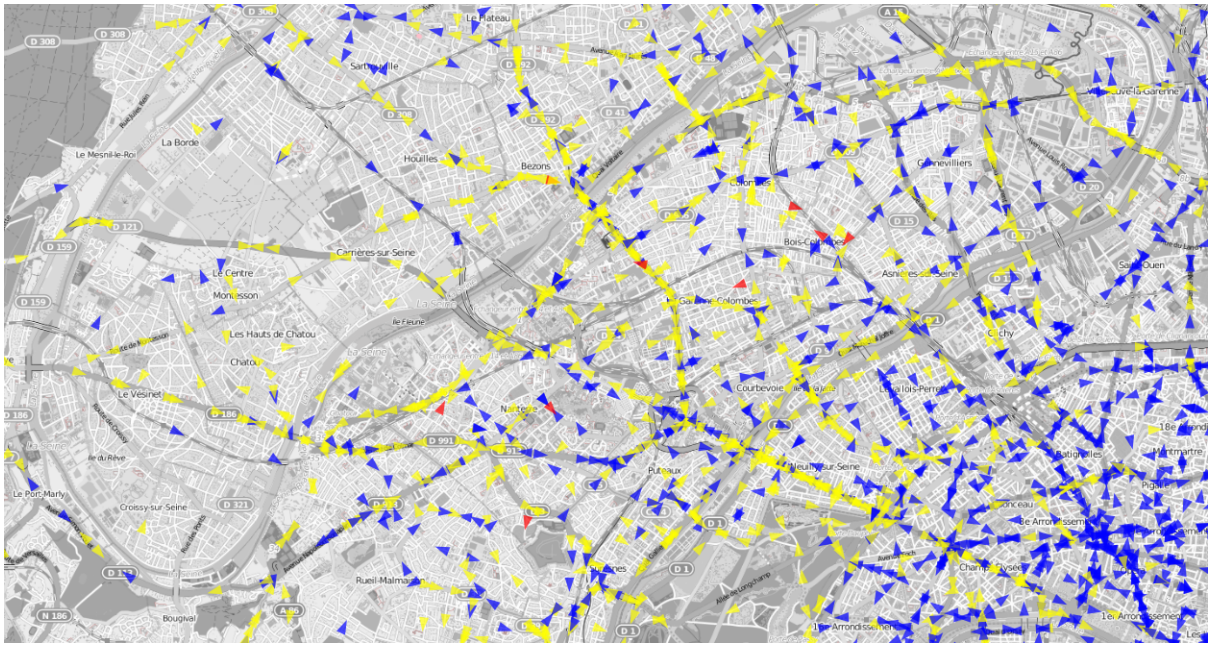


A MATSim scenario for Autonomous Vehicles in La Défense and Île-de-France

Sebastian Hörl

Working paper **10XX**

Institute for Transport Planning and Systems **February 2017**

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Contents

1 Introduction **1**

2 Baseline Scenario: Île-de-France **1**

2.1 Facility Sampling 2

2.2 Population Sampling 7

2.2.1 Agent Population 8

2.2.2 Skeleton Activity Chain 11

2.2.3 Secondary Activity Chain 13

2.2.4 Location Choice 14

2.2.5 Mode Choice 18

2.3 Network Conversion 18

2.4 Public Transport Conversion 19

3 Derived Scenario: La Défense **20**

3.1 Research scenario and sub-sampling 20

3.2 MATSim Integration 22

3.3 First Simulaton Results 22

4 Discussion and Outlook **23**

5 References **25**

Working paper Institute for Transport Planning and Systems 10XX

A MATSim scenario for Autonomous Vehicles in La Défense and Île-de-France

Sebastian Hörl
IVT, ETH Zürich
Stefano-Franscini-Platz 5
CH-8093 Zürich
phone: +41-44-633 38 01

sebastian.hoerl@ivt.baug.ethz.ch

February 2017

Abstract

The synthesis of a scenario for the agent-based traffic simulation framework MATSim for the Île-de-France region is documented. A reduction of the scenario to an area around La Défense in Paris is proposed and next steps towards a study of autonomous vehicle fleets in that area are shown.

Keywords

autonomous vehicles, agent-based, simulation, scenario, MATSim, La Défense, Île-de-France

Preferred citation style

Hörl, S. (2017) A MATSim scenario for Autonomous Vehicles in La Défense and Île-de-France, *Working paper Institute for Transport Planning and Systems*, **10XX**, Institute for Transport Planning and Systems (IVT), ETH Zurich, Zurich.

1 Introduction

Autonomous vehicles [AVs] have become an important topic in the discussion of transport experts all over the world. In order to assess the expected impacts and to prepare the transport system for the presumably disruptive new technology, more and more literature about AVs is in the making.

Traffic systems are always highly individual structures, which depend on the regional demand, geography, economoy and legislation. Furthermore, the overall traffic situation is composed by the actions of large numbers of travelers with individual travel plans and objectives. Therefore, the use of agent-based modelling offers strong advantages over classic macroscopic traffic simulations.

The agent-based traffic simulation framework MATSim (Horni et al., 2015) has recently been extended with functionality for the simulation of autonomous vehicles (Hörl, 2017). Based on this extension case studies on the introduction of AVs to an area around La Défense in Paris are planned.

In this report, focus is put on the generation of a simple simulation scenario in that setting. With heavy reliance on assumptions, an artificial population of agents for Île-de-France is created, the road and public transport system in the region is modeled and finally an example simulation in the scenario is presented.

The process in this report is meant to be a basic blueprint for the creation of a much more detailed and evidence-based simulation scenario. By showing which steps have been taken for this simple version of the scenario it should become evident what is necessary for a more complex setup.

2 Baseline Scenario: Île-de-France

In order to realistically simulate not only the travelers within the La Défense area, but also the traffic going into and out of the area, a baseline scenario for the surrounding areas needs to be developed. For the purpose of this paper, it has been decided that a scenario for the entire of Île-de-France should be sufficient. Still, one needs to keep in mind that long-distance travellers are not covered by this approach. Commuters, e.g. from Lyon arriving by TGV in Paris and continuing their journey to La Défense are not covered here. Nevertheless, embedding La Défense into a bigger scenario of Île-de-France should give enough level of detail while possible discrepancies may still be corrected afterwards.

For the generation of the Île-de-France scenario in this paper open data has been used, which is publicly available on the web. Most definitely, the scenario may be improved significantly with higher detailed data sets. In this sense, the paper at hand can give a practical introduction to the generation of simulation scenarios with MATSim, but does not aim for a highly realistic result.

A couple of steps are necessary to arrive at a full MATSim scenario. The generation of the Île-de-France scenario is structured as follows: First, facilities in the area are sampled so it is known where activities of the agents may take place. Second the population is sampled with the main components of generating their homes and demographics, generating their activity chains and assigning locations to their activities. Finally, the road and public transit networks are generated.

