Step Model (FSM), which uses time-*in*dependent OD matrices. The route assignment step of the FSM can be replaced with Dynamic Traffic Assignment (DTA), which can handle the time-dependent matrices as input. This traffic assignment process allocates routes to the time-dependent OD matrices, resulting in time-dependent traffic patterns such as link volumes and travel times (Balmer, 2007; Rieser et al., 2007). The traffic patterns are returned as input to the AcDG, and the feedback is iterated until the resulting transport demand is consistent.

The final transport demand is time-dependent and disaggregated into *trips*, but the trips are not connected to specific travellers. The combination of AcDG and DTA gives up disaggregation into *travellers*, as well as the connection between individuals and their performance in the simulation (Charypar and Nagel, 2005).

The method of AgDG, however, disaggregates travel demand into the trips made *per*

individual, allowing the simulation to track their performance on a regular basis. Figure 2.2 shows that AgDG feeds the information regarding the synthetic population and the individuals' activities into the traffic assignment process, ensuring that travellers are maintained as individual entities with individual attributes and decision making processes throughout the entire simulation (Balmer, 2007).

2.3 Conclusion

This chapter reviewed literature on the travel behaviour of individuals and on possible methods to represent this behaviour in a model. The first step to representing individual travel behaviour is to create a synthetic population with transport related attributes and the second step uses the synthetic population to model everyday transport decisions.

The most straightforward population modelling method is choroplethic mapping, which assumes that the population is equally distributed over the study area. Since the probability of having an equally distributed population is very low, dasymetric mapping refines choroplethic mapping by utilising population data that is available for smaller land areas. This method reveals the intrinsic boundaries of the population by accounting for sparsely and densely populated areas. Each individual is represented separately in the population, allowing the allocation of individual attributes to the population.

Initial demand modelling is used to create daily activity plans that model the travel behaviour of individuals. Daily activity plans can be generated with two processes: AcDG and AgDG. The former feeds time-dependent OD matrices into DTA, but loses the disaggregation into travellers, as well as the performance of individuals in the simulation. Conversely, AgDG keeps the disaggregation into travellers and their daily activities, allowing the simulation to monitor the performance of individuals.

This study uses dasymetric mapping to create a synthetic population that demographically represents the actual population. AgDG is then used to model the travel behaviour of individuals and create a daily activity plan for every individual.

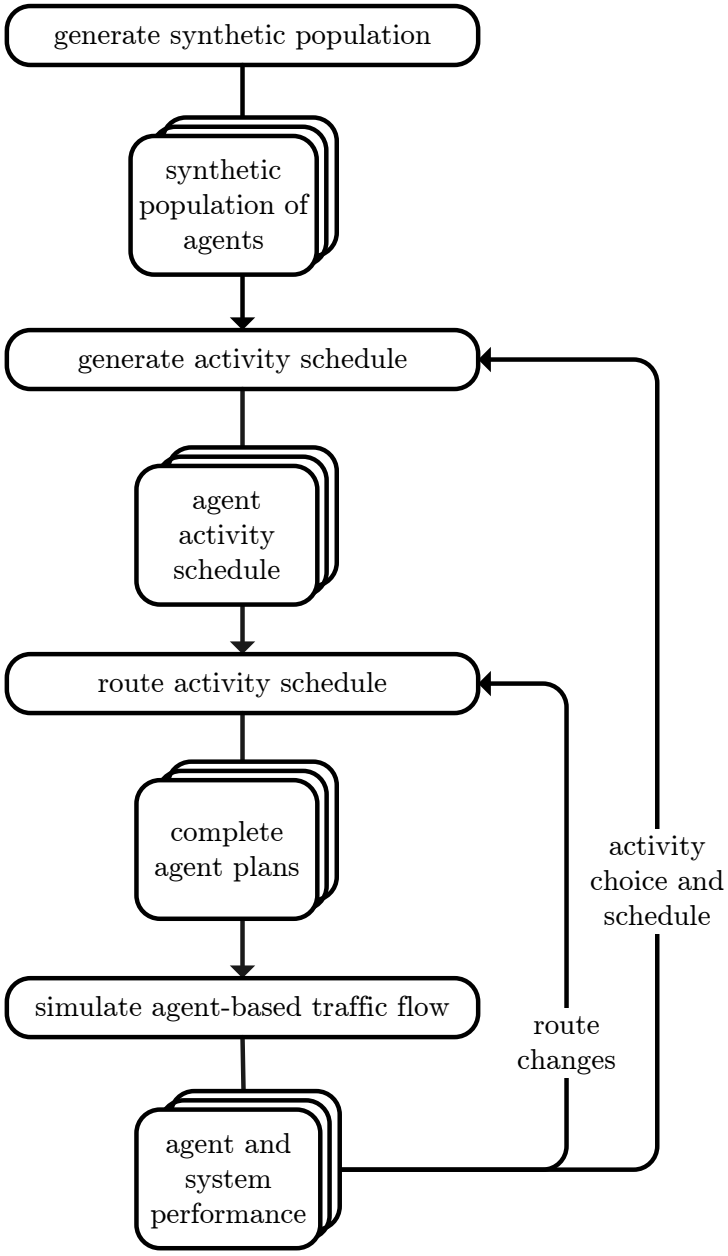


Figure 2.2: Fully agent-based approach (Adapted from Balmer (2007))

Chapter 3

Population modelling

Population modelling is the first step towards an executable scenario of Multi-Agent Transport Simulation (MATSim). This chapter explains the population modelling method used in this research project. The requirements of the method are discussed, the procedure is executed for Gauteng and the results are validated against available population data.

3.1 Methodology

Beckman et al. (1996) have developed a method for the construction of a synthetic population of households. Their method allocates households in an area according to the corresponding census data. The first step of their method uses Iterative Proportional Fitting (IPF)—an iterative algorithm that generates a multiway table in order to expresses the relationship between the categories of several variables (Meulman and Heiser, 1998). This step is used to estimate the proportion of households in a census tract with a desired combination of demographics. To obtain a combination of demographic characteristics, available disaggregated data is used to assign a set of probabilities to each household. The second step constructs a synthetic population of households by selecting entire households from the disaggregated population data based on their demographics, in proportion to the estimated probabilities of demographic combinations as obtained in the multiway table. The number of households to be generated for each census zone can be determined in two ways: the total number of households can be multiplied by the probabilities in the multiway table; or the number of households can be sampled from these probabilities. A set of demographic characteristics is probabilistically assigned to each household.

IPF is fast and has good convergence, therefore Frick and Axhausen (2004) also use it to generate a synthetic population of agents. IPF is able to disaggregate and synchronise data to generate a large number of agents, each with an appropriate set of attribute values. It is stressed that the synthetic population should reflect the demographic structure and the travel behaviour of the actual population.

Research has shown that it is possible to attach the various socio-demographic attributes to the agents in the synthetic population according to the distribution that

those attributes have among the actual population (Ciari et al., 2007). This technique is known as probabilistic sampling. If a random draw from the Cumulative Distribution Function (CDF) of a specific attribute is used to assign a value for that attribute to every agent, the distribution can easily be reproduced in the synthetic population at an accuracy level similar to source data.

Logit models can also be used to add socio-demographic attributes to every agent by using information that is already in the population, even though it is not stated explicitly. Ciari et al. (2007) describe that logit models add attributes to agents with proxy variables that ensure the optimal usage of available data.

Instead of following the IPF method to create a multiway table, this study followed an experimental approach to determine whether sampling from a CDF would yield results of an acceptable standard. The approach involved a combination of probabilistic sampling and rule-based models to attach various socio-demographic attributes. *ArcGIS*®, an Environmental Systems Research Institute (ESRI) product, was used to construct the synthetic population due to its ability to combine the probabilistic approach that uses the distribution of every attribute, with rule-based models that define when specific values of a variable is valid.

3.1.1 Data

A number of data sources have been used to construct the synthetic population for Gauteng. The first and most important dataset is the Population Census of 2001, as documented by Statistics South Africa (Stats SA). The 2001 census counted South Africans for the second time as citizens of a democracy. The national census is the largest and most accurate population survey in the Republic of South Africa (RSA), providing individual, household and labour market data. A Post-Enumeration Survey (PES) was undertaken after the census count to determine the degree of overcount or undercount in 2001 census. The numbers and percentages in the census data reflect the combination of the actual census count and the PES.

One of the limitations imposed by the census data in the RSA is that the population census is only repeated every ten years. The Population Census of 2001 is thus the only and most recent dataset of its kind that is available for the country and it was therefore used to construct the synthetic population. The entire dataset can only be adjusted and updated as soon as the results of the Population Census of 2011 are available, but the population size can be updated more often by using extrapolation based on intermediate population growth estimates.

To ensure the confidentiality of the census data, attributes are aggregated per Sub-Place (SP) before it is made available to the public. Apart from the SP tables, research institutions—such as the University of Pretoria—also have access to a 10% disaggregated sample of the census data. The sample is compounded by incorporating the survey records of every tenth household in the RSA.

The advantage of the 10% census sample is that it provides individual values for every attribute in the census survey for 10% of the South African population. These individual combinations of attributes can be used for the conditional sampling of attributes. For example, if the age of an individual is known, a CDF of employment status—based on the individual's age and home location—can be calculated from the 10% sample. Every population attribute can either be sampled from the respective conditional CDF, or the conditional CDFs can be used to validate attributes.

Gauteng is represented by 19.5% of the 10% sample. However, only 2.79% of these records contain values for all the fields required to create a synthetic population. If used in this study, the 10% sample will only represent 1.18% of the population of Gauteng. Another downside is that the finest geographic representation in the 10% sample is municipalities, whereas the census data are available per SP and the National Household Travel Survey (NHTS) data are available per Travel Analysis Zone (TAZ). The 10% census sample was hence not used to construct or validate the synthetic population in this study. The exclusion of the 10% census sample does not impair this research study, since the synthetic population can be constructed and validated from a combination of the aggregated census data and the NHTS data.

Another limitation of the census data is that it provides lower estimates of labour force participation than labour specific surveys. It is suggested that under-reporting of employment in the 2001 census took place in the informal and subsistence agricultural sectors, especially under part-time workers. This limitation can be overcome by incorporating an additional data source—the NHTS.

The NHTS is a Department of Transport (DoT) initiative undertaken in 2003, with the main objective being to understand domestic travel behaviour in the RSA, as well as the travel needs of individuals and households. The survey was designed by Stats SA, who also executed the fieldwork and collected travel data from over $50,000$ households in the RSA. The Minister of Transport released the results in Parliament in 2005.

A sample size of $50,000$ households in a country with a population of approximately 45 million seems irrelevant at first glance, but it exceeds the normal sample sizes used for national travel surveys—even in developed countries. The reasoning behind the increase in sample size was that the survey should cater for the diversity of the population, and for a variety of geographic circumstances.

Even with the limitations of the available data, the study provides a synthetic population that represents the demographic attributes of the actual population with improved accuracy.

3.1.2 Geographic projection

The most common Geographic Coordinate System (GCS) used for South African geospatial data is Hartebeesthoek 1994, and the data are projected according to the Transverse Mercator datum.

A datum provides a frame of reference for measuring locations on the surface of the earth, since it defines the origin and the orientation of latitude and longitude lines (Kennedy, 2000). Transverse Mercator is a cylindrical projection where the cylinder is longitudinal along the equator, and specifically placed on the region to be highlighted.

MATSim, however, requires all shapefiles to be projected according to the Universal Transverse Mercator (UTM) datum. Kennedy (2000) describes this datum as a specialised application of the Transverse Mercator projection. The globe is divided into 120 zones, where each zone spans six degrees of longitude and has its own central meridian. Gauteng falls within UTM zone 35S, and therefore all the necessary geospatial data are projected according to that zone.

3.1.3 Population modelling procedure

The process of constructing a synthetic population with individual socio-demographic attributes in *ArcGIS* requires three main steps. The first step creates the synthetic population with individual home locations, and the other two steps allocate individual attributes to the population (see Figure 3.1).

