**D6.1 Case study 1: a travel demand model for London**

<table>
<thead>
<tr>
<th>Project acronym</th>
<th>EUNOIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project title</td>
<td>Evolutive User-centric Networks for Intraurban Accessibility</td>
</tr>
<tr>
<td>Grant Agreement number</td>
<td>318367</td>
</tr>
<tr>
<td>Funding Scheme</td>
<td>Collaborative project</td>
</tr>
<tr>
<td>Project start date / Duration</td>
<td>01 Oct 2012 / 26 months</td>
</tr>
<tr>
<td>Call Topic</td>
<td>FP7-ICT-2011-8</td>
</tr>
<tr>
<td>Project web-site</td>
<td><a href="http://www.eunoia-project.eu/">http://www.eunoia-project.eu/</a></td>
</tr>
<tr>
<td>Deliverable</td>
<td>D6.1 Case study 1: a travel demand model for London</td>
</tr>
<tr>
<td>Edition</td>
<td>1</td>
</tr>
<tr>
<td>Date</td>
<td>12/01/2015</td>
</tr>
<tr>
<td>Status</td>
<td>Final</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td>Confidential, only for members of the consortium</td>
</tr>
</tbody>
</table>
D6.1 Case study 1: a travel demand model for London
Edition 1

Authoring and Approval

Prepared by
Joan Serras (CASA – Researcher)
Melanie Bosredon (CASA – Researcher)
Vassilis Zachariadis (CASA – Researcher)
Camilo Vargas-Ruiz (CASA – Researcher)
Thibaut Dubernet (ETH – Researcher)
Mike Batty (CASA – Principal Researcher)

Reviewed by
José Ramasco (Deputy Scientific Coordinator)
Ricardo Herranz (Deputy Management Coordinator)

Approved by
Maxi San Miguel (Project Coordinator)

Record of Revisions

<table>
<thead>
<tr>
<th>Edition</th>
<th>Date</th>
<th>Description/Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edition 1 Draft 1</td>
<td>25/11/2014</td>
<td>Initial version</td>
</tr>
<tr>
<td>Edition 1</td>
<td>12/01/2015</td>
<td>Minor additions and corrections Approved for submission to the EC</td>
</tr>
</tbody>
</table>
Table of Contents

1. INTRODUCTION ........................................................................................................................................... 1
   1.1 ORGANISATION OF THE DOCUMENT .......................................................................................................... 1

2. BUILDING THE CASE STUDY .......................................................................................................................... 3
   2.1 SUPPLY NETWORKS: REPRESENTING THE TRANSPORT INFRASTRUCTURE FOR LONDON ......................... 3
   2.2 AN ACTIVITY-BASED TRAVEL DEMAND FOR LONDON .................................................................................. 6
      2.2.1 The synthetic population ........................................................................................................................... 6
      2.2.2 Generating activities for individuals .......................................................................................................... 9
      2.2.2.1 Assigning skeletal activity patterns ....................................................................................................... 9
      2.2.2.2 Activity location assignment ................................................................................................................. 12
   2.3 RUNNING MATSIM .......................................................................................................................................... 16
      2.3.1 Using Oyster card data to calibrate modal preference ............................................................................. 19
   2.4 CASE STUDY VALIDATION ............................................................................................................................ 20

3. CONCLUSIONS .................................................................................................................................................. 23

ANNEX I. REFERENCES ........................................................................................................................................ 24

ANNEX II. ABBREVIATIONS AND ACRONYMS .............................................................................................. 27

ANNEX III. CAPACITY VALUES FOR ROAD LINKS IN THE LONDON CASE STUDY ...................................... 28

ANNEX IV. THE LAND-USE CONFIGURATION FOR THE LONDON CASE STUDY ........................................... 31

ANNEX V. MATSIM, AN OVERVIEW ................................................................................................................... 34
1. Introduction

The development and usage of the first travel demand-type models in the UK goes back around 50 years, since interest on urban growth, and therefore transport planning began to emerge after World War Two. Since then, a series of key advances have occurred in the field, which have led to a situation where transport-related issues in cities can now be looked at from a varied range of analytical tools (Jones, 2009). One such tool are travel demand models and, in the present time, many cities use them on order to study the impacts involved on different policy scenarios. London is no different and so Transport for London (TfL) owns a range of modelling tools to help estimate impacts around transport-related issues. The current suite of models used by TfL include: the London Transportation Studies (LTS) model, the Highway Assignment Model (HAMs), the Railplan model, the London Regional Demand model (LoRDM). Perhaps the most relevant component in the context of this project is the LTS model which essentially is a strategic multimodal model which aims at forecasting travel demand for London. LTS has a traditional trip-based four-stage approach.

