Creating an open MATSim scenario from open data: The case of Santiago de Chile

Benjamin Kickhöfer[a,*], **Daniel Hosse**[b], **Kai Turner**[a], **Alejandro Tirachini**[c,*]

[a] Transport Systems Planning and Transport Telematics Group, Technische Universität Berlin

[b] Innovation Centre for Mobility and Societal Change, Berlin

[c] Transport Engineering Division, Civil Engineering Department, Universidad de Chile

* Correspondence addresses: benjamin.kickhoefer@gmail.com, alejandro.tirachini@ing.uchile.cl

March 1, 2016

Preferred citation style: Kickhöfer, B., D. Hosse, K. Turner, and A. Tirachini (2016). "Creating an open MATSim scenario from open data: The case of Santiago de Chile". VSP Working Paper 16-02. See http://www.vsp.tu-berlin.de/publications. TU Berlin, Transport Systems Planning and Transport Telematics.

Abstract

MATSim is an activity-based transport simulation framework designed to simulate large-scale scenarios. This paper describes the creation process of the publicly available MATSim scenario of Santiago de Chile. Three open data sources are used: (i) car network information from OSM, (ii) public transport supply data from GTFS, and (iii) travel diaries from Santiago's 2012 Origin-Destination Survey. The first version of the resulting scenario is described, which is meant to provide a platform for researchers and practitioners in the public and private sector. It can be used to answer different research questions on transport policy interventions (e.g., public transport reforms, road pricing, emission modelling), to obtain accessibility measures, to solve location problems or to develop business ideas based on the simulated mobility of individuals in Santiago. One goal is to constantly increase the quality of the scenario with the help of future users who invest time to make it more sophisticated, and feed their improvements back to the original version. The open availability of such detailed scenario is rather unique. It might become a role model for administrations all around the world to realize the power of open data initiatives when it comes to transparent decision making and the stimulation of innovation activity in the private sector.

Keywords: Agent-based simulation, Open data, Scenario generation, Policy analysis

1 Introduction

Santiago, the capital city of Chile, was founded in 1541 by Spanish conquerors in Chile's central valley. Today, the Great Santiago conurbation (Santiago, Puente Alto and San Bernardo) has a total population of 6 million people in an area of 641 square kilometers (Muñoz et al., 2015), making it the most populated city of Chile by far. Several stand-alone hills, the Andes Mountain Range and the Mapocho River are natural boundaries of the city and help to shape its landscape. The city suffers from air pollution particularly in the winter season, due to emissions coming from fixed and mobile sources which concentrate in the Santiago valley.

Socioeconomic differentiation reflected in residential segregation is one of the most common forms of urban segregation in Latin America, and Chile is no exception with high levels of residential segregation in Santiago and regional capitals like Concepción and Valparaíso (Sabatini et al., 2001). Sabatini et al. (2001) describe a common pattern of residential segregation in Latin American cities in countries like Argentina, Brazil and Chile, in which socioeconomic elites concentrate in a single area that usually grows like a cone from the city center towards the periphery in one chosen direction. In the case of Santiago, this "chosen direction" is the east and north-east (towards the Andes Range), in *comunas* (municipalities) such as Providencia, Las Condes, Vitacura and Lo Barnechea. Steady economic growth has modernized Santiago over the past decades, and has slowly moved the Central Business District (CBD) from the historical geographical center (the Santiago municipality) towards the wealthier eastern districts, which induces long travel distances due to the uneven distribution of work and education opportunities across the city.

Multiple interventions in Santiago's transport system in the past 20 years make this city an interesting case study for the analysis of alternative transport policies. Santiago has a Metro (subway) network of five lines over 104 kilometers, with two new lines to be launched in 2017 and 2018, adding 37 kilometers to the network. In the city there is a full-scale integrated public transport system launched in February 2007 – the Transantiago system (Muñoz et al., 2014), which has fare integration between all bus routes in the city and Metro, through the use of a single prepaid (smartcard) payment method. Transantiago has improved its service (including e.g., the provision of user information, more direct routes and express services) after a beginning full of problems, however the quality-of-service issues still attached to the system (mainly passenger crowding in Metro and busy bus routes, unreliability of bus services and deterioration of older buses) are usually put forward as one of the causes for the observed decline in public transport

modal share in Santiago. On the other hand, between 2004 and 2006 a network of 200 kilometers of tolled urban highways started to operate, which benefit car drivers (the higher income group) at the expense, in some cases, of segregating local communities in lower income suburbs (like the Vespucio Sur highway in the south of the city, Tirachini (2015)). The air pollution problem is tackled, in part, by introducing plate-number based car driving bans on the most polluted days (Barahona et al., 2015). All these elements make Santiago an appealing case study for the application of a metropolitan-scale transport and activity simulator, such as MATSim[1].

The remainder of the paper is organized as follows: Sec. 2 briefly describes the model that is used in the present paper for running simulations of the scenario. Sec. 3 presents the input data for scenario generation as well as the necessary steps to convert that data into meaningful MATSim input. Sec. 4 provides detailed information of the simulation approach and shows results of first calibration/verification efforts. Finally, in Sec. 5, the paper is concluded and possible future applications of the scenario are proposed.

2 MATSim

MATSim is a transport simulation framework designed to simulate large-scale scenarios in reasonable computation time. The following text provides a short overview of the basic functioning of MATSim, for detailed information, please refer to Horni et al. (2016a).

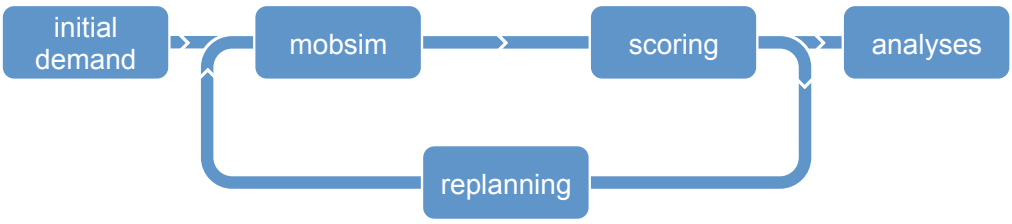


Figure 1: MATSim cycle (Horni et al., 2016b)

Minimal inputs to the model are network data and daily plans of all individual travelers (= agents), forming the initial demand (see Fig. 1), as well as various configuration parameters. Every agent in the simulation, learns and adapts to the system within an iterative process. This process

[1] See www.matsim.org.

is composed of the following three steps:

1. **Simultaneous execution of plans (mobsim)**: In the first step, daily plans of all individuals are executed simultaneously on the network. For all vehicles, a first-in-first-out queuing model is used to describe the traffic dynamics on each street segment (= link) of the network.