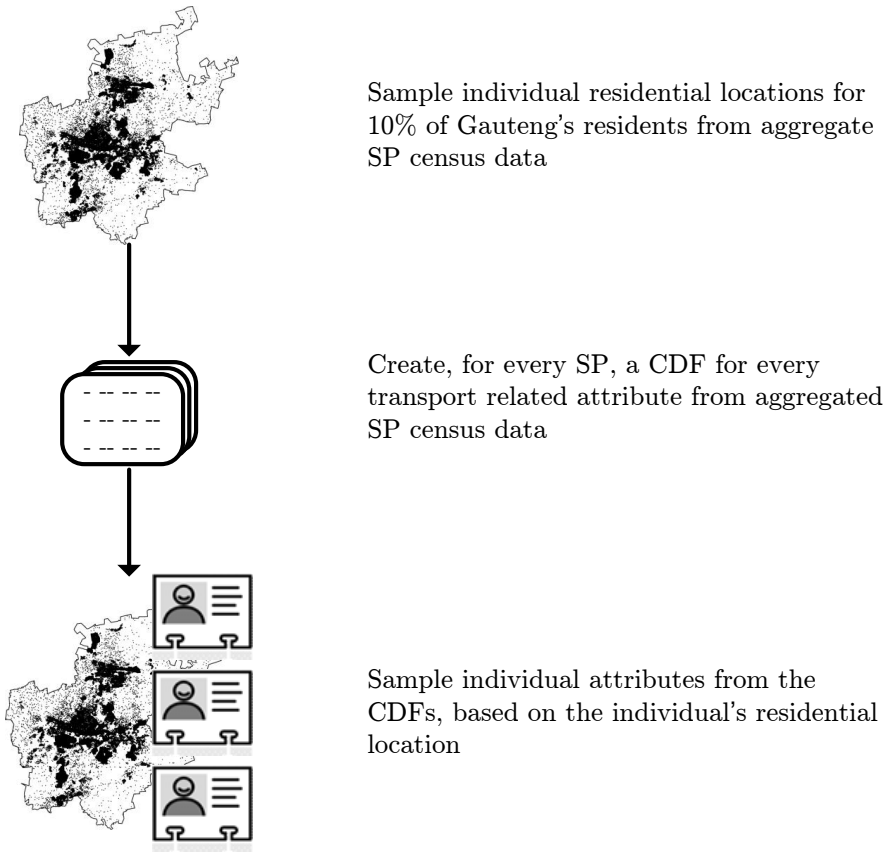


Figure 3.1: Overview of the population modelling procedure

In the first step, the census SP tables are used to sample 10% of the residents' home locations within Gauteng. Every sampled home location is assigned to a synthetic agent

to ensure that the synthetic population is a geographic representation of the actual population. The number of synthetic agents within every SP corresponds to 10% of the census count in that SP.

The second step of the procedure uses the census SP tables to create, for every SP, a CDF for the following transport related attributes in the census data: gender, age, employment status and mode of transport. The CDF depicts the distribution of people in the specific SP with certain values or categories of the attribute.

The last step of the procedure firstly uses rule-based models to determine the valid attribute values for every agent. Examples of rule-based models include that individuals below the age of 15 must attend an educational institution, individuals between the ages of 15 and 65 are eligible for employment and individuals must be 18 years or older to have a valid driver's license. The last step of the procedure then uses the probabilistic approach to sample the individual socio-demographic attributes from the respective CDFs.

The census data do not provide information on driver's license ownership or car availability, and therefore these attributes are still to be allocated. The NHTS reports the percentages of individuals with driver's licenses and access to cars, and these percentages are used to sample the two remaining attributes. Table 3.1 shows a number of agents from a synthetic population for Gauteng, with their respective individual attributes.

There is a level of randomness in the procedure followed to create a synthetic population. Beckman et al. (1996) suggest that multiple synthetic populations are constructed to investigate the inherent uncertainty in the results. This study therefore repeated the above procedure to construct five separate synthetic populations that can be compared to each other.

3.2 Population validation

This section investigates the inherent uncertainty in the five constructed synthetic populations to determine the repeatability of the population modelling procedure and the accuracy of the respective attribute values.

3.2.1 Repeatability

The first validation measure used in this study represents the statistical dispersion of a data set, and is known as the Mean Absolute Deviation (MAD). If a data set $\{x_1, x_2, \ldots, x_n\}$ exists, and $m(X)$ is the preferred measure of central tendency, the MAD is calculated with Equation (3.1). In this study, $m(X)$ refers to the mean of the specific attribute values among the five synthetic populations.

$$\text{MAD} = \frac{\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|}{m(X)}\% \tag{3.1}$$

The MADs of the sampled individual attributes are depicted in Table 3.2. All the

Table 3.1: Sample from a synthetic population for Gauteng

Home coordinates	Gender	Age	Employment status	Driver's license	Car availability	Transport mode
576960.83; 7106525.03	Female	32	Not economically active	Yes	Never	Bus
576986.33; 7106463.66	Female	18	Unemployed	No	Not applicable	Foot
577341.45; 7106383.19	Male	10	Not applicable	Not applicable	Not applicable	Minibus taxi
577132.46; 7106311.56	Female	12	Not applicable	Not applicable	Not applicable	Foot
577280.13; 7106828.01	Male	43	Unemployed	No	Not applicable	Bicycle
577370.90; 7106873.59	Female	13	Not applicable	Not applicable	Not applicable	Foot
577411.66; 7106834.96	Female	20	Unemployed	No	Not applicable	Other
578089.41; 7104298.60	Male	43	Employed	Yes	Always	Car (Driver)
577888.78; 7104238.06	Female	23	Not economically active	No	Not applicable	Train
577476.75; 7104222.56	Male	9	Not applicable	Not applicable	Not applicable	Car (Passenger)
577938.57; 7104387.44	Male	31	Unemployed	No	Not applicable	Other
578116.72; 7104238.53	Female	2	Not applicable	Not applicable	Not applicable	Car (Passenger)
577422.91; 7104270.84	Female	67	Not applicable	No	Not applicable	Train
578260.13; 7104055.85	Male	4	Not applicable	Not applicable	Not applicable	Minibus taxi
577375.97; 7106343.67	Female	11	Not applicable	Not applicable	Not applicable	Minibus taxi
577515.80; 7106414.29	Male	28	Employed	Yes	Always	Car (Driver)
577657.54; 7106655.29	Female	25	Employed	No	Not applicable	Car (Passenger)
577558.20; 7106865.76	Female	8	Not applicable	Not applicable	Not applicable	Bus
577490.11; 7106725.31	Male	50	Employed	No	Not applicable	Motorcycle
577742.93; 7106299.26	Female	1	Not applicable	Not applicable	Not applicable	Minibus taxi

Table 3.2: MAD of multiple synthetic populations' individual attributes

Attribute	MAD%
Gender	0.09
Age	0.53
Employment status	0.20
Driver's license	0.43
Car availability	0.11
Mode to work or school	0.36

attributes' MADs are below 1%, implying that the allocation procedures of the attributes are repeatable.

3.2.2 Accuracy

Beckman et al. (1996) suggest that synthetic populations be validated by comparing the variables and attributes in the synthetic population to that of the actual population to determine the accuracy of the procedure. The second validation in this study is therefore the two-sided t-test. The t-test is a statistical hypothesis test that assesses whether the means of two groups of data are statistically different from each other.

In this study, the H_0 hypothesis is that $\mu_1 = \mu_2$, where μ_1 is the estimated mean of the specific attribute value for the actual population, and μ_2 is the estimated mean of the attribute values over the five synthetic populations. The H_1 hypothesis is that $\mu_1 \neq \mu_2$, and the significance level of the test, α, is assumed to be 5%. The test statistic, t_0, is calculated from Equation (3.2), and the degrees of freedom, v, from Equation (3.3).

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s}{n}}} \tag{3.2}$$

$$v = n - 1 \tag{3.3}$$

In both the equations, x_1 is the specific attribute value as taken from either the census or NHTS data and x_2 is the mean of the specific attribute value over the five synthetic populations. The standard deviation of the attribute value over the synthetic populations is s, and n is the number of synthetic populations. Since n is equal to five for this study, v is always equal to four.

The t-test states that the H_0 hypothesis of an attribute is rejected and the H_1 hypothesis accepted if $|t_0| > t_{\alpha/2,v}$, and for this study $t_{\alpha/2,v} = 2.776$. The t-test results of the various individual attributes are shown in Table 3.3.

According to Ciari et al. (2007), driver's license ownership and car availability are

Table 3.3: T-test results of multiple synthetic populations' individual attributes as measured against the census data

Attribute	$\|t_0\|$	t-test result
Gender	6.52	H_0 Rejected, H_1 Accepted
Ages 0-59	0.20	H_0 Accepted
Employment	0.07	H_0 Accepted
Driver's license	0.82	H_0 Accepted
Car availability	2.45	H_0 Accepted
Car (Driver) mode	1.18	H_0 Accepted

two important attributes in a synthetic population. Both these attributes accepted the H_0 hypothesis, indicating that the allocation procedure is statistically accurate. NHTS data state that 31.3% of the population in Gauteng exceeding the age of 18 years are in possession of a valid driver's license, and that 18.37% of the population in Gauteng have regular access to a car. From the latter percentage, it can be calculated that 25.8% of the population in Gauteng exceeding the age of 18 years has access to a car. A total of 21.23% of the 25.8% is in possession of a valid driver's license—this translates to 82.3% of the population in Gauteng over the age of 18 years. In the synthetic populations, the mean percentage of driver's licenses is 31.28% with a standard deviation of 0.047%. The mean percentage of car availability to those in possession of a valid driver's license is 82.34% with a standard deviation of 0.026%.

The largest discrepancy between the individual attributes and the census data is the gender distribution. According to census data, the population of Gauteng consists of 48.5% females and 51.5% males. The synthetic populations have an average of 49.9% females and 50.1% males, and both standard deviations are 0.06%. The gender distribution is, however, not critical in simulating transport behaviour, and is only used to personalise the individual agents in the simulation. The difference from the census data is therefore not of great concern for this study.

The distribution of all the age groups below the age of 60 is accepted by the t-test. There are discrepancies between the synthetic populations and the census data for the age groups above 60, with an average $|t_0|$ value of 9.06. Individuals above the age of 60 constitute only 6.2% of the population and the majority of people in these age groups are retired. Since retired individuals have a neglectable impact on the daily traffic in Gauteng—especially during peak periods—these discrepancies are tolerable. The mean percentage of people per age group in the synthetic populations is compared to that of the census data in Table 3.4.

In the last discrepancy between the synthetic populations and the census data, some of the transport modes are rejected by the t-test. The census data only report on the

Table 3.4: Comparison in age distribution between multiple synthetic populations and census data

	.	Synthetic populations	
Age group	Census %	$\overline{x}^a$	σ^b
00-04	8.21	8.22	0.01
05-09	7.68	7.71	0.02
10-14	7.73	7.73	0.02
15-19	8.51	8.56	0.01
20-24	11.04	11.07	0.02
25-29	11.85	11.85	0.02
30-34	9.93	9.90	0.02
35-39	8.75	8.69	0.01
40-44	7.36	7.30	0.02
45-49	5.63	5.61	0.03
50-54	4.21	4.19	0.02
55-59	2.94	2.97	0.01
60-64	2.20	1.84	0.02
65-69	1.47	1.89	0.02
70-74	1.08	1.13	0.01
75-79	0.68	0.74	0.01
80-84	0.46	0.51	0.01
85+	0.27	0.10	0.00

[a] Mean percentage
[b] Standard deviation

Table 3.5: Comparison in transport modes between the synthetic populations and census data

	.	Synthetic populations	
Transport mode	Census %	$\overline{x}^a$	σ^b
Foot	34.01	33.32	0.06
Bicycle	1.07	1.22	0.01
Motorcycle	0.57	0.68	0.01
Car (Driver)	19.18	18.34	0.03
Car (Passenger)	12.64	14.72	0.02
Minibus taxi	20.75	20.33	0.04
Bus	5.92	5.59	0.03
Train	5.19	5.08	0.01
Other	0.66	0.73	0.01

[a] Mean percentage
[b] Standard deviation

mode of transport used to either work or school, and not on any other trips. The census data assume that individuals below the age of 15—constituting 23.6% of the population in Gauteng—attend school, and that individuals between the ages of 15 and 65—constituting 72.4% of the population in Gauteng—are eligible for employment. However, only 45% of the eligible individuals are employed. If only work and school trips are simulated, less than 57% of actual trips made in Gauteng would be accounted for. The mode of transport to work or school—as reported by census data in the SP tables—is therefore applied to all trips made in Gauteng by assuming a similar modal distribution. Because of this change, it is expected that the t-test will not yield perfect results. The t-test accepts the distribution of the transport modes that occurs the most frequent in Gauteng: *foot*, *private vehicle driver*, *minibus taxi*, *bus* and *train*. The following modes of transport were rejected by the t-test: *bicycle*, *motorcycle*, *private vehicle passenger* and *other*. This discrepancy between the synthetic populations and the census data is acceptable, since this study simulates transport in Gauteng for the *private vehicle driver* mode only. The mean percentage of people per transport mode in the synthetic populations is compared to that of the census data in Table 3.5.

The deviations in above mentioned attributes are not of great concern, since the sub-sections of the attributes rejected by the t-test are not imperative to the study.

3.3 Conclusion

This chapter discusses the data required to construct a synthetic population for Gauteng in *ArcGIS* and the procedure followed to construct a single synthetic population. Five

separate synthetic populations were constructed from a combination of census data and NHTS data, and the individual attributes of the synthetic populations were compared to each other and to the source data.

The individual attributes of the synthetic populations proved to have good precision if compared to one another, and are acceptably accurate if compared to the source data. Sampling individual attributes from a CDF proved to be a consistent procedure that provides repeatable results. Each of the five synthetic populations represents the demographic structure of the actual population, and one of them will be used in the initial demand modelling of MATSim.

Chapter 4

Initial demand modelling

Initial demand modelling is the second and final step towards achieving an executable scenario of Multi-Agent Transport Simulation (MATSim). This chapter describes the procedure followed to transform a synthetic population with individual attributes into a set of daily activity plans—one for every synthetic agent. The data requirements of the method are discussed, the method is explained and executed, and the results are validated. The primary measure of simulation quality used in this study is how accurately the results compare to reality, as interpreted from actual traffic counts.

4.1 Methodology

The first step in initial demand modelling is the allocation of primary and secondary activities to agents. Primary activities traditionally have longer durations than secondary activities and include *Home*, *Work* and *Education*. This study continues on the assumption of the national census that all individuals below the age of 15 attend school and that individuals between the ages of 15 and 65 are eligible for employment.