For our case study, and after discussions with TfL, we are interested in exploring the issues of interest to EUNOIA through an activity-based travel demand approach. The main difference between trip-based and activity-based approaches lies on activities being central to the analysis of transport related issues as opposed to the trips per se. This kind of approach started to take shape around 30 years ago (Jones, 1983) and it has slowly developed to a stage in which many cities in the US (Donnelly, Erhardt, Moeckel, & Davidson, 2010) and in the Netherlands have operational models in place. For some reason, not much development around activity-based models seem to have occurred in the UK (Jones, 2009). One particular effort in this direction is the one by the Centre for Transport Studies at Imperial College by (Sivakumar, Le Vine, & Polak, 2010).

In order to test out the issues of interest within EUNOIA, we had to build a travel demand model of London as comprehensive as possible. We took the opportunity provided by the project to build the model around the following concepts: activity-based and dynamic traffic assignment. The model, when built, will provide a platform upon which to test many of the new developments around alternative data sources and potential new scenarios.

It is through this framework that we will explore the integration of non-conventional data sources for our case study. In this sense, we provide an application for the active use of the Oyster card sample made available by TfL to CASA, in the model design of our case study. We will also explore the usage of data collected on the London bike sharing scheme.

Ultimately, the main aim of the work done in this work package, is the possibility to explore the impact of planned changes in London through the use of the joint scheme provided by our models and unconventional datasets.

1.1 Organisation of the document

This deliverable aims to describe the work done in the London case study during EUNOIA. Most of the project time was spent putting together all the inputs needed to run the travel demand model. Consequently, the next chapter in this deliverable will start by describing how we produced a representation of the transport infrastructure for London and its corresponding activity-based travel demand. We will mention and explain both, the datasets used and the methodologies adopted to produce them. We will also describe, still in this
chapter, our experiences with MATSim and our efforts to use it in order to integrate unconventional datasets. We will finish this chapter by comparing the outputs from the model to real life to get a sense of the validity of our approach.
2. Building the case study

Before describing into detail each of the various modules composing the model, we will start by defining the main points upon which our model is based.

- Our case study area is the region contained within the M25 with 9.4M inhabitants (Census 2011). A map of the region can be seen in Figure 1 or Figure 5 for instance, later in in this deliverable.
- Our base year is 2011 as the data needed was mostly available around that year.
- The activity types available to our agents are: home, work, shop, education, leisure and other –as vague as this definition may sound, some of the activities included within this category might include visits to hospitals or religion-related activities among many others–.
- We have not included freight and we have also not included commuters from outside the case study area or passers-by crossing the London region due to time constraints.
  - The impact of this assumption will be obvious on the counts. So far, we can quantify the number of in-commuters and out-commuters in the Greater London Authority to around 1M in 2007 (London Travel Report, 2007).
- The model includes four travel modes: private car, public transport, walk and bike. The public transport mode includes the usage of bus, underground, rail, the Docklands Light railway and the London Overground. Taxis have not been included.
- Finally, our analysis zones in most parts of the model have been the English Census 2011 Wards, which from now on we will refer to as wards. The case study area is composed by 850 wards. The population figures for these wards follow a normal distribution around 12000 individuals.

In this chapter we will present how an activity-based travel demand model has been defined for London going through each of the various steps involved in the process. We will start off by describing how we produced a representation of the transport infrastructure for our case study area in terms of the multimodal networks for all travel modes in London and also in terms of the land use configuration which shapes the City. This part of the model is key to the rest of the model components: information from the infrastructure is used as one of the inputs in all of them. We will then describe the methodology used to create a representation of the population in London which from now on we will call our synthetic population. In the next section we will describe how we assigned the sequence of activities for each synthetic individual, their corresponding timings and a location in our case study. In the following section we will present the results once we run all of the plans of each synthetic individual in our synthetic population using MATSim and we will finish off by comparing how well the current model reproduces observed data.

2.1 Supply networks: representing the transport infrastructure for London

A key component to these models is the representation of the supply networks where the agents will be simulated. This is a two-fold effort which includes not only the definition of the multimodal transport networks but also the definition of the land-use configuration in the case study area.
Our multimodal network is composed of a rather detailed representation of the road network where vehicles will operate and by timetable information of all public transport services operating within the boundaries defined for our case study area.

In terms of the road network, various issues need to be mentioned in order to better grasp the assumptions behind our implementation.

