LARGE-SCALE MULTIMODAL TRANSPORT MODELLING
Part I: Demand Generation

Johan W. Joubert[a] and Quintin van Heerden[a,b]

[a] Center of Transport Development, Industrial and Systems Engineering, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa.
Tel: 012 420 2843, Fax: 012 362 5103, Email: johan.joubert@up.ac.za

[b] Transport and Freight Logistics, Built Environment, Council for Scientific and Industrial Research, Meiring Naudé Road, PO Box 395, Pretoria, 0001, South Africa.

ABSTRACT:

Recent developments in agent-based transport simulation provide promising results. However, the agent-based approach is frequently criticized for its apparent dependence on vast amounts of, mostly unattainable, data. In this paper we show that the required data is mostly available, even for a country like South Africa. This paper addresses demand generation, the first step in a two-part series, and focuses on three components. Firstly, we demonstrate how a synthetic population of private individual agents is generated using existing and available data. Using iterative proportional fitting (IPF), the population is generated for the Nelson Mandela Bay Metropolitan, and includes 24-hour activity chains for both primary activities such as home, work and education, and also secondary activities: shopping, leisure and other. Secondly, a commercial vehicle population is generated, a novel contribution in an agent-based setting. The commercial vehicles include both intra- and inter-provincial vehicles with different activity chain characteristics. Lastly, the paper demonstrates how an accurate road network is extracted for the Multi Agent Transport Simulation (MATSim) using open data.

1 INTRODUCTION

Two contributions, one by Fourie (2010) and the other by Gao et al. (2010), showed the benefit of using disaggregate agent-based transport simulations over the state of practice equilibrium assignment models. Agent-based models are indeed slower, but the overall duration is still more than acceptable in large decision-making scenarios. Agent-based models are, for one, capable of providing more accurate estimations of the overall travel time distributions. One reason is that the Multi-Agent Transport Simulation (MATSim) model is able to realistically model spillbacks, which is a major contributing factor in the travel times of individuals. Also, the disaggregate approach allows for much richer result sets that can assist in the decision-making process that are not possible with current models.

In this paper we present the typical process followed to generate the necessary inputs required to run an agent-based transport simulation. Instead of using fictitious data, we opt for a real case study, and will use the Nelson Mandela Bay Metropolitan (NMBM). Our choice for study area is based on the scale of the scenario: the metro is home to more than a million inhabitants, making it a fairly large city. Also, the NMBM is rather isolated with not too much through-traffic that may require additional modelling tricks.

The contribution of this paper is methodological. We describe in detail what is necessary and how modellers should proceed to prepare the input data for agent-based models. The objective is to make the new generation of modelling tools more accessible to planners

and modellers, and debunk some of the myths, most notably that developing countries do not have the necessary data to establish such state of the art transport models. Indeed, these models remain sophisticated and require experts to maintain and execute them. But, given the significant infrastructure decisions that are influenced, is more than justified.

The paper is structured as follows. In Section 2 we generate a synthetic population of agents, and we distinguish between private individuals and commercial vehicles, describing the two distinct processes followed. Section 3 describes the use of free and open network data in the form of *OpenStreetMap*, and how the data is converted into a MATSim network, where after we conclude in Section 4.

2 POPULATION SYNTHESIS

A transport simulation model is used to support decisions and what-if analyses regarding scenarios that are considered in real life. It is therefore plausible and commonly expected that the model of choice should accurately represent the reality that it is supposed to influence.

In the case of an agent-based simulation, each agent represents an actor participating in transport in the real world, such as a person who wants to travel from home to work. In this paper we distinguish between two sub-populations, namely private individuals and commercial vehicles, because the two are generated quite differently, and their travel demands are distinct. In the remainder of this section we will address the two sub-populations separately.

The result for both are the same: a set of agents, each with an activity chain made up of *activities* at (most likely) different locations, with each activity pair linked with a travel *leg*. The activity-leg-activity-leg-… sequence for each agent is referred to as the agent's *plan*.

2.1 Private Individuals

When generating a synthetic population of agents that represent the private individuals, there are two steps. Firstly, creating the individuals and the households they represent to resemble the demographics in reality. Secondly, assigning travel demand for each individual so that each behaves in a realistic manner in the model. For example, in reality a child of age 7 is most likely (legally required) to go to school. An agent that is of age 7 is therefore also expected to have an educational activity in its activity chain.