Fourie (2009) initially simulates only *Home* and *Work* activities, producing a feasible solution for the morning peak period. The solution can be improved by including other primary and secondary activities in the daily activity plans of agents.

Once activities have been assigned to the agents, the second step of initial demand modelling—activity chaining—is introduced. According to the study of Balmer (2007), there are approximately 1670 known activity chains, but 21 of these activity chains account for approximately 93% of the 100 most frequently occurring chains. These activity chains are all home-based, as Ciari et al. (2007) suggest to simplify the process. Every activity chain will thus start and end at the home location of the specific agent.

Activities were chained in this study by first establishing a list of valid activity chains for a specific agent, based on the agent's individual attributes. The agent's activity chain was then assigned by sampling from the list of valid activity chains.

The third step allocates locations to all activities that are to be performed in the simulation. Every agent in the simulation already has a *Home* location. For the other two

primary activities, Ciari et al. (2007) use Origin-Destination (OD) matrices at municipal level, indicating the number of individuals living in a zone and the number of individuals that commute to a specific zone for either work or education. These matrices are respectively known as a work commuter matrix and an education commuter matrix. For every agent with *Work* or *Education* as an activity, the activity zones are first sampled according to the distributions in the matrices, and subsequently specific locations are assigned to the activities according to the capacity of the relevant facilities within the zone.

Individual locations are also required for secondary activities. Ciari et al. (2007) assign these locations with a neighbourhood search. If *Home* is the only primary activity, the neighbourhood in which the search is performed is a circular area where the radius is related to the size of the municipality. If there are two primary activities with separate locations in the activity chain, the neighbourhood is created as the union of two circles, with the two locations as centres and the radius proportional to the distance between the two locations.

This study assigned primary and secondary activity locations by allocating a specific location either from an external data source, or by allocating an activity region from external data and then randomly selecting a specific location within that region. The approach followed depends on the availability of the necessary data.

The fourth step of initial demand modelling determines the mode choice of every agent—this has already been completed in Chapter 3. The fifth step establishes the activity starting times for the first non-home activity in every activity chain and the sixth step allocates initial activity durations.

The final daily plan of each agent includes the activities that will be performed, the order in which the activities will be performed, the location of every activity, the transport mode used by the agent, the starting time of the first activity in the chain, and the duration of every activity. The majority of the procedures used for the initial demand modelling were written in the *Java* programming language.

4.1.1 Data

Various data sources were used for the initial demand modelling. The most important dataset—the National Household Travel Survey (NHTS)—was used to assign a work Travel Analysis Zone (TAZ) to every employed agent and to generate the starting times of the first non-home activity in every chain. Background information regarding this dataset is provided in Section 3.1.1.

Geospatial data, supplied by Business Connexion's BCX Geographic Information System (GIS) division, was used to assign actual locations to primary and secondary activities. Business Connexion (BCX)'s dataset pinpoints various locations in the Republic of South Africa (RSA), such as popular attractions, roads, accommodation and banking facilities.

The road network used in the simulation was also supplied by BCX, and contains the national highways and major roads in Gauteng. Fourie (2009) considers this reduced

network to be sufficient for exploratory work, such as this study. The full network contains smaller roads as well, and is recommended for decision support when large monetary investments are considered.

The last dataset is the 2001 counting station data of a selection of 21 pairs of network links within Gauteng. The selected counting stations can be seen in Figure 4.1. This dataset was used to validate the results obtained from MATSim against the actual traffic counts on the various roads in Gauteng. The data were supplied by South African National Roads Agency Limited (SANRAL)—a company that maintains and improves the national road network of the RSA.

4.1.2 Activity choice procedure

Various socio-demographic attributes were assigned to every agent in the synthetic population in Chapter 3. These attributes can be used to assign the primary activities, *Work* and *Education*, to agents.

The *Work* activity was assigned to agents that are employed and the *Education* activity was assigned to all agents below the age of 15. It must be noted that an agent of any other age group has the possibility being assigned an *Education* activity, either as the only primary activity or as a second primary activity.

4.1.3 Activity chaining procedure

Table 4.1 contains a list of the most frequently occurring activity chains along with the percentage occurrence of each chain in a population, as obtained from the Microcensus 2000 in Switzerland. It is anticipated that the composition of the individual activity chains, as well as the number of occurrences per activity chain, might be different in the RSA as a result of the different transport environment and population composition. Due to a lack of activity chaining data in the RSA, the distribution of activity chains from Switzerland was used and adjusted according to the rules and information provided by the census data in the country.

The activity chains that have a significantly different percentage occurrence after the adjustment are *h-w-h*, *h-e-h*, *h-l-h* and *h-s-h*. The RSA had an unemployment rate of 25.3% in 2010, compared to the 3.6% in Switzerland for the same year. This confirms the reduced percentage occurrence of the *h-w-h* chain in the RSA. The fertility rate in the RSA is an estimated 3.3 children per household, which is more than double Switzerland's 1.5 children per household. The higher fertility rate increases the number of *education* trips required, clarifying the higher percentage occurrence of the *h-e-h* chain. The remaining two activity chains—*h-l-h* and *h-s-h*—do not contain any primary activities and their percentage occurrences are influenced by that of the *h-w-h* and *h-e-h* chains.

One activity chain was sampled from the adjusted distribution of activity chains for every agent. The activity chain of an agent must contain the primary activities assigned to the agent (if applicable), and can also contain some secondary activities. Various

Figure 4.1: SANRAL counting stations in Gauteng selected for this study

Table 4.1: Activity chain distributions for Switzerland and the RSA, indicating the most frequently occurring chains. The characters symbolises the following: 'h' = home, 'w' = work, 'e' = education, 's' = shopping and 'l' = leisure.

Nr	Activity chain	Original % occurrence[a]	Adjusted % occurrence[b]
1	h-w-h	26.34	23.78
2	h-w-l-w-h	3.07	2.77
3	h-w-s-w-h	1.75	1.58
4	h-w-w-h	1.76	1.59
5	h-l-w-h	0.72	0.65
6	h-w-l-h	0.99	0.90
7	h-w-s-h	0.79	0.72
8	h-s-w-h	0.47	0.43
9	h-w-e-h	0.09	0.08
10	h-e-h	12.15	30.48
11	h-e-l-h	0.41	1.02
12	h-e-e-h	0.21	0.54
13	h-l-e-h	0.12	0.29
14	h-e-s-h	0.07	0.17
15	h-l-h	27.67	18.98
16	h-s-h	16.59	11.38
17	h-l-l-h	2.44	1.67
18	h-s-l-h	1.58	1.08
19	h-l-s-l-h	1.09	0.75
20	h-s-s-h	0.88	0.61
21	h-l-s-h	0.80	0.55
	Total	100.00	100.00

[a] Distribution from the Microcensus 2000 in Switzerland(Balmer, 2007)

[b] Distribution as adjusted according to RSA rules and information provided by the census data

individual attributes were considered before activity chains were sampled. For employed agents, activity chains were sampled from the first 9 chains—each of these chains contains at least one *Work* activity. For agents younger than 15 years, activity chains were sampled from chains 10–14. Each of these chains include at least one *Education* activity, but no *Work* activities. The remainder of the agents are older than 15 but not employed and their individual activity chains were sampled from chains 10–21, catering for agents who attend educational institutions and for agents who are only performing secondary activities.

4.1.4 Activity location procedure

A normalized OD matrix was created from NHTS data for *Home–Work* trips. The matrix consists of 57 origin and 58 destination TAZs in Gauteng, as well as the probability of occurrence for the combination of every origin and destination.

The work TAZ of the first *Work* activity of an agent was sampled from the OD matrix, where the home TAZ of the agent is the origin and his work TAZ is the destination. A specific work location was obtained by executing a neighbourhood search that randomly selects a work location within the agent's work TAZ from a list of more than $300,000$ actual work locations—as supplied by BCX. There is no dataset available that indicates all possible work locations in Gauteng and the random selection method is therefore an acceptable option.

Where there are two or more consecutive *Work* activities in an agent's activity chain, the neighbourhood search identifies possible work locations within a 50 km radius from the agent's home location and randomly selects the additional work locations from that list.

The locations of the primary activity *Education*, as well as the secondary activities *Shopping* and *Leisure*, were assigned with a similar neighbourhood search. The search identifies all relevant activity locations in the BCX dataset that is within a 20 km radius from the agent's home location and randomly selects one of these locations for the specific activity.

4.1.5 Mode choice modelling procedure

A transport mode was allocated to every agent in Section 3.1.3. It was assumed that an agent uses only one mode of transport to commute between the various activities performed during the day. This assumption was made for two reasons. The first reason is that the MATSim simulation executed in this study simulates only private vehicle drivers, and does not accommodate multiple modes in an activity chain or mode combinations to an activity. The second reason is that no sufficient data are currently available on multiple mode use within one activity chain.

4.1.6 Trip starting time procedure

The NHTS data report on the time that individuals in Gauteng leave home for work in the morning—the distribution of trip starting times is illustrated in Figure 4.2. Approximately

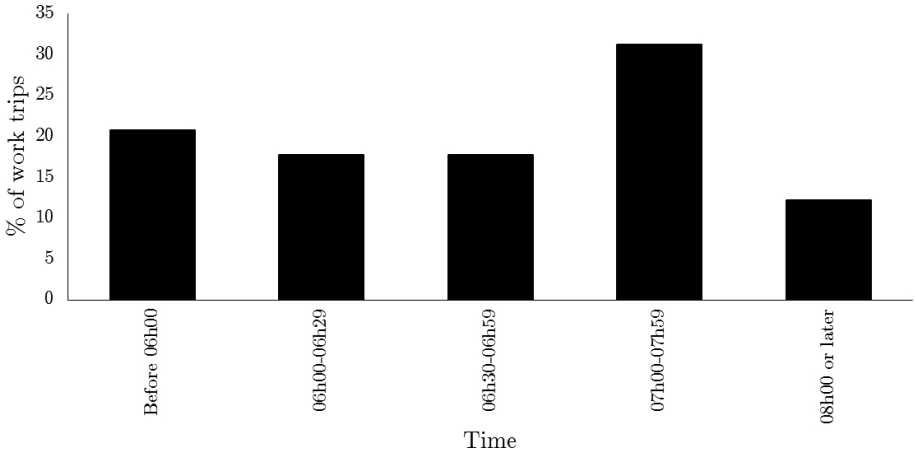


Figure 4.2: Distribution of work trip starting times in Gauteng, as reported by NHTS

35% of activity chains have *Work* as the first non-home activity, and the NHTS data are thus directly applicable to these chains. For the agents constituting the 35%, *Work* trip starting times are sampled from the Cumulative Distribution Function (CDF).

No information is available on activity starting times for the 65% of the activity chains that start with activities other than *Work*. As a preliminary solution, four scenarios were considered to allocate the activity starting times to non-work activities:

Scenario 1 The NHTS work activity trip starting times are directly applied to non-work activities, assuming that the distribution of agents leaving home is the same for work and non-work activities in the morning.

Scenario 2 Non-work activity trip starting times are delayed with one hour from that of work trips.

Scenario 3 Non-work activity trip starting times are delayed with two hours from that of work trips.

Scenario 4 Non-work activity trip starting times are delayed with three hours from that of work trips.

4.1.7 Activity duration procedure

Every activity in an agent's activity chain requires a duration. To determine the duration of *Work* activities, we refer back to the law governing employment in the RSA—the Basic Conditions of Employment Act, No. 75 of 1997 (BCEA). The prescribed duration of a workday is 9 hours, which typically consists of 8 hours working time and 1 hour lunch and/or tea time.

The prescribed workday duration was followed, and for every employed agent with only one *Work* activity in his activity chain, that activity was assigned a duration of 9 hours. If, however, an agent has two *Work* activities in his activity chain, each of the two activities were assigned a duration of 4.5 hours, adding up to a total of 9 hours per day.

No information is available on the typical durations of activities other than *Work* and the following arbitrary assumptions were made about the activity durations:

- Pre-school *Education* activities typically have a duration of 5 hours, whereas the duration increases for primary, secondary and tertiary *Education* activities. The duration of *Education* activities is therefore uniformly distributed between 5 and 9 hours in the simulation.
- *Shopping* activities can vary in purpose and duration, ranging from a couple of minutes to stop at the garage to a number of hours for an extended shopping spree. The duration of *Shopping* activities is therefore uniformly distributed between 0 and 5 hours in the simulation.
- The duration of *Leisure* activities can vary from a couple of minutes to a couple of hours according to the type of activity. The simulation assumes that the duration is uniformly distributed between 0 and 5 hours, since it will then cater for the majority of entertainment activities, outdoor activities, sports, restaurants and other attractions.

4.1.8 Complete individual daily plans

After the successful execution of the initial demand modelling procedures in Sections 4.1.2–4.1.7, the results were combined to create four complete individual daily plans for every synthetic agent—one for every non-work trip starting time scenario. All the agents' daily plans for a specific scenario were consolidated and written as a Extensible Markup Language (XML) file—`plans.xml`. Appendix A is an extract one of the files, showing a number of agents' complete individual daily plans.