- The spatial extent of the network includes a detailed representation within the boundaries of the case study area and a less detailed network to represent the road network for the rest of the UK. The latter one only includes road links of type: ‘motorway’, ‘A road’ and ‘B road’ and the former one includes road links of type: ‘motorway’, ‘A road’, ‘B road’, ‘minor road’ and ‘local street’\(^1\). Figure 1 below shows the extent of the current road network for London. The total number of links included for London is \(\sim 520k\).
- Each directional link has been represented as a single entity in the model. This includes those road links which only have one direction.
- The resulting road network is strongly connected and has no loops.
- Routing information on ITN’s node grading system (related to intersections, bridges or tunnels) and ‘one way’ roads has been dealt with to get as close as possible to a true navigation network.
- In order to quantify the capacity of each road link in the network, MATSim requires various values,
  - the length of the link: taken from the original Integrated Transport Network database which considers curvature,
  - the total number of lanes: not available and so we deduced it taking into account the type of road it was part of – such as ‘motorway’ or ‘B road’– and its “nature”\(^2\) – such as ‘single/dual carriageway’ or ‘slip road’,
  - the speed limit: again not available so we also deduced it from information on the type of road link and its “nature” and
  - the maximum traffic flow allowed in the link in the form of vehicles per hour: in this case, we have used the guidelines proposed in \((\text{The COBA Manual, 2002})\) by the UK’s Department for Transport in order to quantify this measure. A description on how we proceeded with the capacity assignment can be seen in the Annex in page 28.

On the other hand, current issues to be improved in the future include:

- Turn prohibitions or enforcements have not been included in the network.
- No bus lane information found for the case study which means that all the lanes from our road links can be used by all sort of cars.
- Bike lanes have not included either.
- No traffic signals at intersections.

\(^1\) The following road types have been discarded: ‘alley’, ‘pedestrianised street’ and ‘private road’.
\(^2\) We have decided to adopt the terminology used in the original ITN dataset to describe this distinct property of road segments.
In regards to the public transport network, very detailed information has been obtained from timetable data held by the National Public Transport Data Repository (NPTDR). This data includes all public transport services running in England, Wales and Scotland between the 5th and the 11th of October 2009. The data is composed by two data sets: the NaPTAN (National Public Transport Access Nodes) dataset and the TransXChange files. The former includes all public transport nodes categorised by travel mode and geo-located in space. The latter one has a series of transport modal files for each county within England, Wales and Scotland (143 counties in total), with information on all services running within the county. The travel modes included in the data are air, train, bus, coach, metro and ferry. Each service includes routing information as a series of NaPTAN referenced stops each with its corresponding departure and waiting time. For our case study, we have selected those services which operate at, at least, one station which is included within the case study area. This approach will allow agents from outside the region the possibility to use public transport in order to get to their destination. The detail from the timetable has been preserved as MATSim allows for this level of detail to be included in the model.

Finally, and still in this section, we also need to define the land use configuration within our case study region which will be used to locate our agents’ households or guide their activity location choices. We have used Ordnance Survey AddressBase Plus layer which keeps address records for all of the UK with a definition of land use for each one of them. In our case, the following actions have been implemented in order to define them in our network:
• We have defined so-called activity locations within our case study area which we have located at the center of each road link in our network. Each activity location will include information on the amount of activity type addresses included based on proximity. For instance, a certain activity location might include 3 residential addresses, no offices, one shop, one school, no leisure centre and no address of type other.

• We have mapped the X different categories into 6 categories which best match the activity types present in the model. These are: home, work, shop, education, leisure and other.

The Annex in page 31 of this document shows visually the extent of the land use configuration resulting from aggregating the activity locations to each of the 850 wards in the case study.

2.2 An activity-based travel demand for London

The definition of the travel demand for this case study follows an activity-based approach. In the following two sections we will describe how it has been built. Two main steps need to be performed: the implementation of a synthetic population representative of the study area and the assignment of activities to each synthetic individual. These steps need to be performed in a manner in which the model outcomes will be somehow similar to what’s happening in the real world. The next two sections will describe our methodology and results after performing the aforementioned two steps.