2.1.1 Generating the population

The 2001 census data in South Africa was captured at the resolution of an enumeration area, a small area of manageable size in terms of both population and land area that a single official was able to handle during the census count. When census data is reported, aggregating the individual records to sub place level ensured anonymity. Sub places, in turn, can be aggregated to main places, which in turn can be aggregated to municipalities, district or metropolitan councils, provinces, and ultimately national level.

The sub place tables provided in census data give us, for example, the number of males and females in the area, or the number of Blacks, Coloureds, Asian/Indian and Whites. The tables also tell us how many households there are with one, two, three, etc. individuals in the household, and how many people speak Afrikaans, English, Sepedi, etc. as first language. The tables, however, report independently on each characteristic. It

therefore does not tell us, for example, how many Setswana-speaking black males between 15 and 18 years of age there are in a household or sub place.

Census did provide the detailed records for an anonymous 10% sample of unit records. The sample is representative of the geographic distribution, as well as the different demographic characteristics. For each record in the 10% sample, we know all of the census attributes: age, gender, race, language, household size, etc. In this paper, we made use of the household size, age, gender, race, relation/role in the household, employment and level of school an individual is currently attending.

To generate an entire 100% synthetic population for a sub place we use a two-stage approach. The first stage uses Iterative Proportional Fitting (IPF), an iterative algorithm that adapts the weights (also referred to as expansion factors) associated with a reference sample of observed individuals, i.e. the 10% sample, so that an overall target population can be created that adheres to some control totals, i.e the sub place table totals. That is, how many times should each individual record be duplicated so that, when aggregated, the totals for each attribute category is within a given error margin of the observed category totals. The observed totals were inflated to account for the population growth between the 2001 census and the base year for which the population is generated. It is assumed in this paper that growth was even across all sub places, although individual growth factors could have been used if reliable disaggregate growth figures were available.

In this paper we use an extension to the IPF called Iterative Proportional Updating (IPU) that addresses two IPF limitations: the zero-cell problem and the fact that IPF cannot control the population at both household and individual levels (Mueller & Axhausen, 2011). We used as control totals the gender and race at the individual level, and the household income at the household level.

The second stage is generating the population from the weights fitted in the first stage. We employ Monte-Carlo simulation to sample (with replacement) from the weights: the higher the weight, the higher the probability of that individual being chosen. If we want to generate a synthetic population of n individuals for an area, the random sampling is repeated n times. We will use small capital letters to indicate MATSim-specific data structures.

We start by sampling an individual, the head of the household, and create a MATSim *PERSON*. We note its household size attribute, h. A MATSim *HOUSEHOLD* is created, and the head of the household is added. Then, h-1 more individuals are sampled. For each, a *PERSON* is created with all the attributes associated with it, and the individuals are also added to the *HOUSEHOLD*. We keep on sampling heads of households, and their subsequent household members, until we have a population that is just larger than n. The difference between the required population size, n, and the actual synthetic population can therefore be as much as one minus the largest household size, which is capped to be 10. This difference is still negligible in the overall population size.

The process of fitting and generation is repeated for each area, and we chose the sub place as appropriate demarcation. In the case of the Nelson Mandela Bay Metropole (NMBM), we created the population for 223 sub places, each with its unique fit. The overall population for the NMBM was XXXX households containing a total of YYYY individuals.

For each household we know the annual income and the members of the household. For each individual we know the age, gender, race and employment status.

The home location for each household was a randomly sampled point within the sub place in which the household was created.

2.1.2 Travel demand

During 2004 a detailed travel survey was conducted for the NMBM that included a travel diary. A sample of 1% of the households was surveyed for demographic data, as well as transport-specific behavior and choices. The travel diary portion of the survey asked the respondents to relay the detailed travel information for the previous day. The result was an activity chain of pre-categorized activities for which the location and start- and end times were known. The location's level of accuracy was the transport analysis zone (TAZ), a demarcation different from the census sub place, and in general having finer granularity. A total of 589 TAZs were used in the NMBM. The travel mode used between activities is also indicated in the diary.