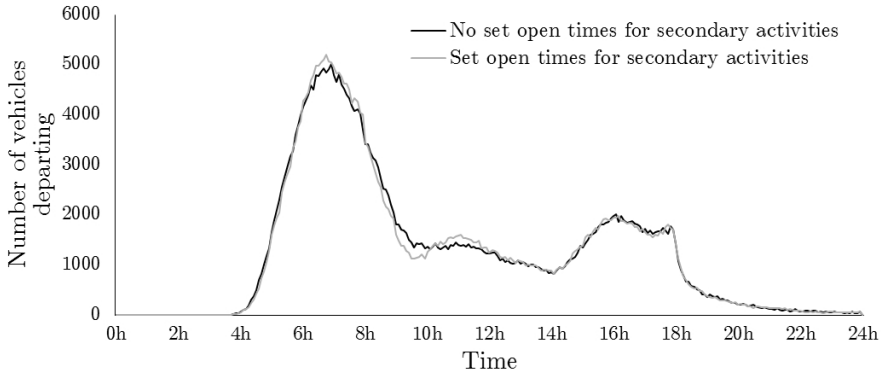
Once the daily plan of every agent is complete, all the agents' plans referring to a specific scenario are executed simultaneously within MATSim and every agent rates the performance of his daily activity plan. The simulation is executed for 100 iterations, since Fourie (2009) suggests that it compares better to the actual system behaviour than a 400+ iteration simulation run. Replanning takes place after every iteration, where an agent can change activity departure times and routes in an attempt to improve his utility score.

4.2 Demand validation

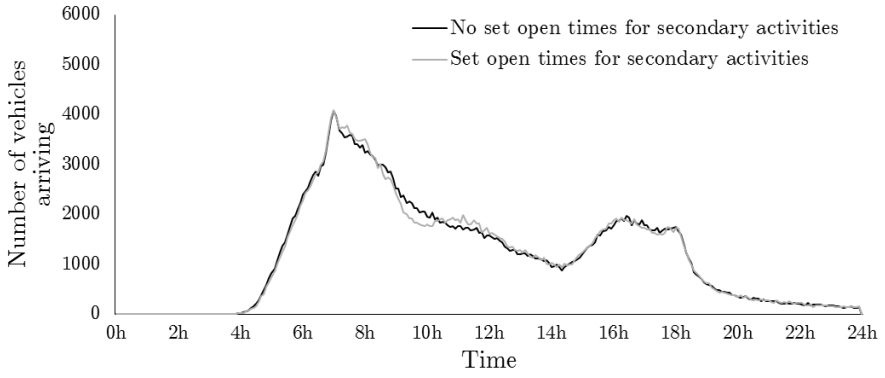
This section evaluates the sensitivity of the simulation for opening and closing times of secondary activities and analyses the impact of four scenarios that allocate the initial activity trip starting times to non-work activities (disucssed in Section 4.1.6). Subsequently,

the repeatability of the initial demand generation procedures and the accuracy of the simulation results are determined.

Two configurations were used to test the sensitivity of the simulation for the opening and closing times of secondary activities. In the first configuration, agents were allowed to perform secondary activities during any time of the day or night. The second configuration specified that *Leisure* activities can be performed between 08h00 and 22h00 and that *Shopping* activities can be performed between 09h00 and 18h00. The comparison between the departure and arrival times of the two configurations dictating the opening and closing times of secondary activities can be seen in Figure 4.3. The simulation proved to have very low sensitivity for the opening and closing times of secondary activities, since there is no significant difference between the departure and arrival statistics of the two configurations.



(a)

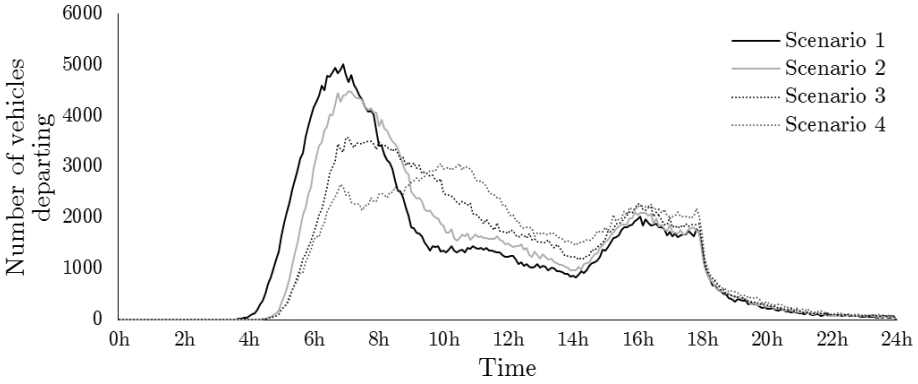


(b)

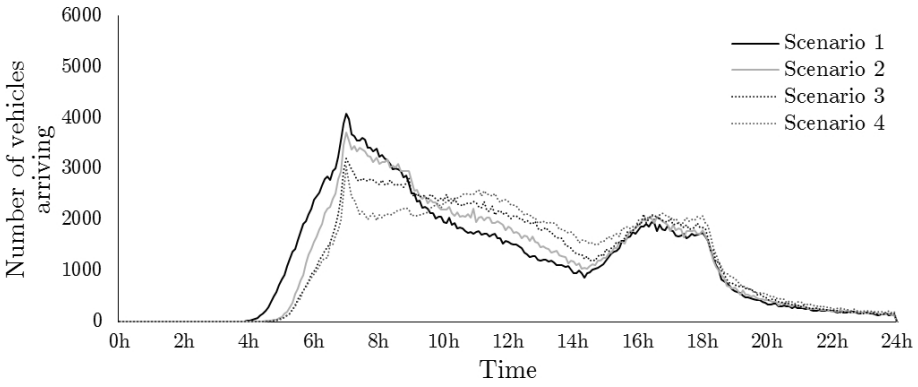
Figure 4.3: Influence of opening and closing times of secondary activities on the simulated departure and arrival times

The departure and arrival times resulting from the four scenarios of allocating the initial activity trip starting times to non-work activities are compared in Figure 4.4. Scenario 1 has a high number of departures in the morning peak period, which is followed by a steep drop that reaches the midday minimum between 10h00 and 14h00. The departures then increase again for the afternoon peak. The same trend is followed for scenarios 2 and 3, each having a lower morning peak and a higher midday minimum than the previous

scenario. The departure distribution of scenario 4 breaks away from the trend with a low morning peak and a daily maximum between 10h00 and 12h00. The trip starting times of non-work activities proved to have an evident influence on the simulated traffic counts, even after 100 iterations. The utility score evolution of the four scenarios is illustrated in



(a)



(b)

Figure 4.4: Influence of initial trip starting times for non-work activities on simulated: (a) Departure times; (b) Arrival times.

Figure 4.5.

4.2.1 Repeatability

The repeatability of the demand generation process was evaluated by creating five different `plans.xml` files for the same synthetic population, where non-work activity trip starting times are equal to work trip starting times. Each of the five files were separately used as input to MATSim, after which the results of the five simulation runs were compared for every iteration.

Equation (4.1) was used, for every iteration, to calculate the percentage difference between the highest and lowest average executed utility scores over the five simulation runs. The utility scores of the different simulation runs are summarised in Table 4.2 by indicating the average, minimum, maximum and 99^{th} percentile of the utility scores, as calculated over the 100 iterations. The same statistics are also used to summarise the

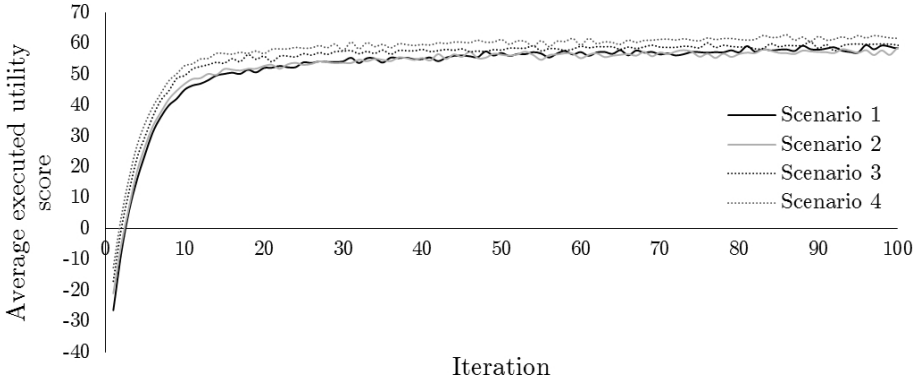


Figure 4.5: Comparison of the utility score evolution of the four scenarios of allocating non-work activity trip starting times

Table 4.2: Statistics of average executed utility score comparison of five different `plans.xml` files

	Utility score over 100 iterations[a]			
Simulation run	Average	Minimum	Maximum	99^{th} Percentile
1	52.35	-26.32	59.98	59.64
2	51.30	-27.39	58.36	58.25
3	51.59	-25.19	59.26	58.54
4	51.93	-27.25	59.47	58.82
5	50.72	-28.11	58.07	58.04
% Difference	4.04[b]	0.94[c]	16.54[d]	7.75[e]

[a] Calculations done for each iteration over all simulation runs
[b] Average difference between the simulation runs over all iterations
[c] Minimum difference between the simulation runs over all iterations
[d] Maximum difference between the simulation runs over all iterations
[e] 99^{th} Percentile of the difference between the simulation runs over all iterations

percentage difference between the highest and lowest average executed utility scores of a single iteration. The demand generation process proved to be repeatable with an average percentage difference of 4% between the utility scores of the different simulation runs.

$$d_i = 100 - \left(\frac{\min\{S_x\}_{x=1}^{r}}{\max\{S_x\}_{x=1}^{r}} \right) * 100 \qquad \forall i = \{1 \dots n\} \tag{4.1}$$

In the above equation, the percentage difference between the simulation runs at iteration i is represented by d_i. The average executed utility score of a simulation run is S_x, where x refers to the relevant simulation run. The number of simulation runs is r and the number of iterations is n; in this case $r = 5$ and $n = 100$.

The vehicle counts for the departure, arrival and *en route* categories of the 100^{th} it-

eration of the five simulation runs were compared over a 24 hour period in 15 minute intervals; the results are shown in Figure 4.6. During the early morning and late night hours, there are definite differences between the vehicle counts of the five simulation runs. There are fewer vehicles on the road during these time periods, causing a small difference between the simulation runs' vehicle counts to have a significant statistical impact. Supporting this statement, the percentage difference between the vehicle counts during the day—when more agents are commuting between activities—is significantly smaller than the difference during the night.

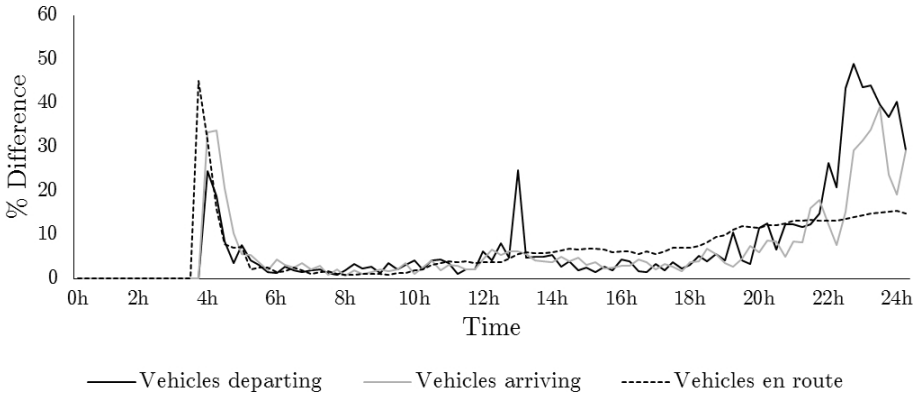


Figure 4.6: Percentage difference between the departure arrival and *en route* vehicle counts for the five simulation runs. Only calculated for simulated vehicle counts larger than 10 per 15 minute interval (as suggested by Fourie (2009)).

The percentage difference between the number of vehicles departing in every 15 minute interval has a standard deviation of 11.06% over a 24 hour period. The midday spike in percentage difference for departing vehicles is the result of the varying number of agents commuting to other activities in this time period, such as a different *Work* or *Education* activity, or a *Leisure* or *Shopping* activity during lunch time. For vehicles arriving at activities, the percentage difference between the simulation runs is below 10% from 04h30 until 21h00, and that of *en route* vehicles gradually increases from 1% at 06h00 until it reaches 13% at 21h00. The increasing percentage difference for *en route* vehicles can be explained by the increasing number of vehicles on the roads during the day that congests the roads and increases the travel time.