The area of interest for this scenario area the 8 departements as show in Figure 1:

- Paris (75)
- Seine-et-Marne (77)
- Yvelines (78)
- Essonne (91)
- Hauts-de-Seine (92)
- Seine-Saint-Denis (93)
- Val-de-Marne (94)
- Val-d'Oise (95)

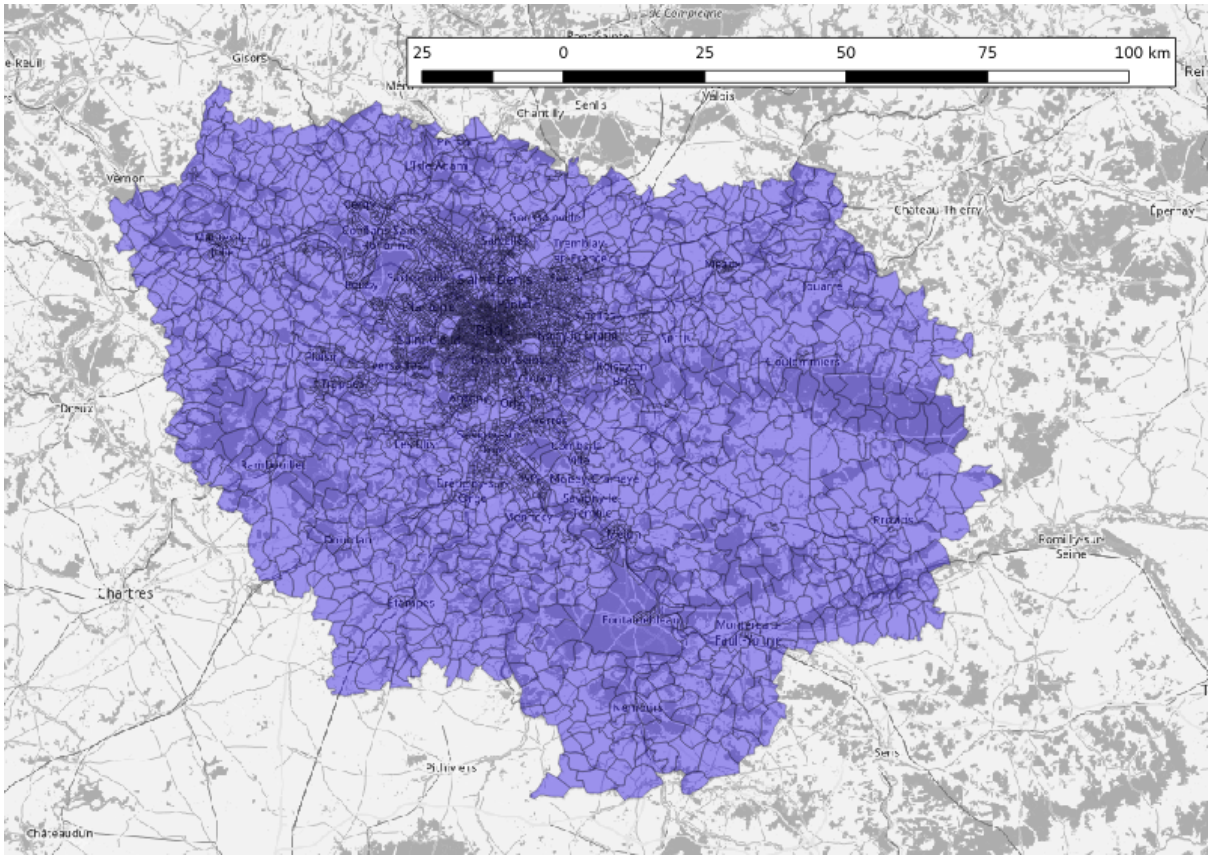
2.1 Facility Sampling

The first step in the scenario generation is the generation of facilities. They are needed to define where the actions of the agents are taking place, i.e. they represent opportunities of performing an action.

Since only a basic scenario is generated, only a small variety of different facilities is used. As is common for basic scenarios, four basic types of facilities will be generated:

- **Work** facilities, where agents may perform work activities. Practically, this can be any company, institution, etc. in the real world
- **Education** facilities, where agents perform education activities. In the real world, this would resemble to schools and universities.
- **Secondary** facilities, where secondary activites are performed. This can be locations for recreational activities (gyms, cinemas, ...), shopping activities (malls, stores, ...) and many

Figure 1: The area of Île-de-France, divided in IRIS segments.



Source: OpenStreetMap (Background)

others.

- **Home** facilities. Those, in contrary to the the latter ones, will directly be created during the population generation, but not in this first step.

The following parts of the chapter will first introduce the available data sets and then describe how the facilities with the given structure have been generated from the data.

The data sets that have been used for the facility generation are:

- Contours IRIS, Septembre 2016 (Institut National de l'Information Geographique et Forrestiere, 2016)
- Dénombrement des équipements de services, santé, enseignement et tourisme en 2015 (Institut national de la statistique et des études économiques, 2016b)

The first provides the shape of all IRIS areas in France (in Lambert-93 projection, Figure 1).

Table 1: Mapping of the INSEE data to simple facility types

INSEE Dataset	Work Facility	Education Faciltiy	Secondary Facility
Sport, loisirs et culture	•		•
Commerce	•		•
Services aux particuliers	•		•
Action sociale	•		
Services du santé	•		
Fonctions médicales et paramédicales	•		
Tourisme	•		
Enseignement du 1er degré	•	•	
Enseignement du second degré	•	•	
Enseignement supérieur, formation et services de l'éducation	•	•	

The second provides, for all IRIS with more than 10k inhabitants, the number of companies and institutions. The data set is divided in categories (e.g. health, education, ...) and within these categories into specific types of institutions. For the educational data set one would for instance have the number of pre-schools or universities per IRIS.

Table 1 shows the available categories in the data set and the type of facility they have been assigned to. One can see that each entry in the database may represent a work facility, but that only a subset is suitable as an educational or secondary activity facility.

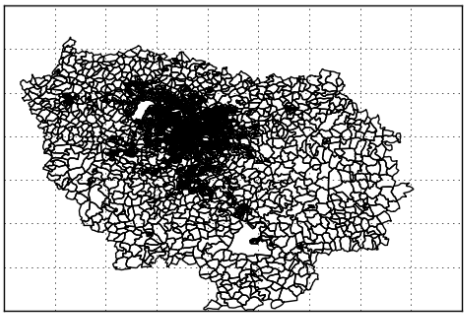
In the first step of the sampling, all relevant IRIS have been filtered from the shape file data. This way, a polygon has been obtained for each IRIS within the 8 departements of Île-de-France (Figure 2(a)). In a second step, axis-aligned bounding boxes have been obtained, which are needed for the sampling later on (Figure 2(b)).

Algorithm 1 has been defined to sample a facility location within a specific IRIS (each IRIS has a unique ID). It first samples coordinates in the bounding box of an IRIS and then returns them if the point is actually covered by the polygon. Otherwise, more coordinates are sampled.

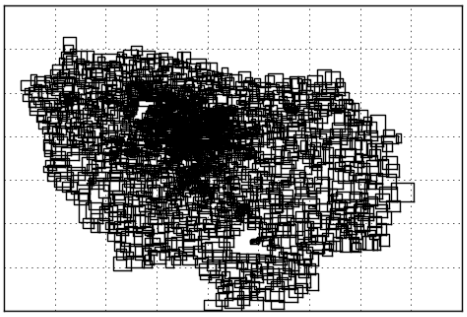
For the actual sampling of the facilities, two cases need to be considered: For IRIS with more than 10k inhabitants, where data is available, a number of facilities according to the counts obtained from the assignment in Table 1 is generated (Algorithm 2). For IRIS with a lower number of inhaibtants a very basic model has been obtained.

Figure 2: Spatial data from IGN

(a) IRIS from IGN shape files



(b) Bounding boxes from IGN shape files



Algorithm 1 Sampling a location within a specific IRIS

Require: IRIS Polygon Area $\mathcal{P}$

Require: $x_{min}, x_{max}, y_{min}, y_{max} \leftarrow$ IRIS Bounding Box

loop

Sample $x \sim \mathcal{U}(x_{min}, x_{max})$

Sample $y \sim \mathcal{U}(y_{min}, y_{max})$

if $(x, y) \in \mathcal{P}$ **then**

return (x, y)

end if

end loop

This model consists of a categorical model on the probability of a facility being of a specific type, as well as a model on the number of facilities in an IRIS. The first one has been obtained from the overall dataset, i.e. it has been measured how many facilities of each type are available (according to Table 1). Their relative frequencies are then used in a categorical model. The result can be seen in Table 2. For the latter model, the total number of facilities for each IRIS over 10k has been obtained and a Gamma distribution has been fitted to give a statistic on the number of facilities per IRIS (Figure 3).

Then, for all the IRIS with less than 10k inhabitants, first a number of facilities n is sampled from the Gamma model, after which n facility types are obtained from the categorical model. A summary of the algorithm can be seen in Algorithm 3.