2.2.1 The synthetic population

The first step starts with the generation of a synthetic population with similar socio-demographic characteristics than the real population in London in 2011. This is called the baseline synthetic population. There are various methodologies in the literature to do this – some examples can be found in (Beckman, Baggerly, & McKay, 1996; Feldman et al., 2007; Harland, Heppenstall, Smith, & Birkin, 2012; Lenormand & Deffuant, 2013; Namazi-Rad, Mokhtarian, & Perez, 2014). In our case, we have followed a rather similar approach than (Feldman et al., 2007) and we have applied a simulated annealing technique based on (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). In fact, a software is available online to apply this technique from (Harland, 2013) which is called Flexible Modelling Framework v1.2. This methodology uses two types of datasets: a comprehensive dataset which includes complete information of the aggregate socio-demographic statistics of our case study area i.e. Census data, and another dataset which includes very detailed socio-demographic statistics on both, individual households and the people living in it, from the same region ideally i.e. typically household survey samples. The microsimulation technique is an iterative process whereby for a single region, a number of households is selected at random which match the overall number of households present in that region – obtained from the Census-. Then, the resulting aggregate socio-demographic statistics from that first sample is compared to the aggregate Census statistics. In subsequent iterations, changes are suggested in the form of a potential replacement of a small percentage of the survey households which already belong to the current region sample and a new one. Only if the selected overall statistics improve the previous fit, then the change is accepted. The process stops when a predefined similarity margin is achieved between the sample and the Census statistics.

In the case of London, we have used a series of tables from the Census 2011 at the level of each of the 850 wards in our region to guide the simulated annealing process. These tables will establish the constraints. On the other hand, we have used Household Sample of Anonymised Records (HSAR) (2001), composed of ~225k surveyed households (~525k persons) in England (this corresponds to a 1% sample). Ideally one would use these
two datasets from the same year, however, in our case, the HSAR from 2011 was not available at the time we started the analysis, and so the assumption here is that household composition will somehow remain relatively similar between 2001 and 2011.

The constraints of the simulated annealing process were obtained from the following 3 tables from the Census 2011, whose unit of analysis relate to households:

- Table KS105EW: Household composition
- Table LC4202EW: Cars and vans by tenure and by ethnic group of household reference person
- Table LC6101EW: National Statistics Socio-economic Classification (NS-SeC) by sex by age of household reference person

Note also that while the first table includes a univariate constraint – only using household composition-, the other two include cross tabulated constraints.

Another very important issue to be mentioned at this stage is that we have adjusted the overall household figures from the three tables to a 10% of their overall value. We have adopted this measure to be consistent with the way MATSim performs the trip assignment step, in which both, the demand and the supply can be adjusted so that the processing will be faster. In other words, when running MATSim for large scenarios, the demand is usually reduced to a 10% of the total demand by selecting a 10% of the overall population in the study area and the capacity values in the transport infrastructure is reduced to a 10% too.

In terms of the HSAR, we could collect the same statistics than those from the Census 2011 table constraints so that the microsimulation process could be run.

The process was run for each of the 850 wards in London using the default settings from the software and we achieved convergence on all zones.

To evaluate the outputs of the software, we selected the following statistics from the Census 2011, which are independent of the statistics used in the simulated annealing process and relate mostly to individuals rather than households:

- Household size
- Age of the population
- Sex of the population
- Employment status

Figure 2 shows the comparison between the Census 2011 and from the London baseline synthetic population using the four variables mentioned. The patterns drawn by the distributions from the synthetic population and the Census 2011 on household composition at the top left part of Figure 2 are very similar. In terms of the population age bands –distribution from the top right-, the pattern defined by the Census data distribution is overall followed by the synthetic population too. The same can be observed for the overall gender patterns at the bottom left part of the Figure. We have also evaluated the relative error between the number of people employed by their ward of residence from the Census 2011 and the synthetic population. The distribution shown at the bottom right of the Figure shows most of the relative error being contained below 20%. This is not surprising considering the fact that we had no direct employment variable guiding the simulated annealing.
approach, and in fact, a previous run without using Table LC6101EW (which is the closest variable to employment through the use of NS-SeC), lead to worst results indeed.

![Comparison between household statistics from the Census and from the London synthetic population](image)

**Figure 2 Comparison between the household statistics from the Census and from the London synthetic population**

At this stage we will stick with the current results even though we are aware of some other methodologies which enable the use of constraints not only at the level of households but also at the level of population which would definitely improve our current results.

The following bit in this section will briefly describe how we have located each single household from each of the 850 wards in our case study into activity locations from our network. This is an important step which will set up the exact location of all potential trip origins in our case study. To do this, we have used the residential land use information in the form of total number of addresses in each activity location to guide the process. In fact, we have run a random generator on the following probabilistic distribution where $p_i$ is the probability of selecting activity location $i$ within a ward and $w_i$ is the weight associated to activity location $i$ in the form of number of residential addresses present in it:
This particular approach allows for the land-use heterogeneity within the ward to play a part on the number of households assigned within its own configuration.

Finally, just a final comment on the fact that we have no processes in place yet to forecast the evolution of the population through time.