Table 4.3 compares the number of vehicles per 15 minute trip duration interval from the 100^{th} iteration of the five simulation runs. This comparison is done for every activity pattern that is related to a primary or secondary activity, such as the following *Work* activities: *education-work*, *home-work*, *leisure-work* and so forth. The percentage difference between the highest and lowest number of vehicles over the trip duration categories is calculated for every activity pattern with Equation (4.1), where n is changed to 1 and S_x represents the average number of vehicles in a trip duration category for simulation run x. The average percentage difference for the activity patterns related to secondary activities varies between 21.60% for *Leisure* activities and 25.50% for *Shopping* activities. No dataset is available to increase the intelligence of selecting secondary activity loca-

Table 4.3: Comparison between the trip durations from the simulation results of five different `plans.xml` files

	Average number of vehicles per 15 minute trip duration interval[a]				
Simulation run	Leisure	Shopping	Education	Work (part-time)[b]	Work (full-time)[c]
1	1010.27	603.55	454.75	608.20	1812.80
2	1003.03	595.63	453.78	591.80	1822.17
3	1007.30	596.30	449.55	600.93	1813.40
4	1007.35	593.73	453.03	602.40	1813.07
5	1014.87	590.20	448.90	612.30	1806.43
% Difference[d]	21.60[e]	25.50[e]	23.75[e]	19.30[e]	16.24[e]

[a] Calculations done for every relevant activity patterns (e.g. *education-shopping*, *work-shopping*, *leisure-shopping* and so forth in the *Shopping* column), over all simulation runs

[b] Workday of 4.5 hours

[c] Workday of 9 hours (as prescribed by BCEA)

[d] Only calculated for simulated vehicle counts larger than 10 per 15 minute interval (as suggested by Fourie (2009))

[e] Average difference between the simulation runs over all relevant activity patterns

tions. The locations were allocated with a neighbourhood search, and different activity locations were assigned in every `plans.xml` file (see Section 4.1.4). The difference between the activity locations in the various simulation runs causes the variance in the trip durations. Similarly, *Education* activity locations were also allocated with a neighbourhood search and different activity locations were therefore assigned in every `plans.xml` file. Trip durations per activity pattern to this activity differ with 23.75% among the five populations.

Trip durations to part-time *Work* activities have a percentage difference of 19.30% among the various activity patterns in the five simulation runs, proving that even limited initial input data improves the repeatability of the simulation. Further justifying the need for input data, the travel time to full-time *Work* activities has the lowest percentage difference over the different activity patterns and simulation runs—16.24%—and had the most available input data. More accurate and complete input data will decrease the percentage differences between the trip durations of the different simulation runs.

The initial demand modelling procedures were executed multiple times and proved to be repeatable. The importance of partial or complete input data is, however, stressed as a restricting factor that could limit the performance of the simulation.

4.2.2 Accuracy

The simulation accuracy was tested by comparing the simulated traffic counts in the peak periods to the traffic counts of the counting stations shown in Figure 4.1. Fourie (2009)

suggests that counting station data for a Wednesday be used to represent a typical workday, as the influence of weekend behaviour is arguably at a minimum on this day. Counting station errors were calculated and the simulated trip durations and home departure times were compared to the NHTS data.

To determine the extent of over- and undercounts for each of the four scenarios of allocating initial activity trip starting times to non-work activities, hourly counting station errors were calculated. The error measures include the mean relative error (Equation (4.2)) and the mean relative bias (Equation (4.3)), where n is the number of counting stations, $x_{sim,i}$ is the simulated traffic count at counting station i and $x_{real,i}$ is the actual traffic count at that counting station. A mean relative error of 100% indicates that the simulated versus actual counts ratio is 2:1, and one of 50% indicates a ratio of 1:2. If the mean relative error falls outside the 50%–100% range, it can be noted as a noticeable under- or overcount. The mean relative bias, on the other hand, is preferred as close to 0 as possible. A negative bias indicates an undercount and a positive bias indicates an overcount.

$$e = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{sim,i} - x_{real,i}|}{x_{real,i}} * 100 \tag{4.2}$$

$$b = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{sim,i} - x_{real,i}}{x_{real,i}} * 100 \tag{4.3}$$

Fourie (2009) defines an additional measure that indicates the degree of deviation from perfect agreement of simulated versus actual traffic counts—the counts ratio error. The error is calculated with Equations (4.4) and (4.5) and the standard deviation for the counts ratio error is calculated with Equation (4.6). These measures were used as an additional accuracy test and the results for the peak and surrounding periods are displayed in Table 4.4. Figure 4.7 represents the hourly counting station error and bias per scenario.

$$r_{c,i} = \begin{cases} \frac{x_{sim,i}}{x_{real,i}} - 1, & \text{if } x_{sim,i} \geq x_{real,i} \\ \frac{-x_{real,i}}{x_{sim,i}} + 1, & \text{if } x_{sim,i} < x_{real,i} \end{cases} \tag{4.4}$$

$$\overline{r_c} = \frac{1}{n}\sum_{i=1}^{n} r_{c,i} \tag{4.5}$$

$$\hat{\sigma}_{r_c} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{c,i} - \overline{r_c})^2} \tag{4.6}$$

If the error statistics of the initial implementation are compared to that of the four scenarios, a number of observations can be made on the addition of primary and secondary activities other than *Work* to the simulation:

Table 4.4: Error statistics for simulated versus actual traffic count comparisons

Hour	e[a](%)	b[b] (%)	$\overline{r_c}$[c]	$\hat{\sigma}_{r_c}$[d]
05h00-06h00				
Original[e]	37.00	17.00	0.04	0.82
Scenario 1	83.08	-4.53	-0.99	2.50
Scenario 2	74.98	-49.94	-2.78	3.20
Scenario 3	73.54	-52.88	-2.96	3.32
Scenario 4	70.16	-55.45	-2.95	3.37
06h00-07h00				
Original[e]	39.00	15.00	0.05	0.72
Scenario 1	58.43	17.19	-0.03	1.55
Scenario 2	55.84	-1.56	-0.42	1.56
Scenario 3	54.62	-21.96	-0.99	1.92
Scenario 4	52.76	-29.73	-1.28	2.43
07h00-08h00				
Original[e]	49.00	-8.00	-0.46	1.08
Scenario 1	111.25	93.72	0.88	3.08
Scenario 2	101.97	77.45	0.68	3.13
Scenario 3	89.85	52.57	0.34	2.83
Scenario 4	69.22	14.13	-0.28	2.29
14h00-15h00				
Original[e]	64.00	-45.00	-2.21	2.61
Scenario 1	87.13	-61.97	-7.01	7.12
Scenario 2	88.16	-25.85	-2.38	5.54
Scenario 3	89.80	-20.63	-2.14	5.76
Scenario 4	89.75	-17.76	-1.88	5.91
15h00-16h00				
Original[e]	35.00	-12.00	-0.44	1.07
Scenario 1	65.22	-30.62	-1.40	1.96
Scenario 2	66.64	-23.89	-1.34	2.26
Scenario 3	65.33	-22.90	-1.16	1.88
Scenario 4	66.78	-19.97	-1.19	2.19
16h00-17h00				
Original[e]	36.00	-13.00	-0.44	1.05
Scenario 1	57.63	-14.16	-0.76	1.72
Scenario 2	55.62	-8.75	-0.58	1.60
Scenario 3	56.20	-2.79	-0.53	1.78
Scenario 4	53.40	-6.30	-0.61	1.72

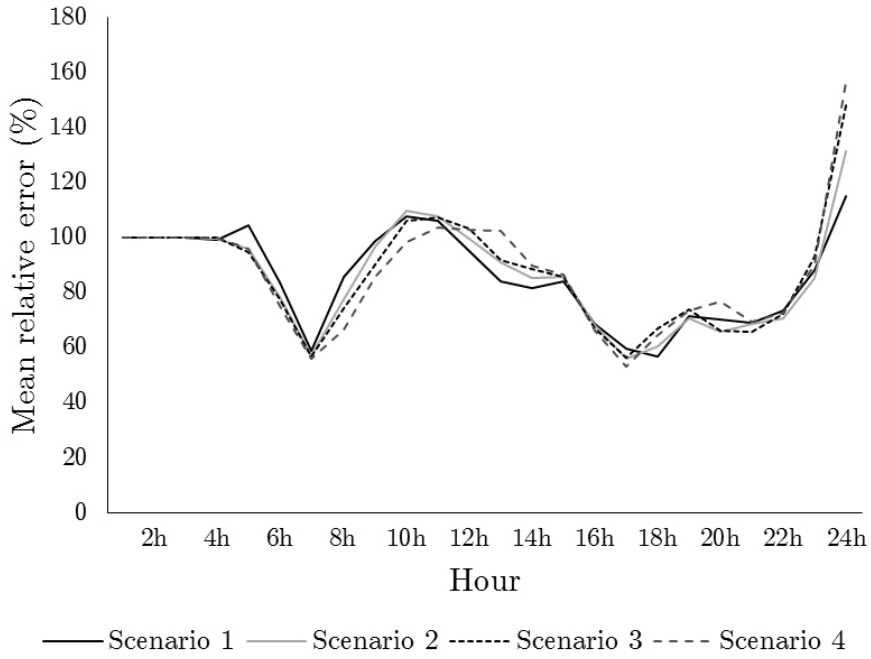
[a] Mean relative error (Equation (4.2))

[b] Mean relative bias (Equation (4.3))

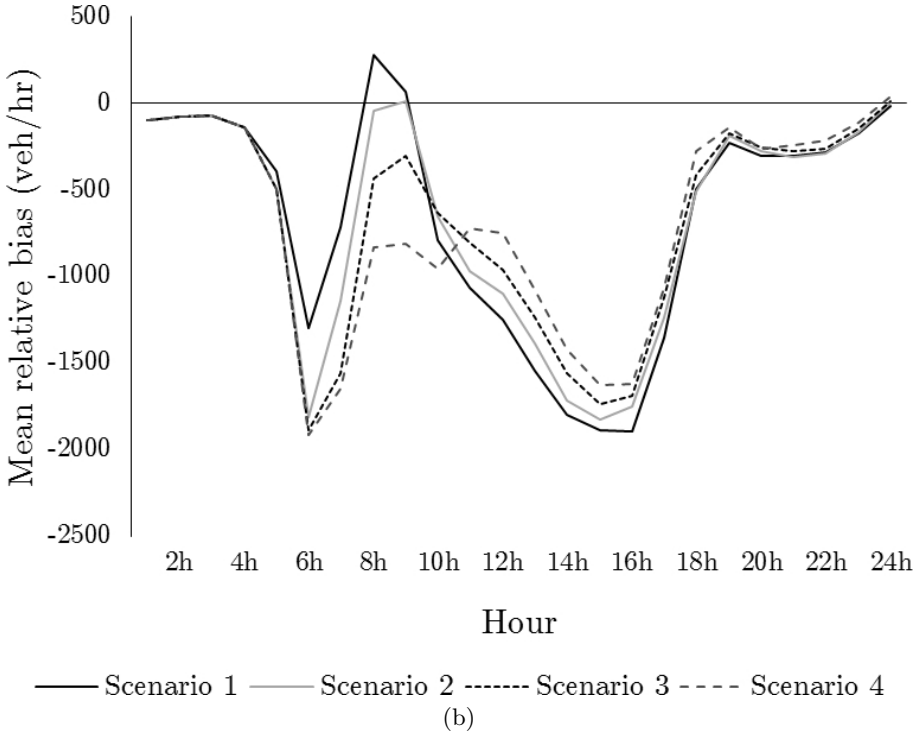
[c] Mean counts ratio error (Equations (4.4)–(4.5))

[d] Standard deviation for counts ratio error (Equation (4.6))

[e] Obtained by Fourie (2009) in the initial *h-w-h* implementation



(a)



(b)

Figure 4.7: Hourly counting station errors: (a) Mean relative error, (b) Mean relative bias.

- The mean relative error of the initial implementation indicates significant undercounts for the peak and surrounding periods, with an exception between 14h00 and 15h00. This study generated mean relative errors that indicate simulated traffic counts that are closer to reality. The only significant overcounts are recorded between 07h00 and 08h00, and only occurs for scenarios 1 and 2. The simulation counts are, however, within acceptable limits for the same period for scenarios 3 and 4. A possible explanation for the overcounts between 07h00 and 08h00 are thus that trips to non-work activities start between two and three hours later than trips to work.

- The mean relative bias of the initial implementation varies between -8 and 17 for the morning peak period and between -45 and -12 for the afternoon peak period. The addition of primary and secondary activities other than *Work* resulted in an improved bias between 05h00 and 06h00 for scenario 1, and between 06h00 and 07h00 for scenario 2. It also improved between 14h00 and 15h00, and 16h00 and 17h00 for scenarios 2, 3 and 4.

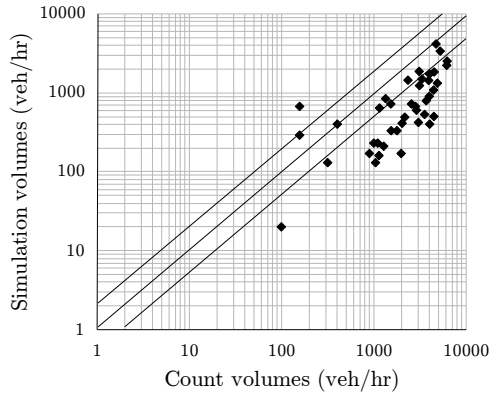
- The mean counts ratio error for the peak and surrounding periods in the initial implementation varies between -2.21 and 0.04. This study improved the mean counts ratio error between 07h00 and 08h00, and 14h00 and 15h00 for scenarios 3 and 4.

- The standard deviation of the counts ratio error varies between 0.72 and 2.61 in the initial implementation. All four scenarios used in this study resulted in a wider spread over the different time periods, varying between 0.72 and 7.12.