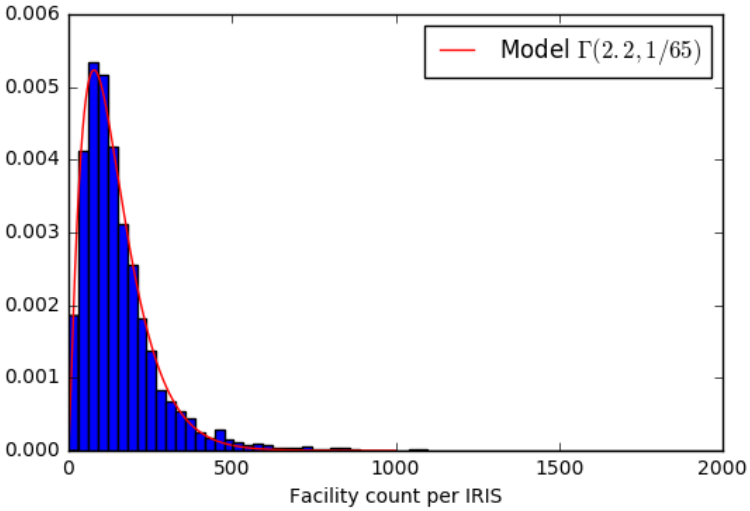
Using these two algorithms, in total 817,419 facilities have been generated.

Obviously, the strongest simplification that has been done is the reduction of all possible facility

Table 2: Mapping of the INSEE data to simple facility types

Facility Type	Absolute Frequency	Relative Frequency
Work	379,268	56.56%
Education	9,231	1.38%
Secondary	282,057	42.06%

Figure 3: Facility count model. $\Gamma(2.2, 1/65)$



Algorithm 2 Sampling facilities for IRIS with more than 10k inhabitants

Require: List of facility types $T = \{\text{work}, \text{education}, \text{secondary}\}$

Require: List of IRIS in Île-de-France $I^+ = \{...\}$ with more than 10k inhabitants

Require: Counts $n(i, t)$ of facilies of type t in IRIS i

for all $i \in I^+$ **do**

for all $t \in T$ **do**

Sample $(x, y) \sim$ Algorithm 1

Save Facility at location (x, y) in IRIS i with type t

end for

end for

Algorithm 3 Sampling facilities for IRIS with less than 10k inhabitants

Require: Facility Count Model $\Gamma[N]$ (Figure 3)
Require: Facility Type Model Cat$[T]$ (Table 2)
Require: List of IRIS in Île-de-France $I^- = \{...\}$ with less than 10k inhabitants
  for all $i \in I^-$ **do**
    Sample $n \sim \Gamma[N]$ (Number of facilities)
    for $k = 1...n$ **do**
      Sample $(x, y) \sim$ Algorithm 1
      Sample $t \sim$ Cat$[T]$
      Save Facility at location (x, y) in IRIS i with type t
    end for
  end for

types to three very broad categories of facilities. This may be reverted to allow for a more rich generation of the population, but is sufficient for the current use case.

More detailed data sets could further enrich the model. First, in terms of spatial resolution: While the IRIS level is rather detailed, no specific coordinates for facilities can be obtained from the data. To create a model that is as accurate as possible, such data would be beneficial.

One point, that has been completely ommitted in this step, are facility capacities. In the real world we observe that there are capacities for specific facilities, i.e. only a certain number of people fits into a bakery or into an office building. This limitation is not covered in this scenario.

Furthermore, facilities are interpreted as "opportunities to perform an activity". However, a more realistic way would be to interpret them as buildings or physical entities where activities can be performed. The difference is that in one building different kinds of activities may be performed. While MATSim allows the definition of different types for one and the same facility with individual capacities, this functionality is not used in this scenario.

2.2 Population Sampling

The sampling of a simple population for the scenario at hand is done from ground up. First, agents are created in the IRIS grid of Île-de-France, second activity chains for their daily plans are created and finally a location choice is performed for the activities.

In general, the agents resemble the facility types described before. Three different kinds of

agents will be generated: work, education and secondary.

Of what type an agent is, will be determined from the demographic structure of the home IRIS, while the type, in turn, largely impacts the shape of the activity chain. The activity chain of a "work" agent will contain elements where he spends time at work, while "education" agents will spend time at "education" facilities.

The characteristics of the activities will - due to the lack of availabe data for Île-de-France during the creation of this scenario - be based on a MATSim model of Switzerland. The choices for work and educational activities will be based on simple assmuptions.

Again, it should be outlined that all the above factors lead to a scenario which is exemplifying the generation of a scenario. It is, to certain extent, useful to conduct studies, but in no way aimed at resembling reality closely.

2.2.1 Agent Population

The available data sets for the sampling of agents are threefold:

- Activité des résidents en 2013 (Recensement de la population) (Institut national de la statistique et des études économiques, 2016a)
- Diplômes - Formation en 2013 (Recensement de la population) (Institut national de la statistique et des études économiques, 2016c)
- Logement en 2013 (Recensemenet de la population) (Institut national de la statistique et des études économiques, 2016d)

These data sets provide counts on different population characteristics on a per-IRIS basis. Included here are all IRIS, not only those over 10k inhabitants.

First, a model on the age structure can be built, because all people younger than 15 years old will generate a "education" agent. The first dataset provides a count on people older than 15, while the second one does that for people which are younger. Therefore, for each IRIS a probability of observing a "< 15y" agent can be obtained:

$$P(\text{Agent younger than } 15|i) = \frac{\#\text{ of agents} < 15 \text{ years in IRIS } i}{\text{of all agents in IRIS } i} \tag{1}$$

For all people over 15, the first dataset provides information on whether they are employed, unemployed, students or retired. Since here unemployed people and retired are combined to the agent type "secondary", these counts can be used to create a categorical model on the agent type per IRIS, which is based on the relative frequencies of the agent type: Cat$[T|i]$.

Similarly, information is available for the modes that people usually use to travel. Available in the data set are "by foot", "by car", "by public transport" and "by bike". These translate directly to the MATSim modes "walk", "car", "pt" and "bike". Again, based on their relative frequencies in each IRIS, a IRIS-based categorical distribution for these modes can be constructed: Cat$[M|i]$. Actually, two of these models are obtained, one where "car" is not an option, denoted as Cat$[M|i|c = 0]$.

This is necessary, because car ownership is considered in the simulation. Data on this is found in data set three, where one can find the number of households per IRIS, as well as the number of households with one or more cars. This way, a car ownership probability for the agents in one IRIS can be defined: $P(\text{Car Ownership}|i)$.

Using these statistics, the algorithm iterates over all IRIS and first determines the overall number of agents $n(i)$ for an IRIS. It is the number of people under and over 15 in the dataset. Then, it is determined if the agent is under 15. If so, he automatically becomes a "education" agent. Otherwise, his type is sampled from Cat$[T|i]$. Then, it is sampled whether the agent owns a car (please note that this is not only done for agents over 15 to compensate for parents escorting their children). Depending on whether a car is available, either the non-car mode distribution or the other one is used to sample a main mode for the agent. Finally, a home location is obtained from the selected IRIS. The algorithm is summarized in Algorithm 4. At the end, each agent has the following attributes: Home Location, Agent Type, Car Ownership, Main Mode.

Using this procedure, 8,733,385 agents have been generated with the number of facilities increasing to 9,550,804. Please note that some sources stat a total number of around 15M inhabitants for Île-de-France. Where this discrepancy is coming from needs to be checked in the future.