2. **Evaluation of plans (scoring)**: In order to model the choice between multiple potential daily plans, executed plans of all agents are evaluated using a utility function, indicating the performance (or score) of the plan. It is represented by:

$$S_{plan} = \sum_{q=0}^{N-1} S_{act,q} + \sum_{q=0}^{N-1} S_{trav,mode(q)} \tag{1}$$

where N is the number of activities, $S_{act,q}$ is the utility from performing activity q, and $S_{trav,mode(q)}$ is the (typically negative) utility for traveling to activity q. In short, the utility earned from performing an activity is given by[2]

$$S_{act,q} = \beta_{dur} \cdot \ t_{typ,q} \cdot \ln(t_{dur,q}/t_{0,q}) \tag{2}$$

where $t_{dur,q}$ and $t_{typ,q}$ are actual and typical durations of activity q, respectively; β_{dur} is the marginal utility of activity duration or the marginal utility of time as a resource; $t_{0,q}$ is the minimal duration, which essentially has no effect as long as dropping activities is not allowed. The mode-specific utility from traveling is described by:

$$S_{trav,mode(q)} = C_{mode(q)} + \beta_{trav,mode(q)} \cdot t_{trav,q} + \beta_m \cdot \Delta m_q \tag{3}$$

where $C_{mode(q)}$ is the Alternative Specific Constant (ASC), $t_{trav,q}$ is the travel time and Δm_q is the change in monetary budget of the trip between activity q and $q+1$; $\beta_{trav,mode(q)}$ is the direct marginal utility of time spent traveling, which comes on top of the marginal utility of time as a resource; and β_m is the marginal utility of money. For the specification of the parameters in the simulation, see later in Sec. 4.

3. **Change of plans (replanning)**: After executing and scoring plans, a new plan is generated for a predefined share of agents. The new plan is generated by modifying an existing plan

[2]See Charypar and Nagel (2005) and Nagel et al. (2016), Sec. 3.2, for a more detailed description.

with respect to predefined choice dimensions (see later in Sec. 4).

The repetition of the above steps eventually results in stabilized simulation output which can then be used for further analysis.

3 Data

3.1 The 2012 origin-destination survey

The travel demand and activity patterns of the MATSim Santiago scenario are based on the travel and activity data collected in the 2012 Origin-Destination Survey (ODS), whose database and results were released to the public in March 2015.[3]

3.1.1 Overview

The surveyed area encompasses 45 *comunas* of the Santiago Metropolitan Region, with an estimated population of 6.65 million people. The survey goes beyond the Great Santiago Area to include the neighboring municipalities of Colina, Lampa, Pirque, Calera de Tango and Melipilla. The total area has 2 million households with an average of 3.24 persons per household. The ODS was conducted between July 2012 and November 2013 by face-to-face interviews at citizens' homes. People were interviewed about all trips within public areas and the conducted activities on one particular day. The sample size is 18 000 randomly chosen households along 866 zones that were defined for the survey. The sampling method is Probability Proportional to Size (PPS) for the selection of blocks; in each block the number of households to be chosen for the survey increases with the number of households that are formally registered in each block.[4] Out of the 18 000 households,

- 11 000 were interviewed about trips and activities on a working day in the normal period.
- 7 000 were interviewed about trips and activities on weekends in the normal period, and on working days and weekends in the summer (holiday) period.

Fig. 2 shows a map of the survey area and zones. The Great Santiago Area is highlighted by an ellipse, in which 91% of the population is concentrated. In the Great Santiago Area, 18% of the surface is allocated to roads (Muñoz et al., 2015).

[3] The survey form, reports and full database are available at the website of Chile's Transport Planning Office (SECTRA), http://www.sectra.gob.cl/biblioteca/detalle1.asp?mfn=3253, accessed 16 August 2015.

[4] For details on the sampling method, see Sectra (2014, p.77).

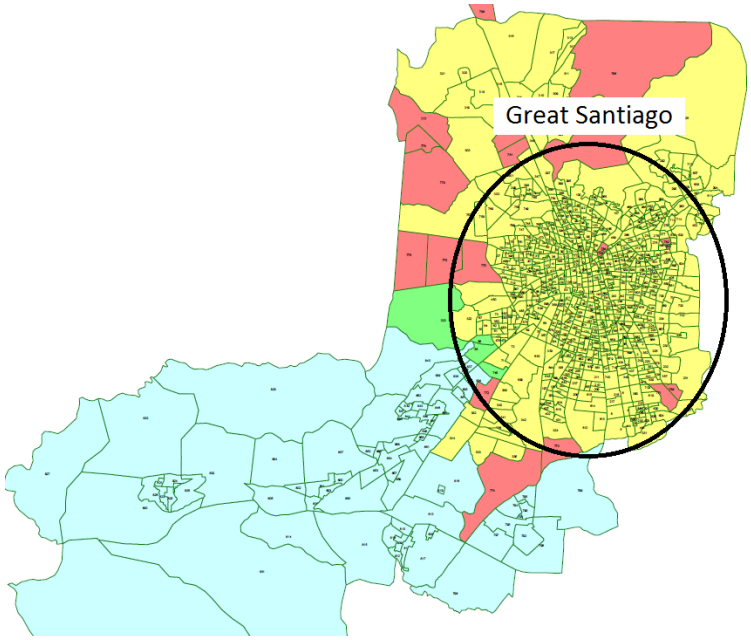


Figure 2: 2012 ODS study area and zones, adapted from Sectra (2014).