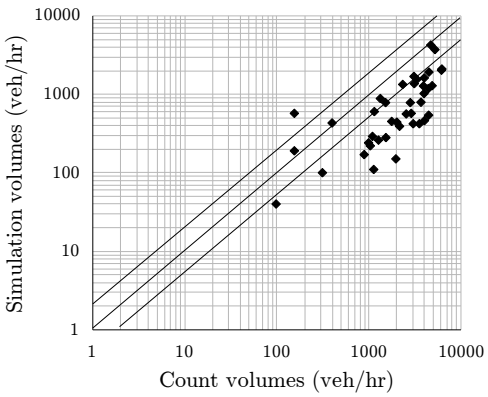
The errors of the different scenarios indicate that, if compared to the initial *h-w-h* implementation, there are less significant undercounts in the simulation if both work and non-work activities are included. The mean counts ratio error and the mean relative bias statistics for the peak and surrounding periods show that the actual traffic counts are best represented in scenarios 3 and 4 respectively. Figures 4.8–4.9 compare the simulated traffic counts of the two scenarios to actual counting station statistics for the peak and surrounding periods. The three diagonal bands in each of the graphs indicate the simulated versus actual count ratios of 2:1, 1:1 and 1:2, starting from the topmost line. Simulated counts within the 2:1 and 1:2 boundaries are deemed acceptable.

Both scenarios 3 and 4 produce undercounts between 05h00 and 07h00 (see Figure 4.8). As the roads become more congested in the peak period, the undercount improves between 07h00 and 08h00. Both scenarios also yield undercounts between 14h00 and 16h00, but reality is more accurately represented between 16h00 and 17h00 (see Figure 4.9).

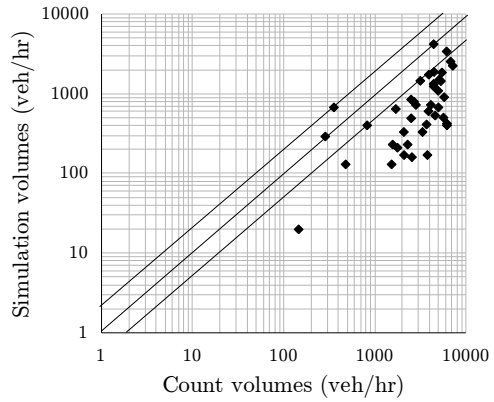
There are a number of reasons for the simulation undercounts. The counting station statistics reflect the sum of the traffic counts of numerous transport modes, of which the private vehicle is one. Since only private vehicles are simulated in this study, the additional vehicles that other transport modes add to the network are not incorporated into the simulated traffic counts.



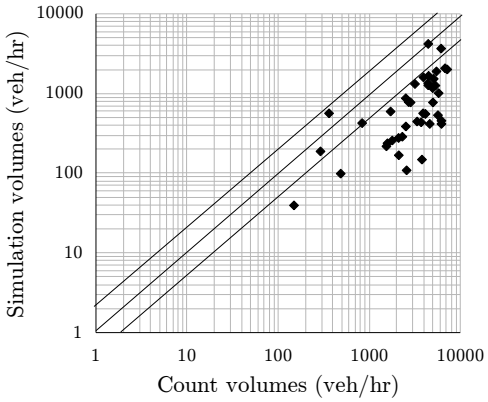
(a) 05h00–06h00, Scenario 3



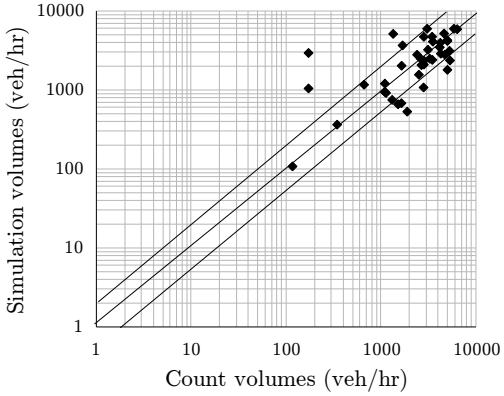
(b) 05h00–06h00, Scenario 4



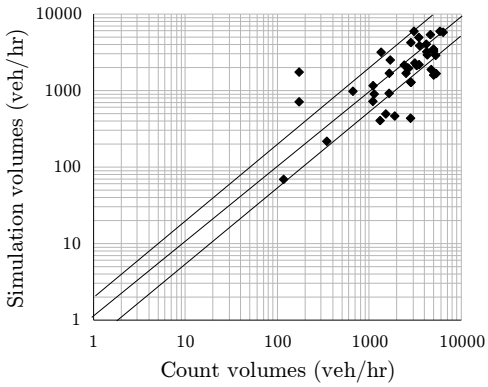
(c) 06h00–07h00, Scenario 3



(d) 06h00–07h00, Scenario 4

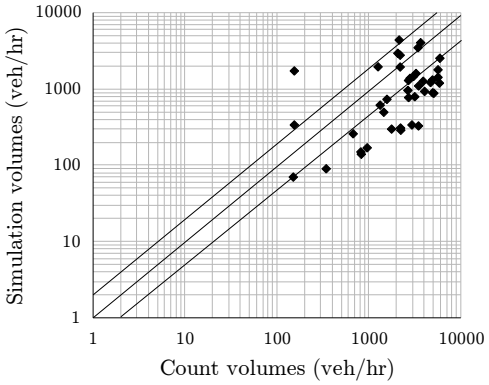


(e) 07h00–08h00, Scenario 3

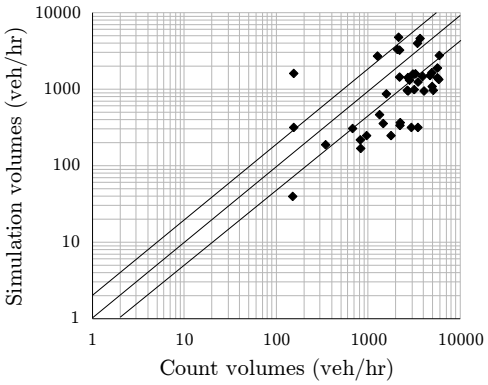


(f) 07h00–08h00, Scenario 4

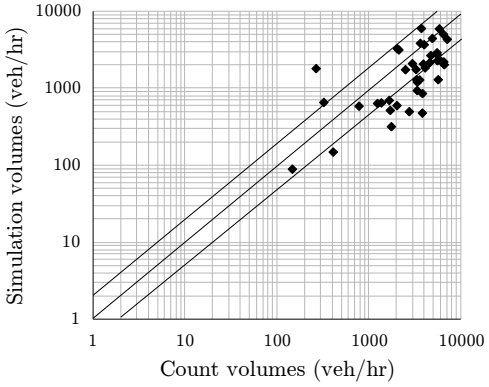
Figure 4.8: Comparison of simulation counts for scenarios 3 and 4 against actual counting station statistics in the morning peak period



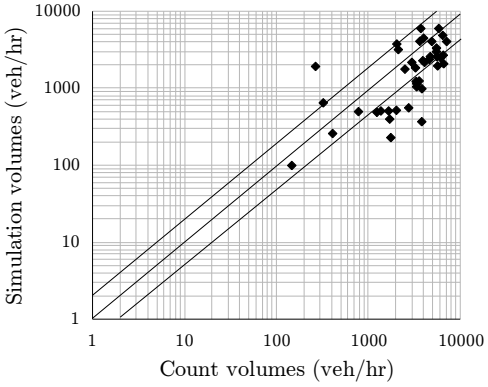
(a) 14h00–15h00, Scenario 3



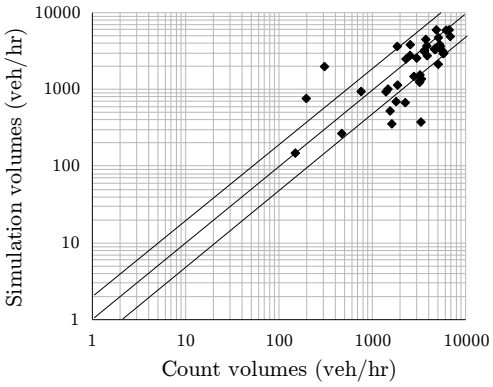
(b) 14h00–15h00, Scenario 4



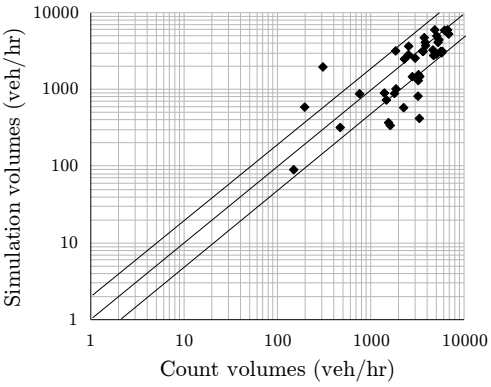
(c) 15h00–16h00, Scenario 3



(d) 15h00–16h00, Scenario 4



(e) 16h00–17h00, Scenario 3



(f) 16h00–17h00, Scenario 4

Figure 4.9: Comparison of simulation counts for scenarios 3 and 4 against actual counting station statistics in the afternoon peak period

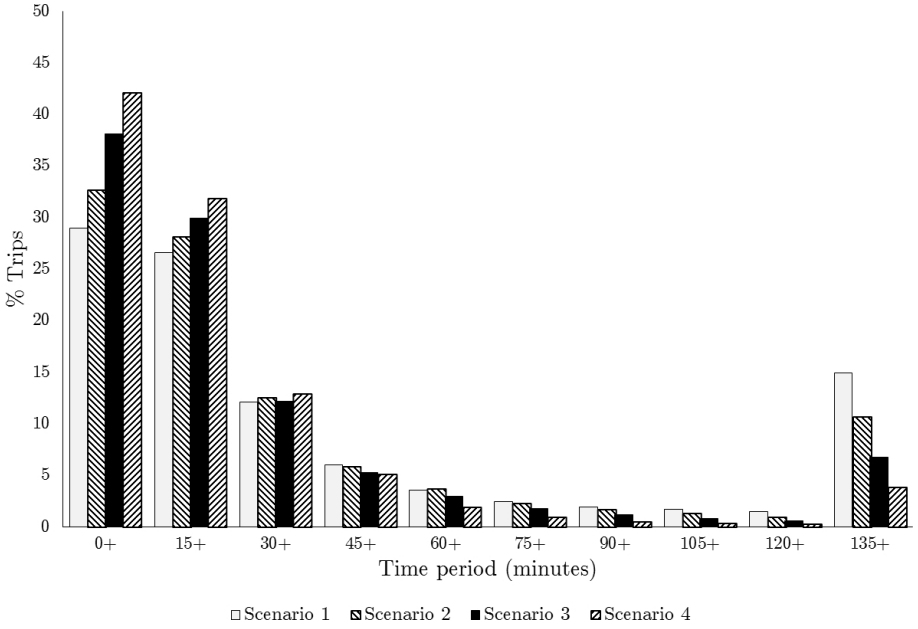
The peak period traffic counts are also influenced by the capacity allocated to the various road links in the network representation. No link has a capacity that exceeds 6,000 vehicles per hour, whereas in reality the capacity sometimes exceeds this and allows for traffic counts above 6,000 vehicles per hour. Roads with higher capacities are usually close to large interchanges where on- and off-ramps are available to improve the traffic flow (Fourie, 2009). The simulated traffic counts are therefore expected to be lower than the actual traffic counts in some instances.

Insofar as outliers are concerned, the majority of overcounts in both scenarios 3 and 4 during the peak and surrounding periods occur at counting station number 4 in Figure 4.1. This counting station is situated on a road classified as a national highway with three lanes per direction, a capacity of 6,000 vehicles per hour and a speed limit of 120 km/h. According to Fourie (2009), some of the roads classified as national highways in the network, in reality only have two lanes per direction and a capacity of 4,000 vehicles per hour. There is thus a discrepancy of 2,000 vehicles per hour between the capacity of the actual road and the road classification in the network. A large number of the feeder routes to the national highways in the network have only one lane per direction and a capacity of 1,000 vehicles per hour in reality, whereas the simulation assumes two lanes per direction and double the capacity. The higher level of available capacity for the highways and their feeder routes in the simulation attracts commuters as a viable alternative route, thereby increasing the traffic counts on these links.

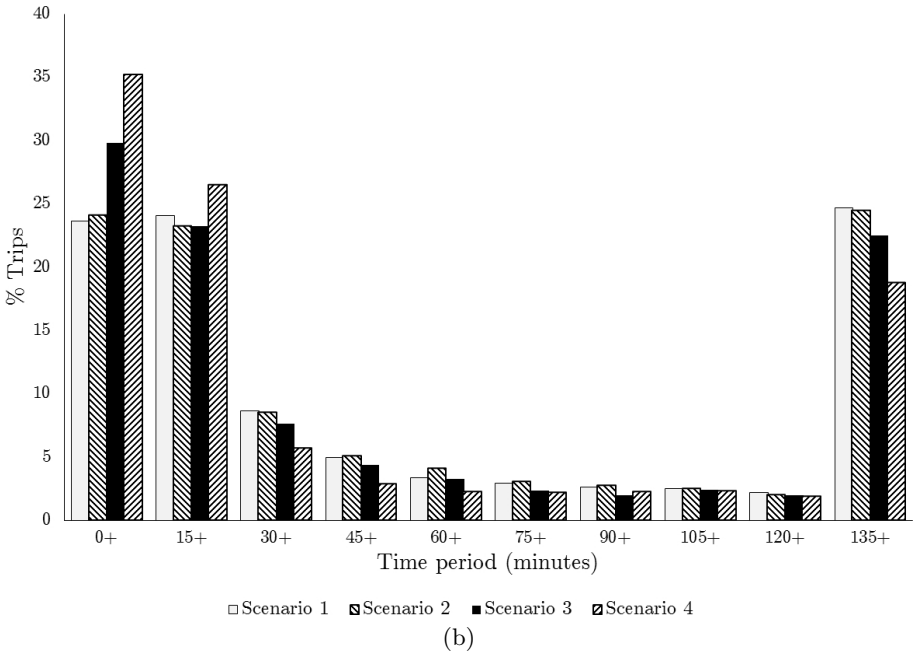
Overall, the morning peak period is best represented between 07h00 and 08h00, and the afternoon peak period between 16h00 and 17h00. Scenario 3 has fewer undercounts than scenario 4 and approximately the same number of overcounts. Figure 4.4 also illustrates that, unlike scenario 3, scenario 4 does not follow the normal demand trend, but reaches a daily maximum in the traditional off-peak period. Scenario 3 simulates both peak and off-peak periods to an acceptable accuracy level, and is thus perceived as the best representation of reality.