2.2.2 Generating activities for individuals

This section will describe the steps taken to fully define the sequence of activities each synthetic individual will go through throughout the simulation period. This effort is twofold and is composed of the following steps:

- Assignment of the sequence of activities, their respective type and timings for each synthetic person
- Assignment of a specific location and the travel mode to get to it for each activity of each synthetic person

These two steps need to be performed so that the outcome is representative of the population in London. The techniques we have used to do this are based on TRANSIMS’ Activity Generator module (Barrett et al., 2001). TRANSIMS is an open-source travel demand model developed in early 2000 in Los Alamos National Laboratory which aimed at re-thinking the way more traditional models worked. In that sense, perhaps its most important features were the adoption of an activity-based approach and its characteristic cellular automata approach to simulate the movement of vehicles in the network in order to optimize computing performance. For our case study we will use exclusively the Activity Generator module. We will describe the techniques behind that particular module and how we have implemented them with contextual data in the following two sections.

2.2.2.1 Assigning skeletal activity patterns

As mentioned earlier, a set of activities needs to be assigned to individuals in some coherent way so that they are somehow indicative of the sequence of activities performed by Londoners. To do this, TRANSIMS Activity Generator uses a technique based on Classification And Regression Tree algorithms (CART) (Speckman, Sun, & Vaughn, 1998). At this stage it is necessary to introduce the concept of ‘skeletal activity patterns’: this is purely the sequence of activities performed by one individual with no other information than just the set of activities, their order and their type (timings, locations and travel mode to get to them are not included in the definition).

In that sense, the CART algorithm is used to produce clusters of survey households whose activity patterns are somehow similar through the use of socio-demographic data. In other words, this technique selects the set of socio-demographic statistics such as household size or number of workers per household, within the survey household population that produce the greatest variation in trip making activity –represented by measures such as time spent at home or time spent at work- between the decision tree categories. Once the tree is built, each end node of the tree will include a set of survey households with similar trip making activity. At this stage, the tree is used to assign each synthetic household to a given survey household through socio-demographic similarities between the two.

---

3 We have used v3.1 of its implementation which corresponds to the one developed in Los Alamos National Laboratory.
Another important step while using these trees is the process of pruning them to avoid over-fitting. In this case, we have used the cross-validation technique which consists on splitting the household sample into \( N \) parts and then building a single tree from each sample alone. From those \( N \) resulting trees, the one leading to the minimum prediction error will be selected as the right tree. This process assumes that by constraining the size of the sample, shorter, less over-fitting trees will be obtained.

In order to produce that tree, we have used the London Travel Demand Survey (LTDS) 2010/11 which we had access to thanks to TfL. LTDS is essentially a travel diary survey of 7901 households (18479 persons) in London which provides very detailed information not only on the socio-demographic characteristics for the surveyed households and persons within them, but also, on details related to the activities performed by each household member and their mobility choices around them such as timings or the mode of transport used to travel between activities. For our case study, we used those households whose survey travel day falls on a weekday (5717 households).

The specific CART algorithm we have used is the multivariate regression tree one and these are the variables collected for each survey household in order to build the tree:

- **Socio-demographic variables (independent variables):**
  - size of household (HHSIZE),
  - number of members under 18 (R18UNDR),
  - number of members over 60 (R60OVER),
  - number of members over 65 (R65OVER),
  - household income (INCOME),
  - number of non-relative members in household (RNONREL),
  - age of the household reference person (HHAGE),
  - number of workers in the household (WORKERS)

- **Trip making activity variables (dependent variables):**
  - These variables refer to the sum of the time spent performing each activity type by all the members in the household; the name of the variables is self-explanatory: THOME, TWORK, TSHOP, TEDUCATION, TLEISURE, TOTHER, NONHOME (the only trip-related variable which refers to the number of non-home based trips from all household members).
Figure 3 on the left, shows the binary CART for London. The first thing to notice from the resulting tree is that household size (HHSIZE) is always selected as the best variable whose splits provide the maximum decrement in the total deviance as compared to the decrements obtained considering all the other independent variables with their own split values. This is not surprising considering the nature of the dependent variables and their dependence on the number of household members. Another thing to notice are the figures in blue and in brackets next to each node; this correspond to the number of survey households included within each specific node. To evaluate the fitness related to the multivariate regression tree process, the coefficient of determination and the relative error using all trip-making related variables is calculated after each split. As splits progress, the pool of households within each node should be more similar, leading to smaller relative error values, as seen on the right-most plot in Figure 3. The final $R^2$ after running the process is 0.6.