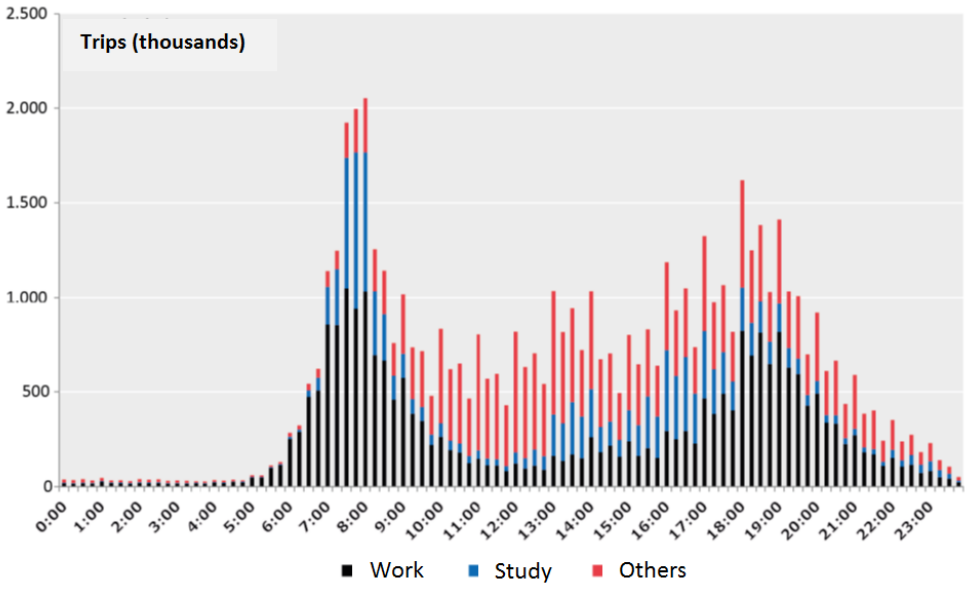


Figure 3: Trip purpose distribution over time-of-day (Sectra, 2014).

Table 1: 2012 ODS general results after expansion (Sectra, 2014).

Households	2 051 310
Persons	6 651 654
Persons/household	3.24
Vehicles/household	0.57
Vehicles/1 000 inhabitants	174.5

The 2012 ODS raw data was in the next step expanded to the full population of the Santiago Metropolitan Region, using a methodology described in Contreras (2015). The general results of that expansion are shown in Tab. 1. Regarding trip purpose, 32.4% of trips are for work (commuting and other work-related trips), 19.5% are for study and 48.1% are for other purposes. The distribution of trip purpose over time-of-day is shown in Fig. 3, distinguishing between work, study and others. It is estimated that on a normal working day, there are 18.5 million trips, from which 38.5% are by non-motorized means (walking and cycling).

Fig. 4 shows different public transport options in Santiago: a Transantiago bus, a Metro train and *colectivos* (shared taxis), which are black vehicles that run on fixed routes and have a fixed fare. Around 25% of the total trips are made using the Transantiago public transport system, out of which 52.4% are bus-only trips, 22.2% are metro-only trips and 25.4% are combined bus-metro trips. Car travel has a modal share of approximately 26% of the total trips (for an overview of the modal split, see later in Tab. 3).

When comparing these numbers with those of the previous (2001) Origin-Destination Survey, one notices that public transport trips have gone down by 2.4%, that car trips have gone up by 39% and that the bicycle modal share has almost doubled from 2.1% to 3.9%. In 2001, modal shares of walking, public transport and car were 38.3%, 30.1% and 21.0%, respectively. Muñoz et al. (2015) present several possible reasons to partly explain the observed shift from public transport to car in Santiago's modal share: (i) the steady increase in car ownership in Santiago, at an annual rate of 4.4%[5], (ii) the development of a network of 200 kilometers of urban highways in Santiago, and (iii) the difficulty of the Transantiago public transport system to provide a reliable service in order to stop user migration, in spite of having increased the Metro network between 2001 and 2012 from 40 to 104 kilometers, and having a fare subsidy estimated at 40% of the annual system costs.

[5]There are 1.4 million motorized vehicles in Santiago, from which 1.2 million are passenger cars.
[6]See https://www.flickr.com/photos/empezardecero, accessed 5 February 2016.

Figure 4: Public transport travel options in Santiago: Transantiago bus, Metro train and colectivos (shared taxis). ©Ariel Cruz.[6]

.

3.1.2 From the 2012 ODS data to MATSim input

In total, 60 054 individuals were interviewed in the 2012 ODS, with a total of 113 591 trips. For the generation of the synthetic MATSim population, it is important that the coordinates of the activity locations and the transport modes for the connecting trips are available. Where no exact coordinates are available, the *comuna* (municipality) tag in the data was used to generate a random coordinate based on shape files.[7] If only the travel time and transport mode to the next activity is available, but not the exact coordinates of that activity, its coordinates are chosen randomly around the previous activity on a circle with the radius of the distance that is traveled by the corresponding mode. Omitting all individuals that do not have two activities plus one connecting trip reduces the sample size to 42 459 synthetic agents (70.7% of all interviewees). Therefore, considering the population of the whole metropolitan area of the sample (6.65 million), the MATSim synthetic population represents a 0.65% sample.

[7] See https://osm.wno-edv-service.de/boundaries/ for shape files of the Santiago comunas. A possible improvement would be to use land use data for allocating activity locations.

The following activity types are considered:

1. home
2. at work
3. work-related
4. education
5. health-related
6. visit someone
7. shopping
8. leisure
9. other

Typical durations $t_{typ,q}$ (see Eq. 2) of the activities are defined according to the reported durations. This assumes that the reported duration of that activity is the desired one, i.e. the personal trade-off of assigning time to different activities is at the individual optimum. That is, for each of the above activity types, there is a sub-activity with the reported duration rounded up to half and full hours. Because of the wrap-around of activities in MATSim for utility calculations (Nagel et al., 2016), agents are differentiated between those who report the first and the last activity of a day to be the same (83%) and those who start the day with a different activity than they end it with (17%). For the latter group, activities that start after midnight are omitted since they would result in negative scores as a consequence of the current definition of the MATSim utility calculation (see Nagel et al., 2016, for details). As a last step, reported activity end times need to be randomized since one can observe over-reporting of activity end times at full or half hours, yielding to implausible behavior of agents who e.g. leave their activities all at 8:00 a.m. The randomized end times follow a normal distribution around the reported end time T with a standard deviation σ of 20 min ($N(T, 20)$).