The simulated *Work* trip durations of the four scenarios that allocate trip starting times to non-work activities are compared to one another in Figure 4.10a, where the trip durations are grouped into 15 minute intervals. The percentage of trips within the 0 to 15 minute interval increases by 15% between scenarios 1 and 4, while the trips in the 135+ minute interval decrease by 10% between the two scenarios. Figure 4.10b compares the simulated *Education* trip durations of the four scenarios to each other. Between scenarios 1 and 4, there is an increase in the percentage of *Education* trips with durations less than 30 minutes, and a decrease in those with longer travel times. Figure 4.11 groups the trip durations to represent the groups used in the NHTS work and education trip duration data. The simulated *Work* and *Education* trip duration distributions are noticeably different than that of the NHTS data. There are numerous possible reasons for this; some of them are discussed below.

It is possible that the activity location procedure assigned the agents' work locations closer to their home locations than they are in reality, resulting in an increased number

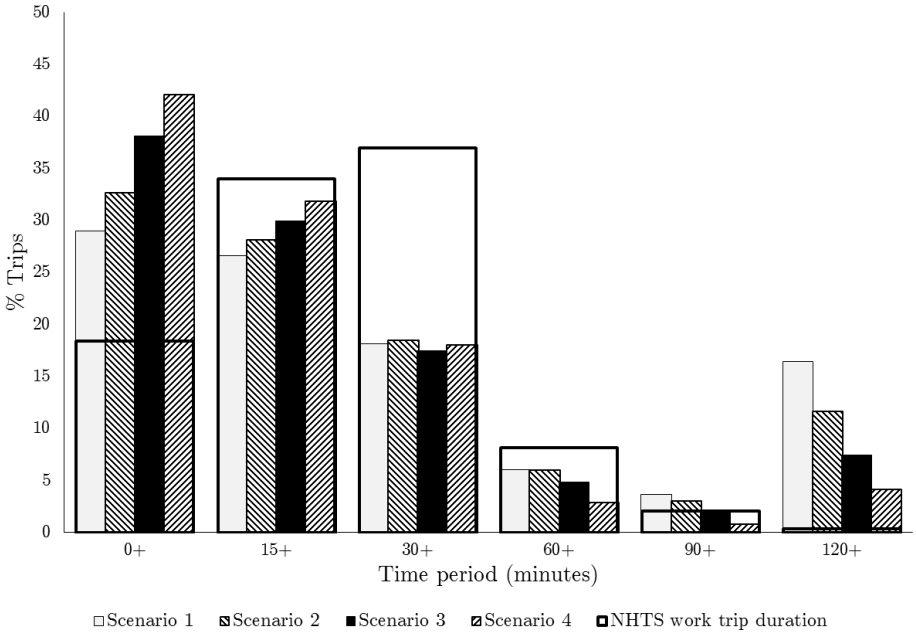


(a)

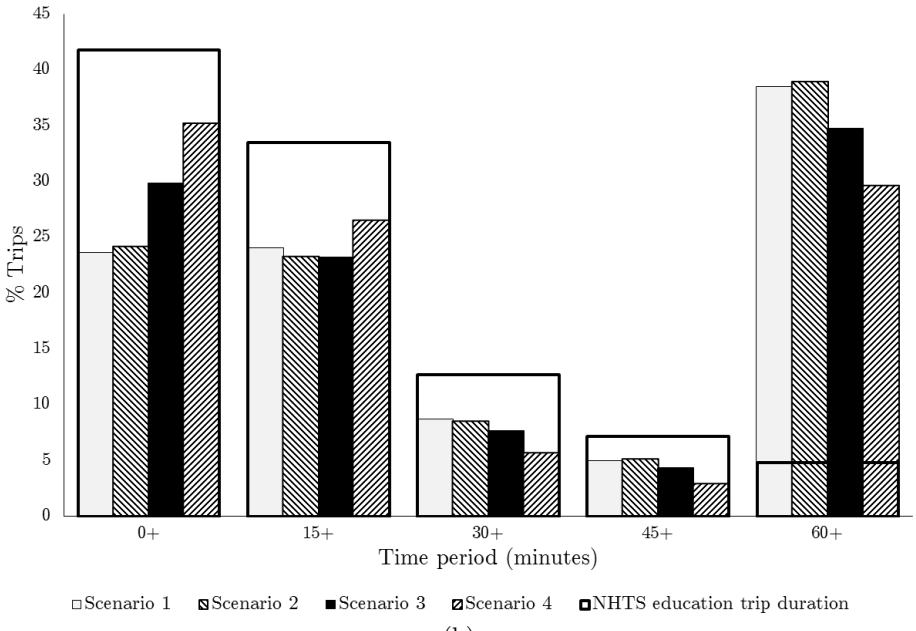


(b)

Figure 4.10: *Work* and *Education* trip duration comparison between the four trip starting time scenarios: (a) Simulated *Work* trip durations for the four scenarios in 15 minute intervals; (b) Simulated *Education* trip durations for the four scenarios in 15 minute intervals.



(a)



(b)

Figure 4.11: *Work* and *Education* trip duration comparison between the four trip starting time scenarios and NHTS data: (a) Simulated *Work* trip durations compared to NHTS work trip duration data; (b) Simulated *Education* trip durations compared to NHTS education trip duration data.

of trip durations to work that is shorter than 15 minutes. The gravity model states that the majority of travel behaviour mimics the gravitational interaction first described by Isaac Newton's law of gravity. This implies that an agent will probably either find work close to his home, or move closer to his work to reduce commuting time between the two. The overestimation of the number of work trips with durations that exceed 120 minutes is the result of the coarser road network representation used in the simulation than the one that exists in reality. The granularity of the road network limits the available alternative routes in the province, since there are fewer roads available for the synthetic agents to use than the actual road network in Gauteng provides. The available roads in the simulation therefore have to carry more traffic than in reality, which inevitably increases the trip duration on these roads.

The simulation underestimates the *Education* trip duration occurrence for most of the travel time categories, except the 60+ minute category. The reason for the overestimation is that most individuals who attend educational institutions on a regular basis, in reality also reside close to the relevant institution. This fact is proven by the high number of short travel times in the NHTS data. It is, however, not portrayed in the simulation, since there is no input data dictating this behaviour. The simulation assigned a *Home* location to an agent before sampling an *Education* location, whereas, in reality, an agent might select a *Home* location based on the location of his *Education* activity.

Another important factor that influences the simulated versus NHTS trip duration comparison, is the sample used in the survey. The income distribution of the sample taken in Gauteng is graphically represented in Figure 4.12. A total of 75% of households included in the survey have a monthly household income of less than R6000, and 17.1% have an income of more than that. From the households included in the sample, only 39.2% of working individuals use private vehicles to commute to work, and 17.1% of individuals commuting to educational institutions use private vehicles. The NHTS also proves that the probability of using a private vehicle increases drastically as the household income increases. It is therefore possible that a more expanded study that covers more households and a broader spectrum of household incomes can influence the trip duration data used to compare the simulated trip durations to in this study.

Even though the trip durations do not correlate with that of the NHTS data, the simulation results represent reality to an acceptable level. The simulated traffic counts are most accurate when non-work trip starting times are delayed with two hours, where the average mean counts ratio error for the peak and surrounding periods is appoximately 71%—which is within the acceptable 1:2 and 2:1 limits. The insight gained from the simulation can be used to support transport related decision making.

4.3 Conclusion

The initial *h-w-h* implementation of MATSim was successfully expanded by including additional primary and secondary activities in the transport demand. Data deficiencies

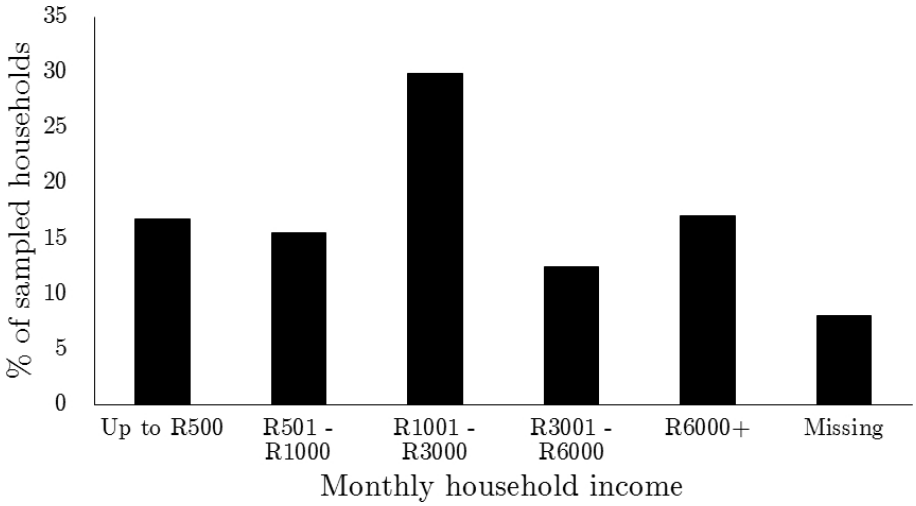


Figure 4.12: Income distribution of NHTS sample

were, to a large extent, overcome by contemplating various scenarios and comparing the results thereof with actual counting station statistics.

The initial implementation simulated private vehicle trips between home and work without considering any other activities, and showed traffic counts for the peak and surrounding periods, but none for the midday traffic between them. This study has improved the simulation accuracy during both peak and off-peak periods. Midday traffic counts are simulated with increased accuracy if non-work activities are considered, indicating that the time of minimum demand between the morning and afternoon peak periods is approximately 14h00. The results of the different non-work trip starting time scenarios also proved that commuters leave home approximately two hours later for non-work trips than for work trips.

Expanding the multi-agent transport model might be simpler and more accurate if all the required data on non-work activities are available. However, the expanded model with limited data can still add value to transport planners and is worthwhile to pursue. The value gained from the addition of other primary and secondary activities surpasses the effort required to include them. There are, however, a number of possible adjustments and improvements that might increase the accuracy of the simulation in the future. The possibilities are discussed in the next chapter.

Chapter 5

Discussion

This chapter summarises the findings of this study and answers the research questions raised in Section 1.3. The study is concluded by discussing possible future research in the field.

5.1 Brief summary of findings

This study created a methodology to generate the initial transport demand for Multi-Agent Transport Simulation (MATSim) in the Republic of South Africa (RSA) and implemented this process for Gauteng province. The transport demand was generated during two phases: population modelling and initial demand modelling.

The first phase constructed a 10% synthetic population from Census data by allocating a home location to every agent. The synthetic population maintained the demographic structure of the actual population to assist in describing the transport behaviour of individuals in Gauteng. Various individual socio-demographic attributes were then assigned to the agents from both census and National Household Travel Survey (NHTS) data by following an approach in *ArcGIS* that involves a combination between probabilistic sampling and rule-based models. The population modelling procedure proved to be consistent and provided repeatable results that were comparable to the source data.

Driver's license ownership and car availability were identified as two important individual attributes in a synthetic population, and the allocation of both these attributes were proven statistically accurate and repeatable.

The second phase transformed the synthetic population into a single `plans.xml` file for every non-work trip starting time scenario. Each of the files contained the daily activity plan for every agent that executes primary and/or secondary activities. The transformation required the execution of six procedures that utilise agents' individual attributes, and census, NHTS and geospatial data to model the transport behaviour of every agent. The first step followed towards a set of daily activity plans was generating an activity chain for every agent. An activity chain consists of the sequential combination of primary and secondary activities, and always starts and ends at the agent's *Home* location.

The primary activities considered in this study were *Home*, *Work* and *Education*, and the secondary activities included *Shopping* and *Leisure*. The second step towards a set of daily activity plans added the necessary transport information to the plan, until it contained the agent's activity chain, the coordinates of every activity, the mode of transport used by the agent to commute to every activity and the duration or starting time of the activity. The agents' daily activity plans succeeded in imitating individual transport behaviour.

Each of the `plans.xml` files were used separately as input to MATSim, which simultaneously executed the daily activity plans of all the agents. This led to emerging phenomena in the transport simulation—macroscopic transport demand patterns and congestion on the roads in Gauteng—that could be compared to the results of the initial implementation and to actual traffic statistics. If compared to the initial implementation that simulated private vehicle trips between home and work, it was confirmed that the addition of non-work activities improves the simulation accuracy during both peak and off-peak periods. The actual traffic on highways and major roads in Gauteng was simulated to an acceptable level, indicating that the initial demand generation procedure is reliable.