Since entire households were surveyed, we had the demographics of each household, as well as the travel patterns of each individual in the household. The first step was to parse the travel survey data into a population in the MATSim format. We created a *HOUSEHOLD* in MATSim and assigned the household income to it. We next created, for each individual in the household, a *PERSON*, and created a *PLAN* for the person that represented that individual's observed travel behaviour. A valid *PLAN* is required to start and end with an *ACTIVITY*, and the *PLAN* elements must be alternating between an *ACTIVITY* and *LEG*.

The survey distinguished between 19 different activity types, which we've aggregated to 8, namely home (h), work (w), primary and secondary education (e_1), tertiary education (e_2), dropping or collecting children from school (e_3), shopping (s), leisure (l) and other (o). The inclusion of e_3 was justified since many children in South Africa are dropped at school using private vehicles, which have a plausibly significant impact on traffic patterns.

A total of 7,129 chains were observed, of which there were 518 unique chains. The five most common chains are indicated in Table 1.

Table 1 – Five most common activity chain structures observed in the Nelson Mandela Bay Metropolitan travel survey.

Activity chain	Number of observations
h-e_1-h	1926
h-w-h	1177
h-o-w-o-h	380
h-o-h	301
h-l-h	243

For each *ACTIVITY* in the *PLAN* we knew the start and end times, and for its location we randomly sampled a point inside the TAZ. For each *LEG* between activities we knew the observed mode. The 14 mode options in the travel survey were aggregated to 8, namely walking, cycling, driving (car driver), driving as passenger, minibus taxi, bus, train or unknown.

Once the survey population was parsed, the next step was to assign an appropriate survey *PLAN* to each *PERSON* created during the IPU's generation stage. For each *PERSON*

we identified the 20 closest survey *PERSON*s that were similar, and with "similar" we mean being in the same age class, income class and household size class. Out of the equally probable 20 *PERSON*s, one was randomly sampled. The sampled survey person's *PLAN* was then assigned to the *PERSON* in the synthetic population.

The location for each *ACTIVITY* in the *PLAN* was then assigned to a physical *FACILITY*, and here we used three sources. Firstly, we used the Spot5 satellite imagery for which we had a data set of physical structures inferred through image processing. Secondly, we captured all of the shopping centres in *OpenStreetMap* that are members of the South African Council for Shopping Centres (SACSC). *OpenStreetMap* is a crowd-sourced map created by contributors under an open data license. The map data is free for download and use and provide rich attributes in XML format. Thirdly, we parsed a variety of amenities from *OpenStreetMap* including educational, shopping and leisure facilities. The facilities are distinguished based on the feature types specified[1].

For home (h) activities, the Spot5 building closest to the household's sampled home location was assigned as *FACILITY*. For work (w) activities, the process is slightly more involved. First we check if the activity's sampled location is within a given threshold of a SACSC shopping centre. We chose an arbitrary threshold of 500m. If it is within the threshold, the person is assumed to work at the shopping centre and the location is changed to that of the SACSC *FACILITY*. If not, we check if the work location is within a given threshold of any other amenity facility parsed from *OpenStreetMap*. Here we used an arbitrary threshold of 20m. If it is within the threshold, the person is assumed to work at that amenity facility and the location is changed to that *FACILITY*. If not, the Spot5 facility closest to the work location is assigned as the place of work.

For educational activities the closest educational facility is chosen, irrespective of the level of education provided. The reason was that the level of education is not known with certainty. We acknowledge that there may indeed be exceptions, but that the majority of scholars attend the closest school, and that primary and secondary schools are in fairly close proximity to one another.

OpenStreetMap amenities were parsed based on their feature type and those that were classified as providing either shopping (s) or leisure (l) activities were used. In the case of the SACSC facilities, these were considered to provide both shopping and leisure activities, as well as other (o) activity types that may include banking, post office or even medical activities as these shopping centres frequently provide a spread of activities. So, for shopping (s) activities in the *PERSON*'s *PLAN*, the closest facility offering shopping as an activity type was selected. Similarly, leisure activities were assigned facilities.

For each *LEG*, the observed mode of the person was retained. In the end we have an entire *POPULATION* of *PERSON*s, each with a *PLAN* that consists of alternating *ACTIVITY*s and *LEG*s.

2.2. Commercial Vehicles

The activity chain structure of commercial vehicles differs quite distinctly from that of private vehicles. Commercial vehicles travel for different reasons, and perform many more activities, than private vehicles. For this reason, it is imperative to model commercial

[1] For a list of features, refer to http://wiki.openstreetmap.org/wiki/Map_Features

vehicles independently from private vehicles and not merely inflate the private vehicle models to account for freight movement, as is the case in many state-of-practice models.