3.2 Road network from OSM

The source data for the MATSim Santiago road network is taken from OpenStreetMap (OSM).[8] Two bounding boxes are defined, one of the Santiago Metropolitan Area including the city of Melipilla and the area of Colina, and one of the Great Santiago Area. For the former, streets down to OSM category 4 (secondary) are considered. For the latter, OSM category 5 (tertiary) is additionally considered. This is necessary in order to avoid broken network connections because of a change in category along an important route. Additionally, some adjustments are made to the network, i.e. removing some unnecessary links or adding some missing links. The lane tag in OSM often exhibits the number of lanes of the cross section (i.e. both directions). Hence, the lane tag was omitted and defaults by street category are taken. Finally, some adjustments to the free speed of the secondary network are made. This is often necessary since the maximum speed (e.g. 30 km/h) is not the typical speed on such street. The reason behind it is that, in reality, there are pedestrians crossing or bicycles slowing down the car traffic in residential streets. On bigger streets, traffic lights, which are not explicitly modeled in the MATSim Santiago scenario, increase the expected travel time. In order to account for this, the free speed $v_{free,OSM}$ from OSM is modified as follows:

- $v_{free,new} = 0.50 \cdot v_{free,OSM}$ for all streets up to 40 km/h and all streets up to 60 km/h with only one lane,
- $v_{free,new} = 0.75 \cdot v_{free,OSM}$ for all other streets up to 60 km/h with two lanes, and
- $v_{free,new} = 1.00 \cdot v_{free,OSM}$ for all other streets.

3.3 Public transport supply from GTFS

The source data for the Transantiago public transport (PT) routes and departure times/service frequencies at the stops over time-of-day is a General Transit Feed Specification (GTFS) file[9], published and continuously updated by Santiago's Metropolitan Public Transport Authority (Directorio de Transporte Pblico Metropolitano, DTPM). The GTFS file includes all bus and Metro services.

In order to convert the GTFS files into MATSim transit schedule, network and vehicles, existing infrastructure is used. The main challenge hereby was to adjust the converter such that it considers

[8] See www.openstreetmap.org, OpenStreetMap and contributors.
[9] See http://datos.gob.cl/dataset/1587, accessed 13 August 2015.

frequencies of departures *and* scheduled departures simultaneously. For example, in the case of the Santiago GTFS data, metro departures are given on a frequency basis for the whole day and as (additional) scheduled departures for the peak hours. The existing converter, however, ignored frequencies as long as scheduled departures are given for a certain line. From the MATSim transit schedule, a pseudo transit network is created along with the transit vehicles. This transit network connects – for each transit line – the stops directly to each other. It is not connected to the car network, and only follows the car network's geometry where the resolution of transit stops is high (i.e. where a transit line has a stop at every corner). To give an example, express buses with only two stops exhibit one long link that might start in the city center and end at the boundary of the city. In consequence, cars and buses run in separate networks; as a result it is currently not possible to analyze, for example, cross-congestion effects between modes. Nonetheless, current congestion patterns of PT are exogenously included, since bus travel times are set to be larger in peak periods, calibrated using historical data from buses that are equipped with GPS devices.

4 Setting up the open scenario

By converting the input data into MATSim format, several files are generated to run the simulation. Since there are no data restrictions, these files are provided as an open scenario.[10] The code for obtaining this data from the input data is also publicly available.[11] If you use the above data or the code for generating it, please make sure that you cite the present paper as indicated on the front page.

4.1 Simulation approach

In the following, information about the simulation approach for version 1 of the open scenario is presented.

4.1.1 Simulation parameters

As explained in Sec. 2, the co-evolutionary algorithm of MATSim compares the options that agents have executed in the simulation environment with respect to a utility function. This function is described by behavioral parameters and attributes of the alternatives.

[10] See https://svn.vsp.tu-berlin.de/repos/public-svn/matsim/scenarios/countries/cl/santiago/.

[11] Currently, see https://github.com/matsim-org/matsim/tree/master/playgrounds/santiago/src/main/java/playground/santiago.

Table 2: Behavioral parameters.

Parameter	Value	Unit
Source: Munizaga et al. (2008)		
Marginal utility of activity duration (β_{dur})	+ 4.014	$utils/h$
Marginal utility of traveling (β_{trav})	– 1.056	$utils/h$
Marginal utility of money (β_m)	+ 0.0023	$utils/CLP$
Approximate average $VTTS$	+ 2204.35	CLP/h
Results from calibration		
ASC car	+ 1.100	$utils$
ASC PT	– 0.883	$utils$
ASC walk	+ 0.000	$utils$

The behavioral parameters are taken from a study by Munizaga et al. (2008) and are depicted in Tab. 2. In their simultaneous model, Munizaga et al. (2008, p.216) find a marginal utility of monetary costs of –0.0023 $utils/CLP$[12]. The absolute value of this parameter is used in MATSim as the marginal utility of money. The authors find a total marginal utility of time of –0.0845 $utils/min$ = –5.07 $utils/h$. According to Nagel et al. (2016), this is the sum of the marginal utility of activity duration (β_{dur}) and the marginal utility of traveling (β_{trav}). This differentiation is reflected in Munizaga et al. (2008) by a value of leisure time of +2.75 USD/h and a value of assigning time to travel of –0.74 USD/h, resulting in a total Value of Travel Time Savings (VTTS) of +3.49 USD/h. Using the implicit exchange rate from above and the marginal utility of money of + 0.0023 $utils/CLP$, the calculation finally leads to the behavioral parameters of β_{dur} = +4.014 $utils/h$ and β_{trav} = –1.056 $utils/h$.