It is at this stage and through this tree that we can now assign each synthetic household to a survey household based on socio-demographic similarities. Once the synthetic household reaches an end node, a survey household is selected at random and the activities from each survey person within the household are assigned to a member from the synthetic household. This step is done again based on socio-demographic similarities between them. It is after this that the skeletal activity pattern is assigned to synthetic individuals.
After running TRANSIMS’ Activity Generator we evaluated the skeletal activity assignment outcome using a similar approach than (Widhalm, Yang, Ulm, Athavale, & Gonzalez, 2014). In that sense, we have plotted the resulting hotspots of activity by their start time and duration using contour maps. It is through visual comparison that we are able to see whether the concentrations of activities occur at similar timings between the LTDS contour map and the one from the London model. The graphs from Figure 4 suggest this is the case. The timings of the activities will change after running MATSim, however, not much change is expected.

In terms of the activity timings, a random limited time increment is applied (post-TRANSIMS) only to the first activity which is sequentially propagated to the rest of the activities. Through this process we aim to relax the bias present on travel diary surveys in terms of activity timings i.e. individuals report parting from home at 8:15am to reach work at 9am while in reality there might be some time discrepancies to this. At the same time, by only changing the timing from the first activity, the time allowed for both the activities and the trips is still preserved.

### 2.2.2.2 Activity location assignment

To assign the specific location where all of the agents’ activities will take place, TRANSIMS runs a two stage assignment:

- The first stage involves a destination assignment at the level of wards
- The second stage selects one activity location from all locations within the previously selected ward

At this point, it is convenient to differentiate between two types of activities: mandatory and discretionary activities. As the naming implies, this classification refers to the difference between activities where there is some sort of obligation to them and those from which the individual could easily dispose of. This classification is a non-trivial one as it affects the choice of methodology. Hence, for mandatory activities we will use one type of activity location model whereas for discretionary ones, a slightly different one will be used. From our activity type categorisation, we have defined as mandatory activities: ‘work’ and ‘education’; and as discretionary: ‘shop’, ‘leisure’ and ‘other’. Note that the location for ‘home’ activities for each household has already been...
defined in the previous section, where each household has been located to the network. In that sense, we call the set of mandatory and home activity types as anchor activities.

The modelling technique to assign a location at the first stage is the multinomial logit choice model, based on discrete choice modelling. This technique is especially suitable in activity-based models as they can be directly applied to one individual. In the case of TRANSIMS, each synthetic individual evaluates the benefit of performing a specific activity at a particular destination as a composite value based on objective metrics associated with this destination (e.g. number of relevant addresses), objective metrics associated with travelling from origin to destination (e.g. travel time) and subjective components following a probability distribution. A full description of this specific model and discrete choice models in general can be found in (Ortuzar & Willumsen, 2011)(Train, 2009).

The utility function used to evaluate the benefit associated to a specific alternative for work-related activities is the following:

\[ V_L = \log(N_L) + b_m \times t_{H,L} \]

Where \( V_L \) is the utility of destination ward \( L \), \( N_L \) is the number of work-related addresses in ward \( L \), \( b_m \) is the parameter related to the travel mode used to get to \( L \) and \( t_{H,L} \) is the travel time between home location \( H \) and work location \( L \).

On the other hand, the utility function to evaluate the benefit of discretionary activities has a slight change. The formal name for the model is multinomial logit choice chaining model and takes into account three locations rather than two for its formulation. In the utility function for the work location model for a zone \( L \) we take into account the location of the household and then the location of \( L \). In the utility function for a certain discretionary activity \( L_j \), we take into account the location of the two anchor activities bounding the discretionary activity from the skeletal activity pattern and the location of \( L_j \). For example, say we want to calculate the utility for a certain shop activity at ward destination \( L_j \). The skeletal activity pattern from the synthetic individual triggering that query has the following sequence: work – shop – home, where work and home are the two anchor activities bounding the shop activity. The utility function for ward \( L_j \) in this particular case is the following one:

\[ V_{L_j} = b_{m_1} \times t_{L,L_j} + \log(N_{L_j}) + b_{m_2} \times t_{L_j,H} \]

Where \( V_{L_j} \) is the utility of destination ward \( L_j \), \( b_{m_1} \) is the parameter related to the travel mode used to get to \( L_j \) from work location \( L \), \( t_{L,L_j} \) is the travel time between work location \( L \) and shop location \( L_j \), \( N_{L_j} \) is the number of shop-related addresses in ward \( L_j \), \( b_{m_2} \) is the parameter related to the travel mode used to get to \( H \) from shop location \( L_j \) and \( t_{L_j,H} \) is the travel time between shop location \( L_j \) and home location \( H \).