A critical look on the procedure reveals some problems: The main ommission here is the household structure, which MATSim is generally able to handle. Probably the INSEE data sets can be used to further develop a model of households and then assign agents to those households. This way it would be possible to capture phenomena such as family members escorting children to school or sharing a car. However, the involved amount of work has been skipped for this basic scenario. It should also be noted that no socio-demographic data has been attached to the agents. Often in the analysis of a scenario this becomes interesting, because it is interesting to

see, out of their travel situation, how different user groups change their travel behaviour once new elements are added to the simulation. For instance, younger agents might be specially attracted by a fleet of autonomous vehicles, just because it inreases their number of travel options next to walking, bikes and public transport, while older user groups with high car availability might not be impacted as much.

It should be outlined, that the step of creating the agents and their attributes is one of the most important steps in the whole procedure of scenario generation. Usually, one would have a small population sample available (from a survey and similar sources) and use it as a kernel for the further development and up-sampling to a whole population. This, along with more detailed models, makes it possible to capture interdependencies between the statistical dimensions. In the scenario here, the counts of agents are correct, on an upper level and independently, but e.g. the joint probability of observing an agent of age older than 15 with the main mode "walk" is not controlled for here! Therefore, a more detailed model of the population (with respective data sources) is strongly advised for a more realistic scenario.

Algorithm 4 Agent Sampling Algorithm

Require: List of IRIS in Île-de-France $I = \{...\}$
  for all $i \in I$ **do**
    for $k = 1...n(i)$ **do**
      Agent Type $t \leftarrow$ education
      Car Ownership $c \leftarrow 0$
      $r \sim \mathcal{U}(0, 1)$
      if $r \leq P(\text{Agent older than } 15|i)$ **then**
        $t \sim \text{Cat}[T|i]$ (based on relative frequencies of agent types in IRIS i)
        $r \sim \mathcal{U}(0, 1)$
        $c \leftarrow r \leq P(\text{Car Ownership}|i)$
      end if
      if c **then**
        Main Mode $m \sim \text{Cat}[M|i|c = 1]$
      else
        Main Mode $m \sim \text{Cat}[M|i|c = 0]$
      end if
      $(x, y) \sim$ Algorithm 1
      Save Home Facility (x, y, home)
      Save Agent (x, y, t, c, m)
    end for
  end for

2.2.2 Skeleton Activity Chain

The next step in the population is the generation of activitiy chain skeletons. While in reality the decision processes for defining which activities to perform during one day are highly complex, here a simple model is proposed. The resulting activity chains should be "believable", i.e. be realistic enough that they "could" be real. Nevertheless, they are not founded on empirical data, but strong assumptions, the main one being that each agents daily plan consists of some "main activities" which are generated first, while "secondary activities" are filled into this skeleton afterwards.

Depending on the agent type, different skeletons are proposed. From a MATSim scenario of Switzerland, statistics on the start time of activities have been measured, as well as their durations. Subsequently, models have been fitted manually to resemble the respective distributions. The result can be seen in Figure 4. There, the respective PDFs are plotted, scaled such that their maximum value in the shown interval is one (for better readability). The distributions are summarized below:

- **Work**
  - Start time: $t_{start} \sim 0.6 \cdot \mathcal{N}(7.5, 1.2) + 0.4 \cdot \mathcal{N}(13.5, 1.4)$
  - Duration: $t_{duration} \sim 0.3 \cdot \Gamma(1, 2) + 0.5 \cdot \mathcal{N}(4.5, 1) + 0.2 \cdot \mathcal{N}(0.5, 1.2)$
- **Education**
  - Start time: $t_{start} \sim 0.7 \cdot \mathcal{N}(8, 0.4) + 0.3 \cdot \mathcal{N}(13.5, 0.4)$
  - Duration: $t_{duration} \sim 0.35 \cdot \mathcal{N}(2, 0.6) + 0.5 \cdot \mathcal{N}(4, 0.5) + 0.15 \cdot \mathcal{N}(8, 1.2)$
- **Secondary**
  - Start time: $t_{start} \sim 0.5 \cdot \mathcal{N}(14.5, 4) + 0.5 \cdot \mathcal{U}(6, 20)$
  - Duration: $t_{duration} \sim 0.3 \cdot \Gamma(0.8, 1.4)$

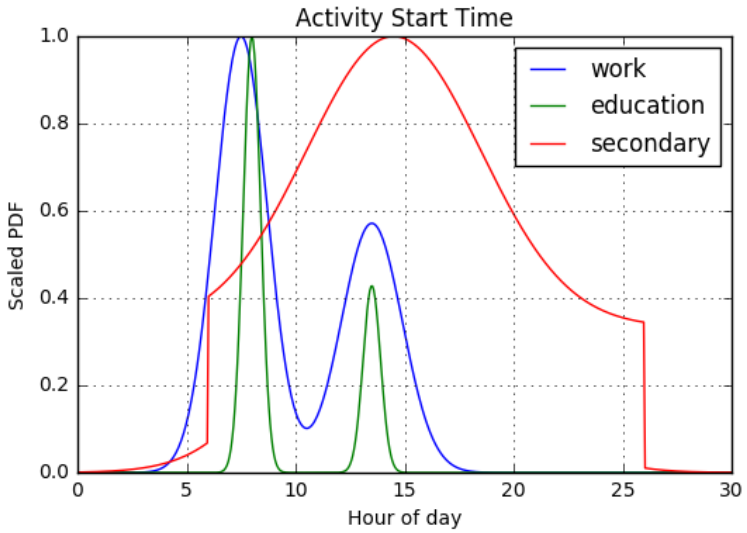
One can see that while for "work" and "education" activities, there are two modes of start times, one in the morning and one in the afternoon, an assumption has been made for "secondary" activities, which are distributed over the day with a preference around 15:00. The durations show that for education either durations of 2h, 4h or 8h are common with different probabilities while for work 4h or around 9h are common. This resembles that some people stay at work during the whole day, while others explicitly insert additional activities for going out to lunch or shopping in between two parts of their daily work.

The skeleton activities for "work" agents are constructed as follows:

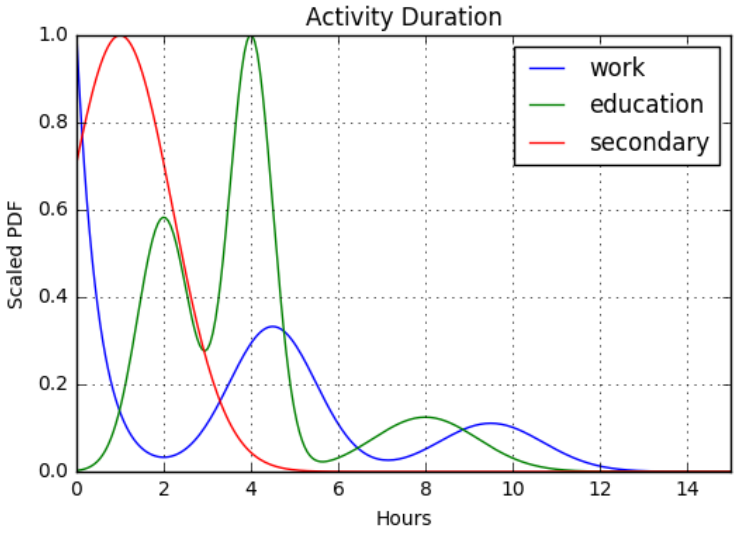
- Sample if agent has a split or combined skeleton (depending on the weights of the two modes in the start time distribution)

Figure 4: Activity Model

(a) Activity Start Time Models



(b) Activity Duration Models



- If it is a split schedule, sample a start time for the morning and a start time for the afternoon. Then insert activities into the schedule of a duration that comes from the 4h duration distribution.
- If it is a combined schedule, sample a start time for the morning and a duration from the 8h distribution. Then insert the activity.

The skeleton activities for "education" agents are constructed as follows:

- Sample if agent has a split or combined skeleton (depending on the weights of the two

modes in the start time distribution)

- If it is a combined schedule, sample a start time for the morning and a duration from the 8h distribution. Then insert the activity.
- If it is a split schedule, two activities will be added. They can either be 4h or 2h. Depending on the distribution weights, sample whether each is 2h or 4h, sample a respective duration and a start time, from the morning distribution for the first one, and from the afternoon distribution for the second one. Then add the activities.