The attributes of the three different modes considered in the present study are travel time (for car, PT, walk) and monetary costs (for car, PT). Travel time for car trips is a direct output of the simulation where vehicles interact on the road network. Hence, the car travel time also includes road congestion. Travel times for PT results from the GTFS data (station-to-station travel times including transfer time) plus access and egress times done by the walk mode. Hence, the PT travel times do only partly include road congestion, i.e. as long as it is approximated correctly by the schedule, which uses longer travel times in peak periods. Travel times for walk are approximated by teleporting agents between their activities q and $q+1$ with a travel time of $t_{trav,q} = \frac{1.3 \cdot d_{trav,q}}{4.0\ km/h}$, where $d_{trav,q}$ is the beeline distance between the two activities.

Monetary costs for the car mode are set to 248 CLP/km (taken from Basso et al. (2011) and updated to 2012), reflecting the variable costs of car usage. Fares for the PT mode are set according

[12]Chilean Peso: 1 USD = 631.61 CLP (Munizaga et al., 2008).

to the integrated Transantiago fare system: it differentiates between off-peak fare (640 *CLP* before 6:30 a.m. and after 8:45 p.m.), peak fare (720 *CLP* from 7:00 to 9:00 a.m. and from 6:00 to 8:00 p.m.), and normal fare (660 *CLP* for the rest of the day). At the time of writing, student and senior fare schemes are not yet implemented in the scenario. Additionally, the modeled fare system does not account for the fact that, in reality, there is no extra peak-hour charge if passengers only use buses for their trip.

Travel times for all other transport modes are approximated by congested car travel times (for colectivo, other, ride, taxi) or by teleportation similar to the walk mode (bike, train) with different teleportation speeds (10.0 and 50.0 km/h, respectively). Monetary costs are also approximated. However, as long as switching from/to these modes is not allowed (see next paragraph), this essentially has no effect on simulation results.

The Alternative specific constants (ASCs) of the different modes (see Tab. 2) are determined in the calibration process which will be described in Sec. 4.2.

4.1.2 Simulation procedure

When simulating large-scale scenarios with MATSim, it is recommended to constraint the number of agents allowed to change plans to avoid large oscillating effects from one iteration to the next. First we run 100 iterations. For 80 iterations, 15% of the agents perform route choice, 15% explore a new transport mode for a subtour in their daily plan, and 70% change between the plans that already exist in their choice set. When performing mode choice, in the present version of the model, agents are only allowed to switch between the transport modes car, PT and walk. Trips performed by any other mode (bike, colectivo, other, ride, taxi, train) remain fixed but can be included in the choice set in future versions. PT captive users are taken into account since agents are only allowed to use a car if they have access to a car according to the survey data. Otherwise their only options are PT and walk. For the final 20 iterations, the choice set innovation is switched off and all agents only change between plans that exist in their choice set (see, e.g., Nagel and Flötteröd, 2012, for more information on choice set generation and choice in MATSim). These warm-up runs can be then used to compare a base-case scenario with policy cases (e.g., including road pricing) for another set of 100 or more iterations. In Sec. 4.2 results of the so-called base case are shown, in which the model is run for another 100 iterations, again with 80 iterations of innovation for the agents to find new options, and 20 iterations for the system to relax.

Since this first version of the scenario does not yet use expansion factors to scale the population to a bigger sample size and therefore uses approximately an 0.65% sample, the flow capacity of all links in the car network is multiplied by a factor of 0.0064. In principle, this also would need to be done to the storage capacity of the links. However, to dampen oscillating effects, the storage capacity of all links in the car network is multiplied by a factor of 0.019. This downscaling is not performed on the PT network since it might yield to undesired congestion effects when simulating the total PT supply from GTFS.

4.2 Validation/Calibration

In this section, first visualizations and the results of first validation/calibration efforts are presented.

4.2.1 Visualization

In Fig. 5, a visualization of the MATSim simulation is depicted. It shows the activities of agents in the whole simulated area at midnight, and the movement of cars and public transit vehicles at 8 a.m. Red triangles indicate cars in traffic jam, whereas green triangles show cars in free flow. Because of the small sample size, the congestion patterns do not fully match the real ones; therefore an expansion of the population is recommended for future studies.

Fig. 6 shows the spatial distribution of boarding and alighting in the public transport system. This is a result of the simulation and could in future studies be used to validate the model against real-world smart card data from the Transantiago system.