It is through this model structure that the location decision on discretionary activities is based not only on the location of the previous activity –in this case, work activity location \( L \), but also on the location of the following anchor activity –in this case, home activity location \( H \). This seems to make sense intuitively as individuals might have higher consideration for shop locations on their way home. Note also that this sort of formulation is possible thanks to the activity-based approach where we know beforehand the skeletal activity pattern for all individuals.
The parameters needed by TRANSIMS to run these two models are the mode-related parameters to get to each activity type. The assumption here is that the travel mode used to get to each destination activity is initially preserved from the original survey person’s activity plan. We will look at mode choice through MATSim in the next section. Hence, a parameter needs to be calculated for each combination of activity types and travel modes. In order to do this, we have used the trip information available from the LTDS 2010/11.

Table 1 Calibration values for the $b_m$ parameter of the multinomial logit choice model

<table>
<thead>
<tr>
<th>purpose</th>
<th>mode</th>
<th>$b_m$</th>
<th>Adj rho-square</th>
<th>samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>work</td>
<td>walk</td>
<td>-0.00197</td>
<td>0.681</td>
<td>726</td>
</tr>
<tr>
<td>work</td>
<td>cycle</td>
<td>-0.00214</td>
<td>0.334</td>
<td>152</td>
</tr>
<tr>
<td>work</td>
<td>car</td>
<td>-0.00255</td>
<td>0.206</td>
<td>1772</td>
</tr>
<tr>
<td>work</td>
<td>pt</td>
<td>-0.00148</td>
<td>0.212</td>
<td>1345</td>
</tr>
<tr>
<td>shop</td>
<td>walk</td>
<td>-0.00243</td>
<td>0.814</td>
<td>2420</td>
</tr>
<tr>
<td>shop</td>
<td>cycle</td>
<td>-0.0061</td>
<td>0.568</td>
<td>72</td>
</tr>
<tr>
<td>shop</td>
<td>car</td>
<td>-0.00451</td>
<td>0.386</td>
<td>1837</td>
</tr>
<tr>
<td>shop</td>
<td>pt</td>
<td>-0.00402</td>
<td>0.395</td>
<td>1007</td>
</tr>
<tr>
<td>education</td>
<td>walk</td>
<td>-0.00233</td>
<td>0.728</td>
<td>660</td>
</tr>
<tr>
<td>education</td>
<td>cycle</td>
<td>-0.00443</td>
<td>0.464</td>
<td>32</td>
</tr>
<tr>
<td>education</td>
<td>car</td>
<td>-0.00558</td>
<td>0.447</td>
<td>444</td>
</tr>
<tr>
<td>education</td>
<td>pt</td>
<td>-0.00351</td>
<td>0.276</td>
<td>507</td>
</tr>
<tr>
<td>leisure</td>
<td>walk</td>
<td>-0.00235</td>
<td>0.729</td>
<td>1408</td>
</tr>
<tr>
<td>leisure</td>
<td>cycle</td>
<td>-0.00527</td>
<td>0.492</td>
<td>100</td>
</tr>
<tr>
<td>leisure</td>
<td>car</td>
<td>-0.0035</td>
<td>0.31</td>
<td>1401</td>
</tr>
<tr>
<td>leisure</td>
<td>pt</td>
<td>-0.00264</td>
<td>0.203</td>
<td>624</td>
</tr>
<tr>
<td>other</td>
<td>walk</td>
<td>-0.00262</td>
<td>0.768</td>
<td>49</td>
</tr>
<tr>
<td>other</td>
<td>cycle</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>other</td>
<td>car</td>
<td>-0.00668</td>
<td>0.474</td>
<td>71</td>
</tr>
<tr>
<td>other</td>
<td>pt</td>
<td>-0.00416</td>
<td>0.328</td>
<td>43</td>
</tr>
</tbody>
</table>

With the information available from the LTDS (only trips done on weekdays), the attractiveness values associated to each ward by activity type and an approximation of travel time between all pairs of wards in the case study area using crow-fly distances and average speeds by travel mode, we proceeded to calibrate the parameters. We run the calibration using Biogeme v2.3 (Bierlaire, 2003). Table 1 shows the values for the parameters. Even though the resulting adjusted rho-square values is mainly promising, some parameters are calibrated using samples of less than 100 occurrences, which is not ideal.

Once a destination ward has been picked running the corresponding discrete choice model, a specific activity location within the ward needs to be picked. This is done through a weighted probabilistic distribution considering the specific attractiveness of each activity location in relation to the rest of the activities within the ward.