2.2.1 Data preparation

In this paper, as in Joubert et al. (2010); Joubert and Axhausen (2011); Joubert (2012); Van Heerden and Joubert (2012), we used a large GPS dataset of 41 711 commercial vehicles that are subscribed to the cTrack tracking service. We followed the extraction process of Joubert and Axhausen (2012) to extract activity chains from the dataset. An activity chain consists of activities, where we typically distinguish between *major* and *minor* activities. Major activities are those activities in excess of 300min in duration, typically depot-stops; and minor activities are those activities shorter than 300min in duration, typically drop-off or collection activities. A complete chain constitutes a chain consisting of at least two major activities, one as the start activity and one as the end activity, with any number of minor activities in-between.

Any activity chain with no activities inside NMBM was omitted. From the remaining chains, all chains with more than 60% of their activities inside the NMBM boundaries were considered to be intra-provincial chains and the rest classified as inter-provincial chains. Chains with activities outside the area were cleaned: the activities outside the NMBM boundaries were removed and replaced with an *entry* or *exit* type of activity where the vehicle crossed the boundary of NMBM, depending on the direction of travel. The inter-provincial chains were further broken down into two types of chains: in-out, which are those chains starting with an *entry* activity, performing some activities inside NMBM and thereafter leaving the area again with an *exit* activity; and out-in chains, which are those starting inside the study area with a *major* activity, leaving the study area with an *exit* activity, returning to the study area with an *entry* activity after which they end with a *major* activity. Table 2 summarizes the number of observed chains by type.

Table 2 – Commercial vehicle activity chains observed in the Nelson Mandela Bay Metropolitan from GPS analyses.

Activity chain	Number of observations	% of total
Intra-provincial	6756	69.6
Inter In-out	1819	30.4
Inter Out-In	1128	

From the activity chains, we generated a complex network as in Joubert and Axhausen (2012), which describes the connectivity between pairs of locations where activities take place. When a vehicle travelled between two activity locations, the weight of the link connecting them was incremented. This process was repeated for all activity chains and two separate complex networks were created for intra- and inter-provincial vehicles.

Artificial *facilities* were created at the locations where activities took place and at each facility, the duration of activities was noted and an *activity duration deciles distribution* was calculated.

We then created a three-dimensional matrix, where the three dimensions are *start hour, number of activities in chain,* and *chain duration*. Separate matrices were created for intra- and inter-provincial vehicles.

2.2.2 Generating the population

In this section we again refer to MATSim objects by utilizing small capitals. Separate populations were generated for the intra- and inter-provincial activity chains since their chain structures differ slightly. For each commercial vehicle, we generated a *PERSON*. Similar to private vehicle commuters, each vehicle also has a *PLAN.*

A plan contains both *ACTIVITY* and *LEG* elements, where an *ACTIVITY* can be of the type *major, minor, entry or exit*; and a *LEG* connects the activities by means of transportation with a commercial vehicle. For a *PLAN* to be valid, it must start and end with an *ACTIVITY*, and the *PLAN* elements must be alternating between an *ACTIVITY* and *LEG*.

An intra-provincial chain's first and last *ACTIVITY* is of type *major*, and we sampled a random starting location from the intra-provincial complex network and allocated this location to the *ACTIVITY*. Next we sampled, from the three-dimensional matrix, a starting hour. Given the starting hour, we sampled the number of activities in the chain. Given both the starting hour and number of activities in the chain, we sampled the duration of the chain. For each of the number of activities sampled, we generated an *ACTIVITY* of type *minor* and sampled a location from the complex network again, weighted by the out degree from the starting location. The duration of the minor activity was sampled from the *activity duration deciles distribution* for that facility where the activity took place. The last *ACTIVITY* in the chain was of type *major* again, and the end time of the chain was calculated from the start time and chain duration sampled, by adding the duration to the start time. The location of the last major *ACTIVITY* was again sampled from the complex network by considering the out degree from the last minor *ACTIVITY*, but only considering those locations with a *major ACTIVITY* type. The start times of minor activities were spread evenly between the start and end time of the chain.