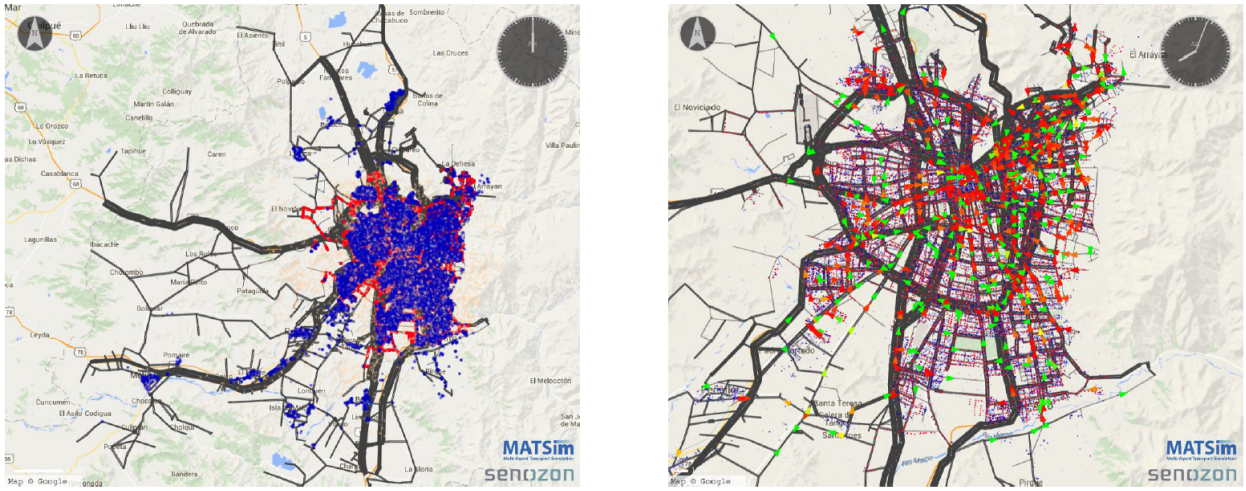
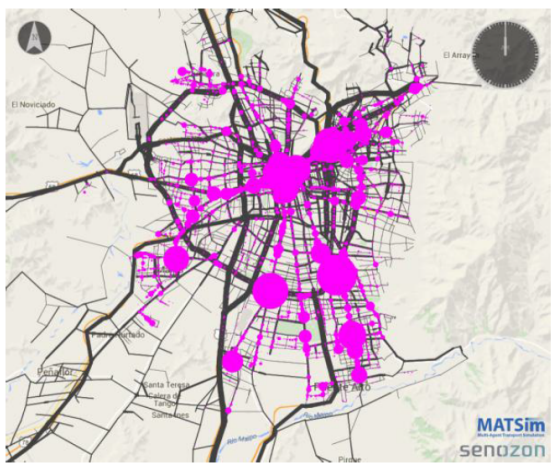
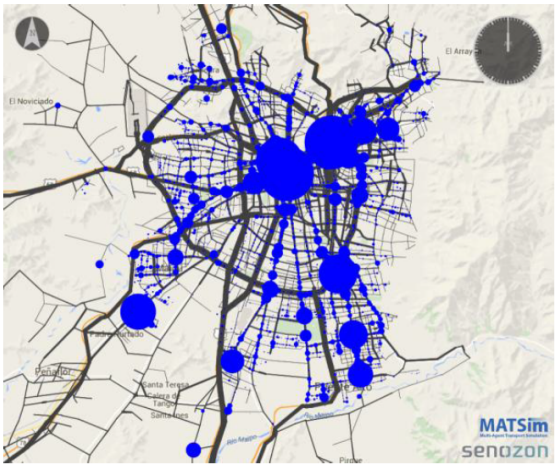


Figure 5: Visualization of the simulation: whole simulated area (left) and Great Santiago (right).



(a) Boarding, Great Santiago

(b) Boarding, central districts



(c) Alighting, Great Santiago

(d) Alighting, central districts

Figure 6: Visualization of simulated boarding and alighting volumes at public transport stations.

4.2.2 Modal split

Tab. 3 shows the modal split of Santiago. The second column depicts the modal split according to Sectra (2014) after expanding the survey to the whole city. The third column shows the modal split in the raw data of the survey. When comparing these two columns, one notices that bike, car, taxi and walk trips are underrepresented in the raw data. In contrast, colectivo, other and PT trips are overrepresented.

As explained in Sec. 3.1.2, some individuals had to be omitted while converting the ODS 2012 raw data into MATSim input. A comparison between column three (raw data) and four (MATSim it.0) of Tab. 3 exhibits that this data cleaning did not introduce systematic errors into the modal split over all trips. When comparing the figures, please note that "Car" in the raw data

Table 3: Modal split: comparison between input data and MATSim synthetic population.

Mode	Sectra (2014)	Raw data	MATSim it.0	MATSim it.200
Bike	4.00	3.41	3.41	3.41
Car	25.70	23.27	14.40	14.28
Colectivo	2.90	3.11	3.73	3.73
Other	6.20	7.74	7.98	7.98
PT	25.00	31.50	29.88	28.19
Ride	in "Car"	in "Car"	8.26	8.26
Taxi	1.70	1.46	1.47	1.47
Train	in "Other"	in "Other"	0.03	0.03
Walk	34.50	29.78	30.83	32.64

(23.27%) includes "Car as a driver" (14.40%) and "Car as a passenger" (= Ride) trips (8.26%) in the MATSim synthetic population. Hence, total car trips in the MATSim population are slightly underrepresented, with 22.66% of total trips. The same is true for PT trips. Colectivo, other and walk trips are slightly overrepresented in the simulation. The share of bike and taxi trips is almost equal.

Column five of Tab. 3 represents the resulting modal split once agents are in the simulation allowed to freely chose between car, PT and walk. As discussed in Sec. 4.1, agents base this decision on a utility function. Since the behavioral parameters are given, and travel times and monetary costs are provided by the simulation (including interaction with other agents), the ASCs of the three transport modes under consideration C_{mode} had to be calibrated to match the initial modal split of the synthetic population (MATSim it.0). This was done by adjusting the constants of every trip q iteratively according to

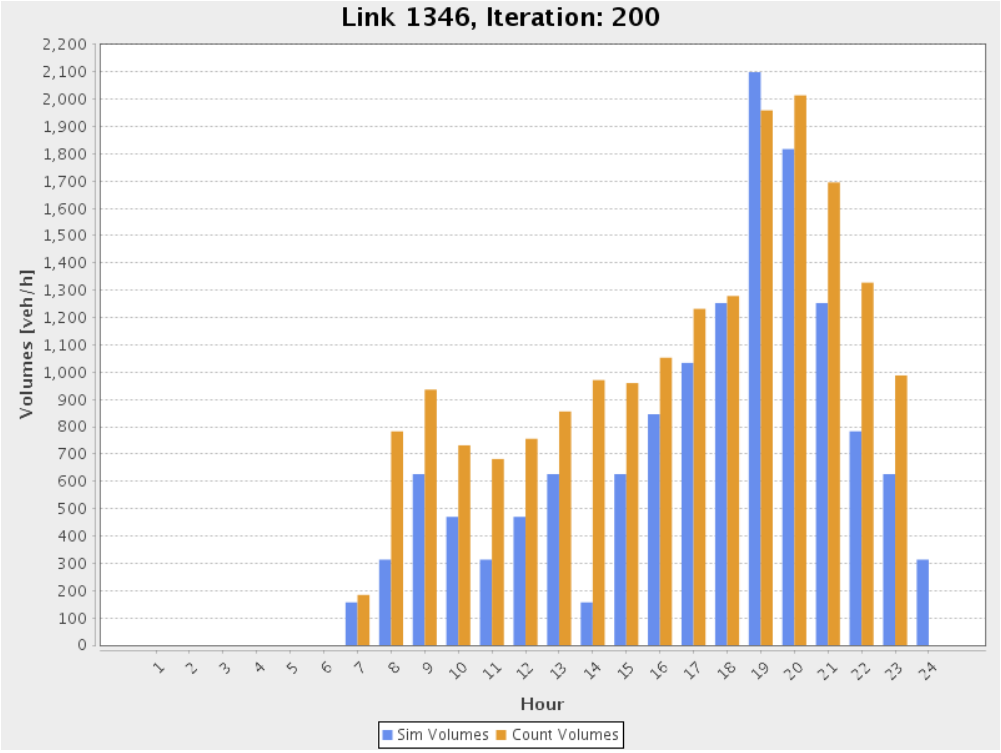
$$C_{mode,n+1} = C_{mode,n} - log\left(\frac{p_{mode,n}}{p_{mode,it.0}}\right) \ , \tag{4}$$