The skeleton activities for "secondary" agents are constructed in a very simple way. One single "secondary" activity is sampled in duration and start time and added to the schedule. It will be used as a "seed" in subsequent steps.

It becomes apparent, that this appraoch adds a lot of assumptions to the population. Here, it would be beneficial if at least the start time distribution and duration distribution of activities in Île-de-France would be available as reference data. Furthermore, it would even be better if this information is given jointly. Here, those two dimensions are sampled independently, but in reality there is a strong correlation. For instance, if one observes a work duration of 8h, it is more likely that the start time of this activity is quite early. While efforts are made to cover such implications in a formal way in this approach, it would be even better if this information could be sampled from empirical data for France.

2.2.3 Secondary Activity Chain

As a second step for the generation of activity chains, secondary activities are added to the schedules. The algorithm searches each schedule for gaps of more than 1h and then samples secondary activities. If they fit into the gaps, they are inserted into the schedule. This process is done iteratively per schedule until no new secondary activity could be inserted. The insertion process also makes sure that a minimum of 30min buffer is found between activities, such that the agent has time to travel between them. Additionally, the insertion is only done with a specific probability.

The algorithm parameters have been tuned to reach at schedule structures, which are not overly cluttered with secondary activities, but still provide a certain amount of individualism. Table 3 shows the statistics of the generated activity chains. One can see that in general "education" agents have more secondary activities on average than working ones, while for both the standard deviation of the number of secondary activities is around one and a half.

Finally, home activities are added to the start and end of the schedule. This assumes that any

Table 3: Statistics on the number of secondary activities

Agent Type	Avgerage number of sec. act.	SD of number of sec. act.
Work	1.01	1.26
Education	1.24	1.36
Secondary	2.46	1.45

agent in the population starts the day at his home locations and also ends it there.

2.2.4 Location Choice

While it is now known, which activities each agent performs during the day, there is no information yet on where these activities take place. Only the home locations and facilities of the agents are known up to this point.

Again, a simple model is proposed, which is based on artificially built distance distributions. They are shown in Figure 5. One can see that walking distances are quite short, while a certain share of people is using "long-distance" travel, i.e. car or public transit, to cover longer distances. Furthermore, a bike distance distribution is available.

The location choice process is also divided in two stages: First, a "main location" is sampled, which determines where an agent's "work" or "education" activities take place, second locations for the secondary activities in between are sampled.

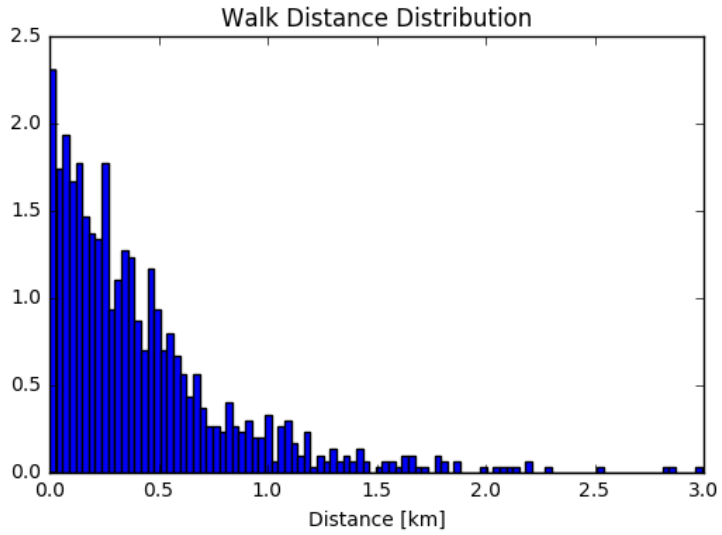
The first algorithm works as follows:

- For an agent the home location is known
- Depending on his "main mode", sample a distance
- From all IRIS, find the one where the distance of the centroid to the agent's home location gets closest to the sampled distance
- Sample a facility (of the right type) within this IRIS and define it as the "main location" for this agent

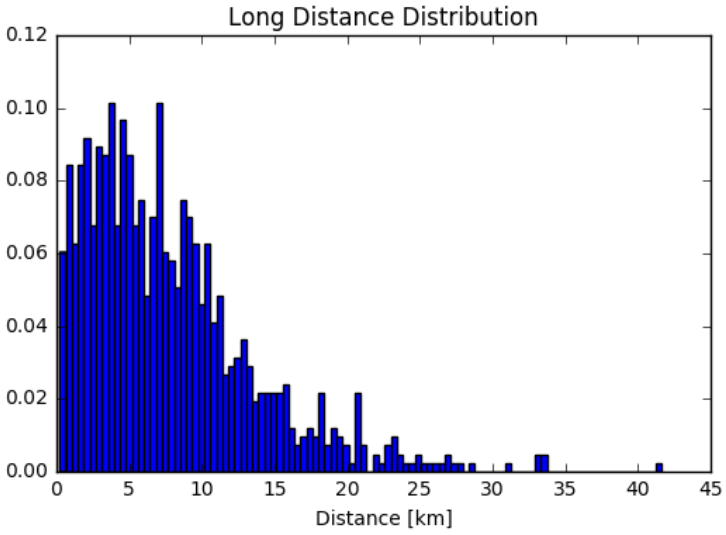
Therefore, the location of work or education is mainly determined by the main mode of the agent. An example can be seen in Figure 6. There, possible main locations have been sampled for walking distances and for long distances.

Figure 5: Distance Model

(a) Walk Distance Model



(b) Long Distance Model



These locations, in turn, highly influence the location of secondary activities, because they need to be consistent (i.e. they cannot be on the opposite side of Île-de-France). Again, a simple model is proposed: Between to main locations (i.e. home to work, or work to work, or work to home, ...) an imaginary line is drawn. Along this line a Gamma distribution is defined with the center of mass close to one of the locations, thinning out towards the other one. The same is done symmetrically on the other side. To add variation around this imaginary line, locations are furthermore scattered in normal direction. The resulting distribution of locations between two such main locations is exemplified in Figure 7. This way, newly sampled locations for secondary activities are likely to be close to either of the main locations, or in between, but they may never

Figure 6: Main Location Sampling. Crosses are IRIS centroids. Blue candidates have been obtained for the walk mode, while the magenta ones stem from the long distance distribution.