The initial *h-w-h* implementation of MATSim was therefore successfully expanded by including additional primary and secondary activities in the transport demand. Data deficiencies were overcome by contemplating various scenarios and comparing the results thereof to actual counting station statistics. The importance of input data has, however, been highlighted by the results, since *Work* activities have the most available data and are simulated more accurately than non-work activities. It can be concluded that the daily activity plans generated in this study provide an improved representation of the travel behaviour of individuals in Gauteng.

5.2 Research agenda

During the execution of the study, a number of areas were identified where further research can improve the quality and accuracy of the study. Each of these areas are briefly discussed below.

5.2.1 Predicting transport behaviour

This study used a set of socio-demographic attributes to predict the transport behaviour of individual agents. Transport behaviour is, however, a complex phenomenon that is best described by human intuition and can thus not be predicted by only allocating a simple set of attributes to an individual. Further research can identify attributes currently excluded from the study, that will enhance the prediction of transport behaviour. Other possible methods of accurately predicting the transport behaviour of individuals can be identified and tested, and additional data sources can be consulted. A possible additional data source is the Gauteng Transportation Study 2000, as used by Fourie (2010) in a comparison between MATSim and EMME/2 .

5.2.2 Multiple plans per agent

In this study, every agent was assigned one daily activity plan for every contemplated scenario of allocating non-work trip starting times. Charypar and Nagel (2005) present an approach where a Genetic Algorithm (GA) is used to keep several instances of daily plans for every agent in the synthetic population. The plans are modified and improved by mutation and crossover, while the bad plans are eventually discarded. This approach could be used in the South African implementation of MATSim.

5.2.3 Activity chain distribution

Balmer (2007) obtains the activity chain distribution in Table 4.1 from the micro-census in Switzerland. The composition of the individual activity chains, as well as the number of occurrences per activity chain, is anticipated to be different in the RSA due to the difference between the two transport environments and the difference in the composition of the populations. Due to a lack of activity chaining data in the RSA, the distribution of activity chains from Switzerland was used and adjusted according to the rules and information provided by the census data in the country. A duplication of the micro-census in the RSA will, however, result in a more accurate activity chaining data for the country, where both the composition of the activity chains and the percentage occurrence per chain are refined.

5.2.4 Activity location

A neighbourhood search that finds activity locations within a specified distance from agents's home locations was used in this study to allocate locations to the second *Work* activity in a chain, as well as *Education*, *Shopping* and *Leisure* activities. Improved intelligence can be used in the neighbourhood search to, for example, assign the activity locations within a specific distance from the preceding activity in the chain and in the direction of the succeeding activity in the chain.

5.2.5 Mode of transport

This study used census data to allocate a mode of transport to every agent. The data, however, only address the modal distribution of individuals attending either work or school. This modal distribution was applied to all agents in the synthetic population. Approximately 43% of individuals in Gauteng do not attend work or school and the procedure can be improved if the relevant modal split data are obtained for secondary activities.

This study allocated one of nine transport modes to every agent in the synthetic population. However, the only mode that was simulated is *private vehicle drivers*. The other transport modes can be included in the simulation to evaluate their influence on the transport system in Gauteng. Another possible enhancement is to allow multiple modes

of transport in an activity chain and mode combinations in a trip, thereby representing reality more accurately.

5.2.6 Trip starting time

The simulation specified that full-time *Work* activities start daily at 08h00 and used NHTS data to assign the trip starting times of these activities. Due to a lack of more complete data, four scenarios were considered to assign trip starting times to non-work activities. The results showed that trips to non-work activities are most likely to start two hours after that of *Work* activities. Improved data on *Education*, *Shopping* and *Leisure* trip starting times can eradicate some of the assumptions and either validate or nullify the various scenarios, thereby further improving the simulation accuracy.

5.2.7 Activity duration

The Basic Conditions of Employment Act, No. 75 of 1997 (BCEA) prescribes a 9 hour duration for the full-time *Work* activity, as used in this study. It is assumed that the sum of the durations of all part-time *Work* activities in an activity chain also equals 9 hours. A study on part-time *Work* activities can validate this assumption and lead to more accurate activity durations.

The duration range of *Education*, *Shopping* and *Leisure* activities were arbitrarily estimated in this study, since no activity duration data on these activities are currently available. The duration estimates can be refined and adjusted if more data are made available, thereby improving the accuracy of the simulation.

The list of future research areas is not deemed to be exhaustive. Many research opportunities that can enhance this study may still arise through further investigations. Even though the research areas investigated in this study can be expanded to improve the simulation accuracy, the integral parts of the study are sufficient to confirm that the addition of primary and secondary activities in the initial demand enhances the quality of the simulation.

Bibliography

Arentze, T. and Timmermans, H. (2005). Information gain, novelty seeking and travel: a model of dynamic activity-travel behaviour under conditions of uncertainty. *Transportation Research Part A*, 39(2-3):125–145.

Balmer, M. (2007). *Travel demand modeling for multi-agent transport simulations: Algorithms and systems*. PhD thesis, ETH Zurich.

Balmer, M., Meister, K., Rieser, M., Nagel, K., and Axhausen, K. (2008). Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. In *2nd TRB Conference on Innovations in Travel Modeling*.

Balmer, M., Nagel, K., and Raney, B. (2004). Large-Scale Multi-Agent Simulations for Transportation Application. *Intelligent Transportation Systems*, 8(4):205–221.

Balmer, M., Rieser, M., Vogel, A., Axhausen, K., and Nagel, K. (2005). Generating day plans based on Origin-Destination matrices: A comparison between VISUM and MATSim based on Kanton Zurich data. In *5th Swiss Transport Research Conference*.

Beckman, R., Baggerly, K., and McKay, M. (1996). Creating synthetic baseline populations. *Transportation Research Part A*, 30(6):415–429.

Charypar, D., Axhausen, K., and Nagel, K. (2006). Implementing Activity-based Models: Accelerating the Replanning Process of Agents using an Evolution Strategy. In *11th International Conference on Travel Behaviour Research*.

Charypar, D. and Nagel, K. (2005). Generating complete all-day activity plans with genetic algorithms. *Transportation*, 32(4):369–397.

Ciari, F., Balmer, M., and Axhausen, K. (2007). Mobility Tool Ownership and Mode Choice Decision Processes in Multi-Agent Transportation Simulation. In *7th Swiss Transport Research Conference*.

Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglone, J., and Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. *Transportation Research Part A*, 41(5):464–488.

Diedericks, D. and Joubert, J. (2006). Towards transportation system integration in the City of Tshwane Metropolitan Municipality. In *Urban Transport XII. Urban Transport and the Environment in the 21st Century.*

Ettema, D., Timmermans, H., and Arentze, T. (2004). Modelling Perception Updating of Travel Times in the Context of Departure Time Choice Under ITS. *Intelligent Transportation Systems*, 8(1):33–43.

Fletterman, M. (2008). Designing multimodal public transport networks using metaheuristics. Master's thesis, University of Pretoria.

Fourie, P. J. (2009). An initial implementation of a multi-agent transport simulator for South Africa. Master's thesis, University of Pretoria.

Fourie, P. J. (2010). Agent-based transport simulation versus Equilibrium Assignment for private vehicle traffic in Gauteng. In *Southern African Transport Conference.*

Frick, M. and Axhausen, K. (2004). Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some new results. In *4th Swiss Transport Research Conference.*

Garling, T. and Axhausen, K. (2003). Introduction: Habitual travel choice. *Transportation*, 30(1):1–11.

Garling, T., Kwan, M., and Golledge, R. (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B*, 28B(5):355–364.

Holt, J., Lo, C., and Hodler, T. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31(2):103–121.

Illenberger, J., Flötteröd, G., and Nagel, K. (2007). Enhancing MATSim with capabilities of within-day re-planning. In *IEEE Ingelligent Transportation Systems Conference.*

Kennedy, M. (2000). Understanding Map Projections. Technical report, ESRI. ArcInfo 8.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation*, 15(1-2):9–34.

Langford, M., Higgs, G., Radcliffe, J., and White, S. (2008). Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, 32(1):66–80.

MATSim Development Group (2008). MATSim controller structure. http://matsim.org/docs/controler. Accessed 12 March 2010.

McNally, M. (2000). *Handbook of Transport Modelling*, volume 1, chapter 3, pages 35–69. Pergamon, 2nd edition.

Merchant, D. and Nemhauser, G. (1978). A model and an Algorithm for the Dynamic Traffic Assignment Problems. *Transportation Science*, 12(3):183–199.

Meulman, J. J. and Heiser, W. J. (1998). *Visualization of Categorical Data*, chapter 20, pages 277–296. Academic Press. Visual Display of Interaction in Multiway Contingency Tables by Use of Homogeneity Analysis: the 2 X 2 X 2 X 2 Case.

Moon, Z. and Farmer, F. (2001). Population Density Surface: A New Approach to an Old Problem. *Society and Natural Resources*, 14(1):39–51.

Recker, W. (1995). The household activity pattern problem: General formulation and solution. *Transportation Research Part B*, 29B(1):61–77.

Rieser, M. (2004). Generating Day Plans from Origin-Destination Matrices. Term project.

Rieser, M., Nagel, K., Beuck, U., Balmer, M., and Rümenapp, J. (2007). Agent-oriented coupling of activity-based demand generation with multiagent traffic simulation. *Transportation Research Board*, 2021(2):10–17.

Schlich, R. and Axhausen, K. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30(1):13–36.

Susilo, Y. and Kitamura, R. (2008). Structural changes in commuters' daily travel: The case of auto and transit commuters in the Osaka metropolitan area of Japan, 1980-2000. *Transportation Research Part A*, 42(1):95–115.

Vovsha, P., Bradley, M., and Bowman, J. (2005). *Progress in Activity-Based Analysis*, chapter 18, pages 389–414. Elsevier. Activity-Based Travel Forecasting Models in the United States: Progress Since 1995 and Prospects for the Future.

Yuan, Y., Smith, R., and Limp, W. (1997). Remodeling census population with spatial information from Landsat TM imagery. *Computers, Environment and Urban Systems*, 21(3/4):245–258.

Appendix A

Extract of `plans.xml` file

```
<plans>
   <person id = "1">
        <plan>
            <act type="home" x="576993.8495" y="7106439.753" end_time="06:34:00"/>
            <leg mode="car"/>
            <act type="leisure" x="590435.221" y="7097923.021" dur="03:19:00"/>
            <leg mode="car"/>
            <act type="home" x="576993.8495" y="7106439.753"/>
        </plan>
   </person>

   <person id = "3">
        <plan>
            <act type="home" x="577124.2275" y="7106486.245" end_time="05:17:00"/>
            <leg mode="car"/>
            <act type="education" x="588192.8331" y="7096916.173" dur="05:00:00"/>
            <leg mode="car"/>
            <act type="home" x="577124.2275" y="7106486.245"/>
        </plan>
   </person>

   <person id = "34">
        <plan>
            <act type="home" x="577037.63" y="7106582.554" end_time="06:29:00"/>
            <leg mode="car"/>
            <act type="leisure" x="589834.3811" y="7105348.157" dur="02:47:00"/>
            <leg mode="car"/>
            <act type="shopping" x="598052.9205" y="7099902.284" dur="03:29:00"/>
            <leg mode="car"/>
            <act type="leisure" x="588455.9928" y="7094940.618" dur="01:51:00"/>
            <leg mode="car"/>
            <act type="home" x="577037.63" y="7106582.554"/>
        </plan>
   </person>

   <person id = "13591">
        <plan>
            <act type="home" x="580438.0817" y="7110669.714" end_time="05:30:00"/>
            <leg mode="car"/>
            <act type="work" x="601661.6042" y="7102571.758" dur="09:00:00"/>
            <leg mode="car"/>
            <act type="home" x="580438.0817" y="7110669.714"/>
        </plan>
   </person>
```

```
    <person id = "120266">
         <plan>
             <act type="home" x="597669.1248" y="7054095.92" end_time="07:02:00"/>
             <leg mode="car"/>
             <act type="work1" x="601474.8875" y="7131284.223" dur="04:30:00"/>
             <leg mode="car"/>
             <act type="shopping" x="607653.7061" y="7124875.543" dur="02:50:00"/>
             <leg mode="car"/>
             <act type="work1" x="601474.8875" y="7131284.223" dur="04:30:00"/>
             <leg mode="car"/>
             <act type="home" x="597669.1248" y="7054095.92"/>
         </plan>
    </person>

    <person id = "213595">
         <plan>
             <act type="home" x="647883.1461" y="7111935.091" end_time="08:47:00"/>
             <leg mode="car"/>
             <act type="shopping" x="642573.5418" y="7104050.765" dur="01:53:00"/>
             <leg mode="car"/>
             <act type="home" x="647883.1461" y="7111935.091"/>
         </plan>
    </person>

    <person id = "450021">
         <plan>
             <act type="home" x="589071.1886" y="7090303.889" end_time="07:39:00"/>
             <leg mode="car"/>
             <act type="work1" x="609377.7819" y="7111986.001" dur="04:30:00"/>
             <leg mode="car"/>
             <act type="work1" x="620922.3383" y="7118250.915" dur="04:30:00"/>
             <leg mode="car"/>
             <act type="home" x="589071.1886" y="7090303.889"/>
         </plan>
    </person>

</plans>
```