where n is the iteration step of this calibration, $p_{mode,n}$ is the modal share of the corresponding transport mode at the end of MATSim simulation n, and $p_{mode,it.0}$ is the modal share of the transport mode according to the corresponding entry in Tab. 3. As a final result of this calibration procedure, the modal split of iteration 200 is very similar to the one of iteration 0. Only some PT trips are still replaced by walk trips. Additionally, a modal split distribution over different trip lengths should be investigated in the future. However, the model output yields a rather stable modal split and is from this point of view suitable for investigating the impact of different policies.

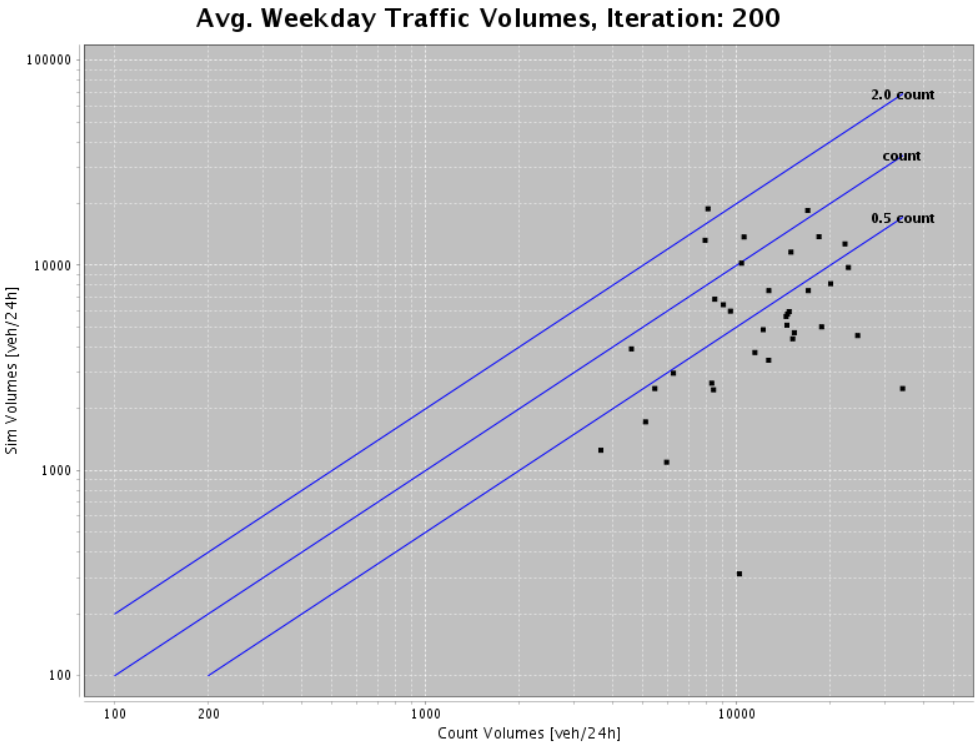
4.2.3 Counting stations

Another standard verification of MATSim simulation output is the comparison of traffic flows to data from real-world counting stations. 49 counting stations are available within the Santiago greater area, 40 on major roads, 9 on (parallel) local roads. The counts data is recorded in July 2011. After cleaning the data, 36 counting stations remain with data from 6:00 a.m. to 11:30 p.m. in 15 min. time bins. Fig. 7(a) shows the comparison between simulated and observed flows for one counting station over time of day.[13] It can be observed that the simulation predicts the overall shape of the load profile pretty well. However, it seems that in most hours, there are not enough vehicles passing the counting station in the simulation (higher yellow bars than blue bars). This effect can be observed as a general issue from Fig. 7(b). It represents simulated over real-world counts over the whole day. Every data point stands for one counting station. If all data points were on the 45 degree line in the center of the figure, the simulation would perfectly reproduce reality. However, most data points are below the 45 degree line. This indicates that – in the simulation – there are systematically not enough vehicles on the roads. Two possible reasons come to mind: First, since overall modal split fits the raw data pretty well, it might be that short car trips are overrepresented and long car trips are underrepresented, yielding to too little kilometers traveled. Second, the simulation is based on the survey population only (approx. 0.65%). This means that many counting stations remain completely untouched for many hours of the day since one simulated vehicle stands for approx. 170 other vehicles, but still can only drive on one road. This additionally is likely to have impacts on the traffic flow model and the prediction of travel times. Also tolled urban highways are not implemented in the simulation yet, which could introduce a systematic error to the measurement as many counting stations are on one of these tolled highways. Hence, in order to obtain more realistic travel times and flows on the roads, the most important next step in this project is the synthesis of a 10% or 100% population from raw data, and to include the tolled highways in Santiago.

[13] Please note, that this comparison is performed for vehicle category "C01" only, which represents passenger cars without taxis and colectivos. The latter two are currently not part of the traffic flow simulation; hence, the counts comparison is consistent.



(a) One station over time of day



(b) All stations for the whole day

Figure 7: Comparison to counting stations.

5 Conclusion and outlook

This paper showed how a MATSim scenario can be set up in a very sophisticated way if the input data is open and publicly available. The resulting scenario provides a platform for researchers, but also for the public and private sector. Possible applications include the (economic) evaluation of planned transport policies and projects and the development of business ideas based on the simulated mobility of individuals in the city. This indicates the importance of open data as a prerequisite for transparent decision making of modern administrations as well as for stimulating innovation activity of the private sector.