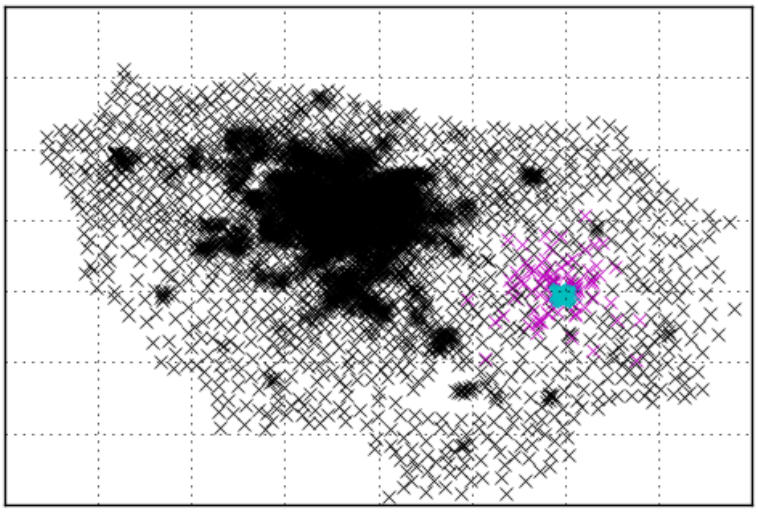
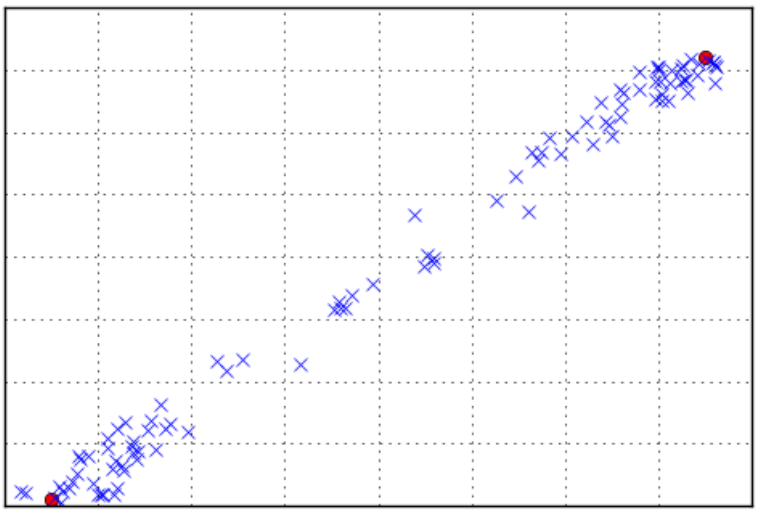
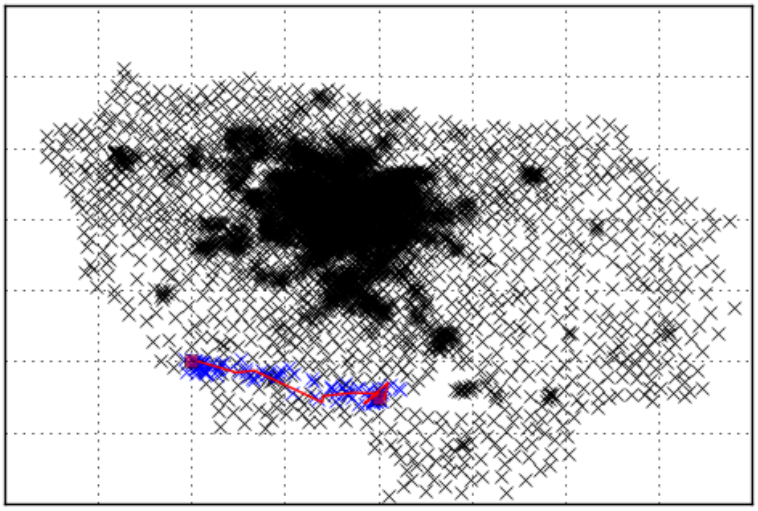


Figure 7: Sampling approach for secondary facility locations using two Gamma distributions and a normal variation.



be too far off.

Figure 8: Example of a constructed secondary activity chain.



The sampling is done the following way:

- Determine all the main activities in a schedule
- Determine the number of secondary activities in each segment of the schedule, which is framed by two main activities
- For each segment of size n, sample n secondary activity facilities between the two framing main activities according to the previously explained scheme
- Order these n locations by distance between the two framing activities and assign them to the secondary activities

This way a consistent chain of locations is put between each pair of main activities. A constructed example can be seen in Figure 8. There, the blue crosses are a large number of locations sampled between the two main locations, while the red line is then a chain of 10 of these locations that connects them.

A lot can be said about the flaws of this location choice approach. In fact, it is solely based on assumption and not on data. However, it is one of the most crucial points of the model. Data, either as single origin-destination observations or in form of aggregated OD-matrices, could be used here to resemble the actual travel patterns in Île-de-France. This would not only include realistic locations, but also result in realistic distances.

2.2.5 Mode Choice

Finally, to conclude the generation of the population, legs are sampled, which connect two activities. The process is rather simple: Given the locations of one activity following the other, a mode for this leg is chosen, based on the distance distributions.

First, the distance between two activities is determined. Then, for each mode, the pdf of the mode distance distribution is evaluated at that point. Given the values for all the modes, a categorical selection is performed based on these quantities (Equation 2). This way, the mode for a connecting leg is sampled.

$$P(M|d) = \frac{P(d|M)}{\sum_M P(d|M)} \tag{2}$$

If the main mode of an agent is "car", all legs have been fixed to be "car" legs, because this way it can be ensured that the car is not abandoned at some stage during the daily plan. More sophisticated models would be needed to allow for sub-tours which are done by walking or public transport.

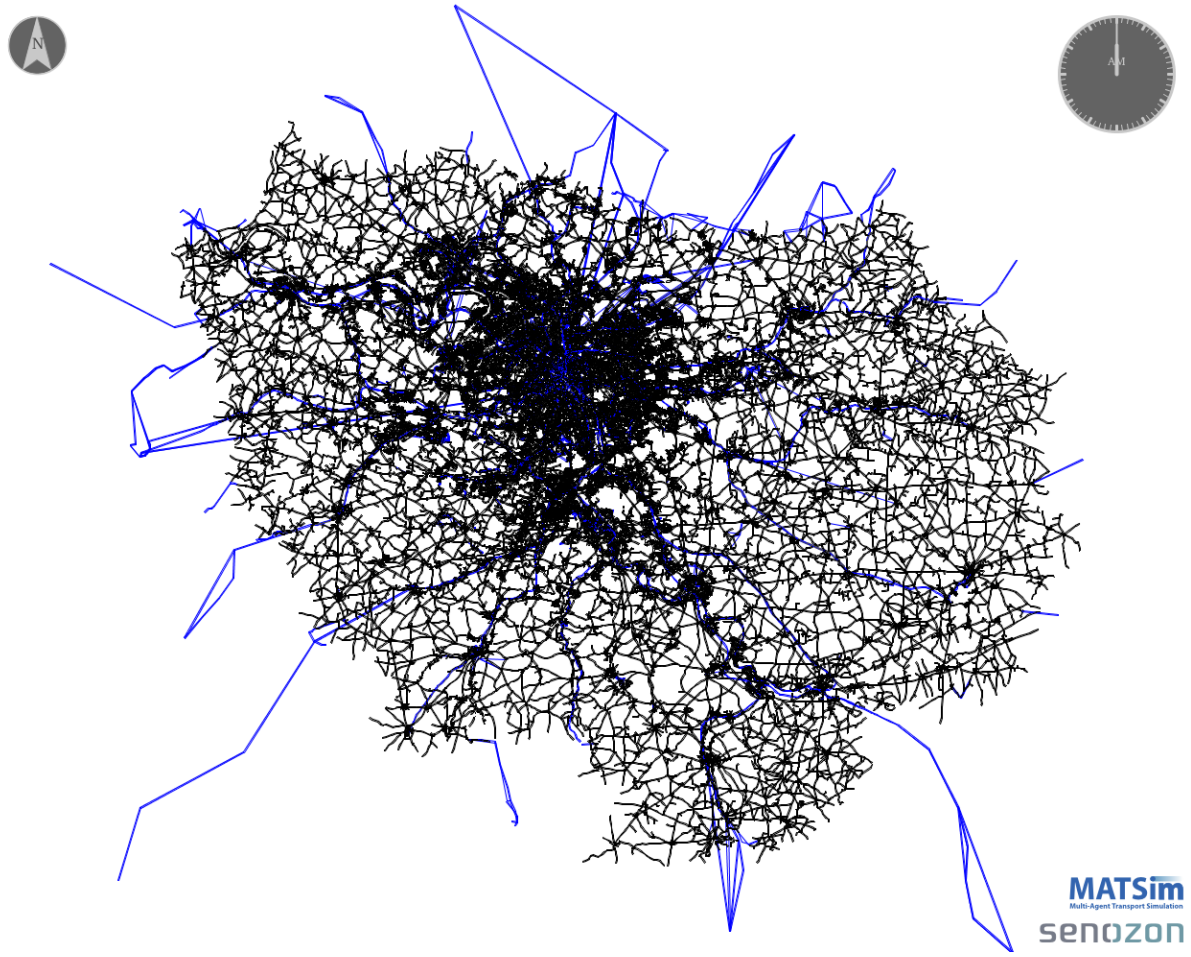
Again, more detailed data on the mode choice of inhabitants of Île-de-France would be beneficial here.