We followed the same procedure for inter-provincial *in-out* and *out-in* chains, but the chain structures differed. The first *ACTIVITY* in an *in-out* chain was of type *entry* and the last *ACTIVITY* of type *exit*. For *out-in* chains, we generated two "parts". The first part was the part of the chain that started inside the area, thus starting with an activity of type *major*, and ending with an *ACTIVITY* of type *exit*. The second part of the chain is the part where the vehicle returns to the study area with an *ACTIVITY* of type *entry* and eventually ending with an *ACTIVITY* of type *major* again.

3 NETWORK

As with assigning facilities to activities in the population, we use *OpenStreetMap* data for the road network. *OpenStreetMap* allows for a variety of tags to be added to each road segment, and there are guidelines specifically for capturing the South African road network[2]. Among these tags may be the number of lanes, and the road classification. There is a standard interface in MATSim that converts the *OpenStreetMap* data into a MATSim *NETWORK*. Conversion defaults used in the NMBM are indicated in Table 2.

The defaults are used only when an overriding tag is not available. For example, if a road segment with tag **`highway=secondary`** has a tag **`lanes=2`**, then the default of one lane will not be used. Also, the user can set the defaults used during conversion if local knowledge deems necessary.

[2] See http://wiki.openstreetmap.org/wiki/South_African_Tagging_Guidelines

During the conversion, bidirectional road segments are replaced with two segments, each accommodating flow in only a single direction.

Table 2: Default conversion values to be used when converting *OpenStreetMap* data into a MATSim network.

Highway type[1]	*Number of lanes*	*Maximum speed (km/h)*	*Capacity (vehicles per hour per lane*[2]*)*	*One way*
Trunk	1	120	2000	No
Motorway	2	120	2000	Yes
Motorway link	1	80	1500	Yes
Primary	1	80	1500	No
Primary link	1	60	1500	No
Secondary	1	80	1000	No
Tertiary	1	60	1000	No
Minor	1	45	600	No
Residential	1	45	600	No
Living street	1	15	300	No
Unclassified	1	60	800	No

[1] Based on the OpenStreetMap definitions.
[2] Per direction, if applicable, based on the one-way attribute.

4 CONCLUSION

In this first paper of the two-part series we showed how available data could be used to create the necessary inputs for an agent-based transport model, specifically MATSim.

We showed how one can successfully generate a synthetic population of private individual agents utilizing this available data. We generated a synthetic population by using an iterative proportional fitting (IPF) method that included activity chains of individuals spanning over 24 hours and included both primary activities and secondary activities.

Next, we described how to generate a commercial vehicle synthetic population that includes both intra- and inter-provincial activity chains. The process involves the extraction of activity chains from GPS logs and the generation of a synthetic commercial vehicle population from the observed activity chain characteristics. We also showed how to extract an accurate road network for use in MATSim from open data on *OpenStreetMap.*

REFERENCES

Fourie, P, 2010. Agent-based transport simulation versus equilibrium assignment for private vehicle traffic in Gauteng. In Proceedings of the Southern African Transport Conference 2010. SATC.

Gao, W, Balmer, M, and Miller, EJ, 2010. Comparison of MATSim and EMME/2 on Greater Toronto and Hamilton Area Network, Canada. Transportation Research Record: Journal of the Transportation Research Board, pages 118-128.

Mueller, K, and Axhausen, KW, 2011. Hierarchical IPF: Generating a synthetic population for Switzerland. Technical report, Transport Systems Planning and Transport Telematics (VSP), Technical University of Berlin.

Joubert, JW, and Axhausen, KW, 2012. A complex network approach to understand commercial vehicle movement, Transportation, NYP, doi: 10.1007/s11116-012-9439-0.

Joubert, JW, and Axhausen, KW, 2011. Inferring commercial vehicle activities in Gauteng, South Africa. Journal of Transport Geography, 19(1):115-124.

Joubert, JW, Fourie, PJ, and Axhausen, KW 2010. Large-scale agent-based combined traffic simulation of private cars and commercial vehicles. Transportation Research Record, 2168:24-32.

Joubert, JW, 2012. Analysing commercial through-traffic. In Procedia - Social and Behavioral Sciences, Vol. 39, pp. 184-194.

Van Heerden, Q, and Joubert, JW, 2012. Commercial vehicle behaviour: analysing GPS records. In: 42nd Computers and Industrial Engineering (CIE) Conference, Paper 99.