When tackling one of the many interesting research questions that come to mind with this scenario, some time should be invested to improve it in the near future. The idea here is that everyone who wants to use the scenario is asked to do some improvement during her/his work. The improvements should then be provided for other researchers by uploading them as a new version to https://svn.vsp.tu-berlin.de/repos/public-svn/matsim/scenarios/countries/cl/santiago/. Please get in touch with the corresponding authors. A non-exhaustive list of possible improvements is the following:

- Synthesize a 10% or 100% population
- Use land use data (instead of random points) for unknown activity locations in the population synthesis
- Add toll to tollways (the tollways are included in the network, tolls are missing but the data is available)
- Add colectivos (a shape file is available)
- Add freight traffic from survey data
- Further calibration of traffic counts, modal shares, and travel times
- Add bicycles as network mode
- Add capacity constraints for PT vehicles and add PT vehicles to road network

A non-exhaustive list of potential research problems to be analyzed with the MATSim Santiago scenario follows:

- The effects of road pricing strategies on travel times, traffic volumes, public transport and demand for non-motorized mobility, air pollution, noise levels and so on.
- The introduction of alternative interventions such as (full or partial) pedestrianization of the city center, zones 30, roads with exclusive right-of-way for public transport, plate-number based car driving restrictions, parking restrictions, road closures and road openings, restrictions on truck traffic, new cycleways and new Metro lines.
- The extraction of accessibility measures to study the land use impacts of transport interventions.

Acknowledgements

This research project highly benefited from the public data policy of the Chilean Government, which led to the publication of the full origin-destination survey and of the public transport supply data. The project has been supported by Chile's National Commission for Scientific and Technological Research (CONICYT) within the FONDECYT project "Social effects and quality of service valuation of public transport services" (Grant 11130227), that funded the stay of B. Kickhöfer at Universidad de Chile in Santiago. The authors would like to thank K. Nagel (Technische Universität Berlin) for supporting and partially funding this research. A. Tirachini also acknowledges support from the Complex Engineering Systems Institute, Chile (Grants ICM P-05-004-F, CONICYT FBO16). Finally, the authors want to thank H. Schwandt and N. Paschedag at the Department of Mathematics (Technische Universität Berlin) for maintaining our computing clusters.

References

H. Barahona, F. Gallego, and J.-P. Montero. Adopting a cleaner technology: The effect of driving restrictions on fleet turnover. Working paper, available at (accessed 25/jan/2016) http://www.um.edu.uy/docs/adopting-a-cleaner-technology-the-effect-of-driving-restrictions-on-fleet-turnover.pdf, 2015.

L. J. Basso, C. A. Guevara, A. Gschwender, and M. Fuster. Congestion pricing, transit subsidies

and dedicated bus lanes: Efficient and practical solutions to congestion. *Transport Policy*, 18(5): 676–684, 2011. ISSN 0967-070X. doi:10.1016/j.tranpol.2011.01.002.

D. Charypar and K. Nagel. Generating complete all-day activity plans with genetic algorithms. *Transportation*, 32(4):369–397, 2005. ISSN 0049-4488. doi:10.1007/s11116-004-8287-y.

R. Contreras. EOD Santiago 2012: procesos de expansión y corrección en ausencia de datos censales. *XVII Congreso Chileno de Ingeniera de Transporte, Concepción, 13-15 October*, 2015.

A. Horni, K. W. Axhausen, and K. Nagel, editors. *The Multi-Agent Transport Simulation MATSim*. Ubiquity, London, 2016a. doi:10.5334/baw. URL http://ci.matsim.org:8080/job/MATSim-Book/ws/matsimbook-latest.pdf.

A. Horni, K. Nagel, and K. W. Axhausen. Introducing MATSim. In Horni et al. (2016a), chapter 1. doi:10.5334/baw. URL http://ci.matsim.org:8080/job/MATSim-Book/ws/matsimbook-latest.pdf.

M. Munizaga, S. Jara-Díaz, P. Greeven, and C. Bhat. Econometric calibration of the joint time assignment–mode choice model. *Transportation Science*, 42(2):208–219, May 2008. doi:10.1287/trsc.1080.0231.

J. C. Muñoz, M. Batarce, and D. Hidalgo. Transantiago, five years after its launch. *Research in Transportation Economics*, 48:184–193, 2014.

V. Muñoz, A. Thomas, C. Navarrete, and R. Contreras. Encuesta Origen Destino de Santiago 2012: Resultados y validaciones. *Revista Ingeniería de Transporte*, 19(1):21–36, 2015.

K. Nagel and G. Flötteröd. Agent-based traffic assignment: Going from trips to behavioural travelers. In R. Pendyala and C. Bhat, editors, *Travel Behaviour Research in an Evolving World – Selected papers from the 12th international conference on travel behaviour research*, chapter 12, pages 261–294. International Association for Travel Behaviour Research, 2012. ISBN 978-1-105-47378-4.

K. Nagel, B. Kickhöfer, A. Horni, and D. Charypar. A closer look at scoring. In Horni et al. (2016a), chapter 3. doi:10.5334/baw. URL http://ci.matsim.org:8080/job/MATSim-Book/ws/matsimbook-latest.pdf.

F. G. Sabatini, G. Caceres, and J. Cerda. Segregación residencial en las principales ciudades chilenas: Tendencias de las tres últimas décadas y posibles cursos de acción. *EURE (Santiago)*, 27(82):21–42, 2001.

Sectra. Actualización y recolección de información del sistema de transporte urbano, Etapa IX, Encuesta Origen Destino Santiago 2012. *Informe Final, Observatorio Social Universidad Alberto Hurtado*, 2014.

A. Tirachini. Probability distribution of walking trips and effects of restricting free pedestrian movement on walking distance. *Transport Policy*, 37:101 – 110, 2015. ISSN 0967-070X. doi:http://dx.doi.org/10.1016/j.tranpol.2014.10.008. URL http://www.sciencedirect.com/science/article/pii/S0967070X14002078.