2.3 Network Conversion

The source of the network for Île-de-France is OpenStreetMap. A pre-packaged excerpt from OSM for the region is available from Geofabrik GmbH (2016). For the conversion to MATSim, the standard MATSIM OSM Converter has been used. Part of this conversion has been the simplification of the network. For instance, large curves in the real road network are represented by a large number of links in OSM. For MATSim, such level of detail is not necessary and thus, such curves are replaced by single links. This is done completely automatically by the software.

The resulting road (and public transport) network can be seen in Figure 9.

Figure 9: Road and public transit network of Île-de-France in MATSim.



2.4 Public Transport Conversion

While the network for the public transport has been obtained from OSM, no further information on the number of lines, vehicles and their timetables is known yet. This information can be obtained from STIF (2017) in GTFS format. On a regular basis, they provide public transport schedules for Île-de-France for the following three weeks.

The tool "pt2matsim" (Poletti, 2016) has been used to first convert the GTFS data from STIF to MATSim schedules. For that, the 10 January 2017 has been chosen as a normal Tuesday (without any special holiday service) to create the schedule. In a second step, the tool has been used to map the transit schedules to the network. This is necessary, because GTFS only provides information on times and stops, also where those stops are, but not, how they relate to the actual road (or rail) network. This step is done automatically by "pt2matsim".

The resulting public transport service consists of 1785 lines with 45,329 distinct routes. In total there are 50,881 stop facilities.

3 Derived Scenario: La Défense

From the upper-level Île-de-France scenario, one can "zoom-in" to the area around La Défense or cut it out, depending on which approach is taken.

The first section will discuss several ways of scaling down the scenario to the La Défense area while maintaining a specific level of detail. Afterwards, one use case with the introduction of autonomous vehicles will be presented, along with first simulation results.

3.1 Research scenario and sub-sampling

In which way the Île-de-France scenario can be scaled down to a simulation of La Défense highly depends on what effects one wants to observe in the simulation.

The first simulation run of the scenario would generate routes for all the agents. Because no information on the travel times in the network is available, they would be assigned according to the fastest route, based on freespeed limitations. However, especially at peak times, this is not how people act in the real world. Because of heavy congestion in some areas, travel times would drop at peak hours and alternative routes would be chosen. MATSim is able to reproduce this behaviour, but in any case the scenario needs to be run for a couple of iteratons. Then, a consistent picture of measured travel times at any time of the day can be drawn.

Once this is achieved, one can think about simplifications to the scenario.

A very simple scenario would be to manually convert all car or public transport trips within La Défense to AVs. This can easily be done by modifying the data set. One could turn off all replanning behaviour (i.e. mode choice, departure time choice, etc.) and just observe if a specific fleet size is able to handle the demand. In such a case, one could find all agents, which have an actual activity within La Défense or cross it at any point during the day. All other agents could be deleted from the scenario, because they would have no influence on the simulation results. The same would be true for a large number of public transport lines.

However, if one wants to explore mode choices, the situation looks different. A study question

here could be: Given that there are AVs available in La Défense, how do people choose from the set of available modes? On one hand, one could do this with the relevant set of agents mentioned above, with the respective public transport infrastructure within La Défense and fix their travel decisions outside of the region. Nonetheless it might also be interesting to see how commuters make decisions. Maybe, because the last mile travel is made significantly easier, people would switch to public transport lines for the main part of their journey in favor of cars. In that case, one would need to re-insert the surrounding public transport system, such that people who have been using the car actually have a choice to switch to transit lines.

Furthermore, if location choice decisions should be observed, the complete surroundings of La Défense need to be included in the simulation. It could happen that people from completely new regions start to find work locations in La Défense interesting because of the introduction of AVs. In order to observe these effects, the whole population of Île-de-France would be needed to be simulated.

The examples above should show that it is possible to significantly reduce the size of the simulation scenario, but that details are lost while doing this. However, it always strongly depends on the research question, because the missing details may not be relevant to the study at hand.

Another important point that needs to be considered when setting up the research scenario is calibration. On one hand, the behavioural parameter of the agents, which lead to their decisions, need to be well-defined for France and Île-de-France, on the other hand there, where it is not possible to find specific values, the scenario needs to be calibrated such that all values are consistent.

Calibration may be done against traffic or public transport counts. By comparing the real-world reference counts with the simulated results one can see how close the simulation gets to reality. By defining behavioural parameters (but also by generating a better suited population) one can get closer and closer. In a way the calibration process therefore also acts as a validation of the scenario that has been created.

One more step that can be done for MATSim is to down-sample the population proportionally. One needs to be careful here, because a 10% sample on a network that is scaled down to 10% of the road capacity may lead to similar results in classic settings, but the results are expected to be different if shared modes are introduced, such as pooled AVs. This is due to the changed demand structure by the sub-sampling.

3.2 MATSim Integration

During the integration of the scenario to MATSim some problems have occured, which need to be solved in the future.

The public transport network of Île-de-France is so complex (especially because of Paris) that it gets very huge for computational purposes. The way that the MATSim PT router works today is that transfer links are created between stop facilities. A radius can be defined that states in which distance stop facilities should be connected by such "walk" links. While for other big scenarios around 700m have worked, this would result in the creation of around 1B transfer links for Île-de-France. This is clearly too much and thus the search radius has been reduced significantly (to 50m). That can be a reasonable choice for the city core of Paris, because not very long walk connection between public transport stops are expected, but it is not ideal for outer regions of the scenario. Therefore, an adaptive way of defining this connectivity radius should be introduced to the simulation software.

For an example run of the scenario, a strategy proposed previously has been applied to the scenario: All agents which either perform at least one activity in their schedule within the area or cross the area of interest during their daily plan are tagged as "relevant" agents. All others are discarded. This way, the scenario is reduced to around 1M agents. Furthermore, the remaining 1M population has been sub-sampled down to 100k agents, leading to a 10% sample.

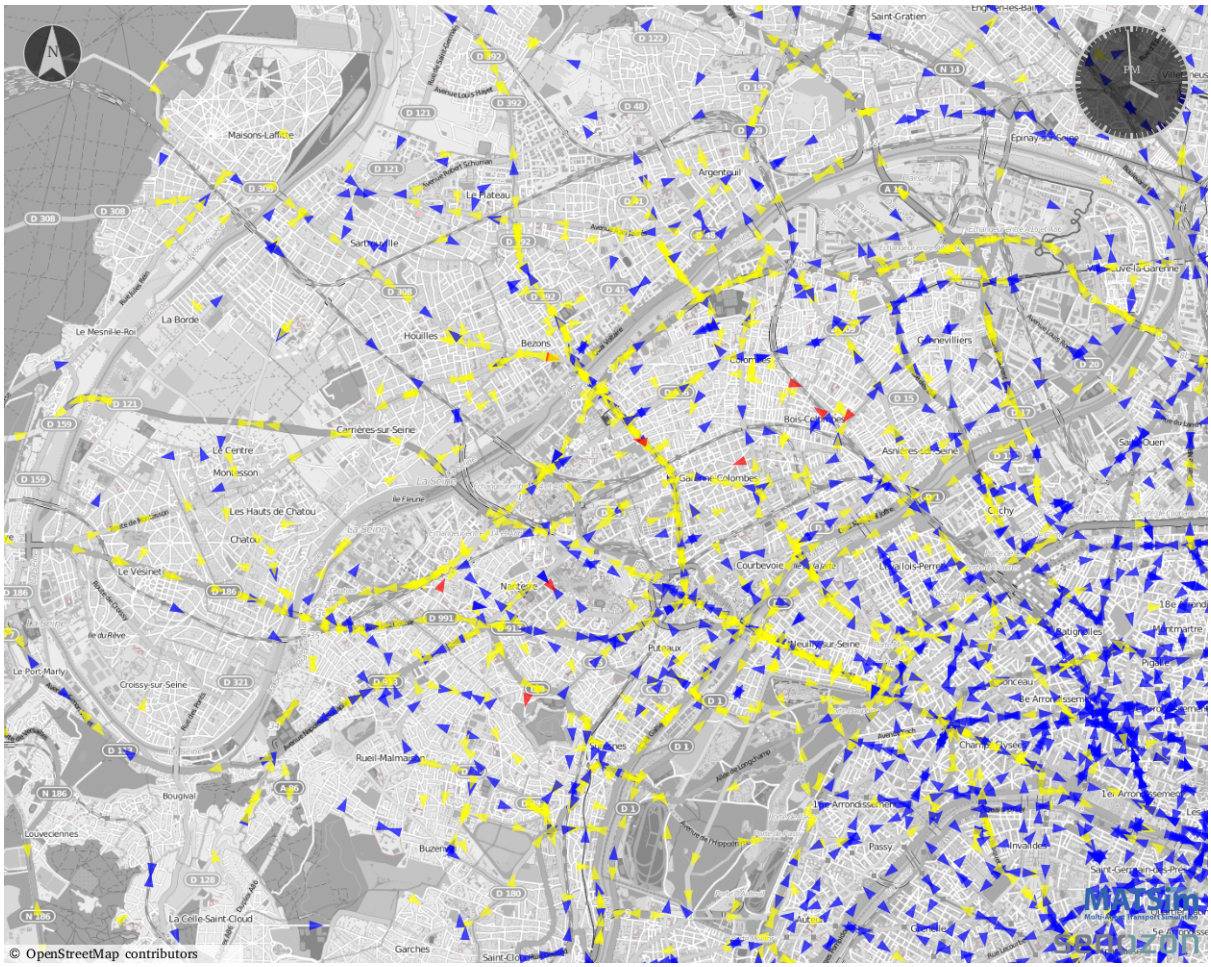
For the presented test case, no initial travel time relaxation has been done, this will be necessary for future applications (but in any case it can be done with a 10% sample of the entire of Île-de-France).

Finally, the autonomous vehicles components of MATSim have been modified to only let AVs be present in the relevant area. The relevant agents have the options of departure time choice and mode choice.

3.3 First Simulaton Results

Figure 10 shows a snapshot of the simulation, visualized by the VIA software. One can see public tranport vehicles in blue, private cars in yellow and AVs in red. It can be seen how they interact in the same network. Figure 11 then gives a glimpse on the analysis of such a service. Iteration by iteration, travel choices are made and one can wait until they lead towards an equilibirum. At this point it can be seen that agents use the AV option, but also that the share of car users has increased significantly from the initial conditions. This may be because of two

Figure 10: Final Result: A simulation of autonomous vehicles in La Défense. Blue: Public Transit Vehicle. Yellow: Private Car. Red: Autonomous Taxi.



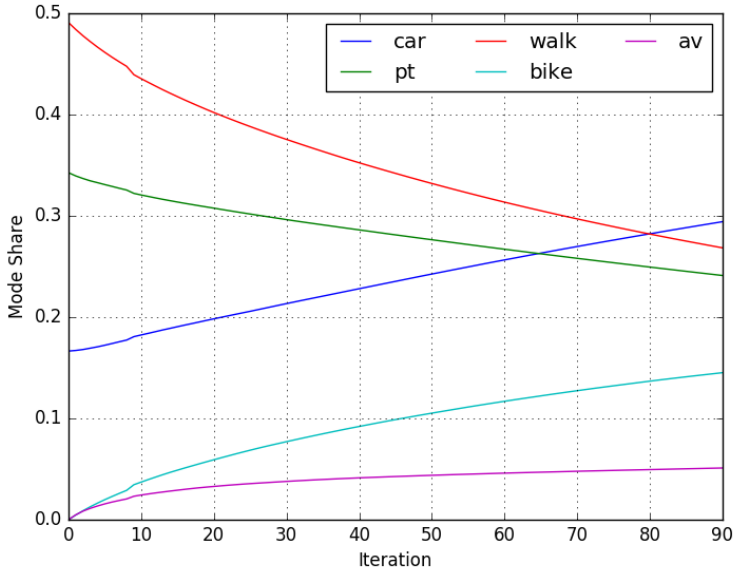
factors: First, no travel times have been inferred for the peak hours, so driving a car does not lose attractivity due to congestion. Furthermore, the scoring parameters, i.e. the behavioural information about the agents, have been adopted from a Switzerland scenario. They would either need to be backed by real data for France and Île-de-France or be calibrated.

4 Discussion and Outlook

The report at hand follows the creation of a scenario for Île-de-France and La Défense in a very basic way. Additional data sources that would be beneficial for a more realistic version are:

- Information on Household structure to capture household related travel behaviour
- Detailed travel survey information to obtain the frequency of trips with specific modes and

Figure 11: Share of different mode over iteratons in a simulation run.



purposes. A person-based source would be beneficial to be used as a sampling kernel for a realistic dependency structure of the statistical dimensions.

- Such a data set ideally would include information on trip purposes and therefore allow the construction of better activity chains. Especially departure times and trip durations in dependency on modes, home and work locations and purposes would be interesting.
- Data on distinct trips and/or aggregated OD matrices would be beneficial for getting the spatial distribution of agents, as well as the distance distributions right
- Information on behavioural parameters (i.e. the value-of-time) for different modes and socio-demographic characteristics is crucial for a well-defined simulation
- Reference data would be needed for the calibraton of the scenario (vehicle counts, passenger counts, ...)

Valuable data sources might (for instance) be:

- *Enquête Nationale de Transports et Déplacements (ENTD) 2007/2008* by INSEE though the data is not very recent and the survey is only done every 10 years
- *Enquête Globale Transport (EGT) 2010* by OMNIL for Île-de-France

In order to arrive at a useful scenario for the assessment of AV fleet algorithms, the following steps would need to be performed:

- Simulation of overall travel times (for the whole of Île-de-France, possibly with a 10%

sample)
- Thinning of the scenario based on the research question and the relevant choice dimensions
- Consistent definition of behavioural parameters
- Calibraton of the scenario against real-world traffic and/or public transport counts
- Definition of behavioural parameters for the new AV mode

Then, different algorithms could be tested and their impact on the artificial population could be assessed.

5 References

Geofabrik GmbH (2016) Ile-de-France, `http://download.geofabrik.de/europe/france/ile-de-france.html`.

Hörl, S. (2017) Agent-based simulation of autonomous taxi services with dynamic demand responses, *Arbeitsberichte Verkehrs- und Raumplanung*, **Upcoming**, IVT, ETH Zurich, Zurich.

Horni, A., K. Nagel and K. W. Axhausen (2015) *The Multi-Agent Transport Simulation MATSim*, Ubiquity, London.

Institut national de la statistique et des études économiques (2016a) Activité des résidents en 2013, `https://www.insee.fr/fr/statistiques/2386631`.

Institut national de la statistique et des études économiques (2016b) Dénombrement des équipements de services, santé, enseignement et tourisme en 2015, `https://www.insee.fr/fr/statistiques/2044564`.

Institut national de la statistique et des études économiques (2016c) Diplômes - Formation en 2013, `https://www.insee.fr/fr/statistiques/2386698`.

Institut national de la statistique et des études économiques (2016d) Logement en 2013, `https://www.insee.fr/fr/statistiques/2386703`.

Institut National de l'Information Geographique et Forrestiere (2016) Contours Iris, `https://www.data.gouv.fr/en/datasets/contours-iris/`.

Poletti, F. (2016) Public transit mapping on multi-modal networks in MATSim, *Master Thesis*, IVT, ETH Zurich, Zurich.

STIF (2017) Horaires prévus sur les lignes de transport en commun d'Ile-de-France (GTFS), `https://opendata.stif.info/explore/dataset/offre-horaires-tc-gtfs-idf/`.