To evaluate the location assignment we have used table WU03EW on location of usual residence and place of work by method of travel to work at the level of Middle Super Output Areas (MSOA) for 2011. That dataset is also referred to as the journey to work matrix and intends to provide a full picture of all work-related trips in England and Wales.
In order to proceed with the comparison, we had to produce OD matrices at the level of MSOA zones. Note that the definition of a ward is different than the one from the MSOA, the latter is smaller than the former (the region covered in our case study is composed by 850 wards or 1156 MSOAs). This is quite straightforward in our case as our origins and destinations are at the level of activity locations in the network so we only need to assign locations to MSOAs based on spatial inclusion.

Figure 5 shows the comparison of the spatial distribution of employment from the journey to work dataset (left) and from the model (right). The values from the choropleth relate to the sum of workers in each MSOA. The overall spatial distribution looks relatively similar for both with concentration of employment in the City of London and the highlight of Heathrow too.

The next comparison we have performed is two-fold on the actual flow values. On one hand, we have compared the flow values associated to each MSOA; that is the normalised number of trips from the journey to work matrix \( p_{ij_{obs}} \) with the normalised number of trips from the model \( p_{ij_{model}} \). We have run a regression (Figure 6, left) which leads to a coefficient of determination of 0.283. On the other hand, we have also looked at the resulting trip length distributions for the normalised journey to work dataset (observed) and for the normalised model home-to-work trips (model) (Figure 6, right). There is a mismatch in the trend for distances lower than 5km and for distances higher than 25km. On both mismatches, the model tends to under-predict the distributions from the data.
To finish this section we will mention the awareness on the margin of improvement considering the current model design and the inputs used. In this sense, we believe the following list of things to do next should increase the fit of the current results:

- Use of a better measure of attractiveness such as retail floorspace; by using only number of addresses, the attractiveness associated to big retail centres, for instance, might be underrepresented.
- Use of more realistic ward to ward travel times; the use of network-based travel times as opposed to crow-fly distances should lead to better accessibility values.
- Use of a more realistic travel cost through the use of generalised travel costs as opposed to distance or travel times; this measure should capture the pros and cons of the various travel modes in the case study in a more realistic manner.
- Use of better cost functions such as Box-Cox rather than the current negative exponential function to enable better calibration results.

At this stage then, the demand is fully defined with daily plans for all of the agents waiting to be loaded into the available travel infrastructure so that more realistic timings are produced.

### 2.3 Running MATSim

A part from the full definition of the travel infrastructure and the travel demand in the London case study which has been achieved and documented in sections 2.1 and 2.2 from this deliverable, MATSim requires a few more settings which we will now discuss.

These settings relate to the three activated modules in MATSim for London:

- Route assignment
- Time assignment
- Modal preference

Figure 6 Comparison between the normalised journey to work dataset and the model estimates using linear regression (left) and trip length distributions (right)
The route assignment is based on Dynamic Traffic Assignment methodologies. The modelling of the drivers’ behaviour on the road (as in car following or lane changing of drivers) is outside MATSim’s capabilities. In terms of the modelling of public transport in our case study, we have used a teleportation-based technique based on the schedule data. In that respect, even though MATSim allows for cars to interact with buses on the road network, this has not yet been tested for London, although we are in the process of testing it in the near future.

In terms of the time assignment, this module enables agents to randomly change of the timings of their activities within a set up time range. The overall plan utility values will respond to those changes taking into account their overall performance in terms of the activity-related utilities and the travelling (dis)utilities. One of the inputs which will affect the utility values associated to activities are the timings associated to each one of the activity types. These timings are in fact the main drivers pushing agents to reach each of their daily activities on time and to perform them throughout their ‘typical duration’ to maximize their utility.

### Table 2 Timings set in MATSim associated to activity types for the London case study

<table>
<thead>
<tr>
<th>purpose</th>
<th>opening time</th>
<th>closing time</th>
<th>Typical duration</th>
<th>minimal duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>home</td>
<td>NA</td>
<td>NA</td>
<td>8h</td>
<td>1h</td>
</tr>
<tr>
<td>work</td>
<td>9:00h</td>
<td>17:00h</td>
<td>8h</td>
<td>1h</td>
</tr>
<tr>
<td>shop</td>
<td>9:00h</td>
<td>19:00h</td>
<td>1h</td>
<td>5min</td>
</tr>
<tr>
<td>education</td>
<td>9:00h</td>
<td>NA</td>
<td>6h 30min</td>
<td>1h</td>
</tr>
<tr>
<td>leisure</td>
<td>9:00h</td>
<td>23:00h</td>
<td>2h</td>
<td>30min</td>
</tr>
<tr>
<td>other</td>
<td>9:00h</td>
<td>23:00h</td>
<td>2h</td>
<td>30min</td>
</tr>
</tbody>
</table>

Table 2 above summarises the activity timings which we have set taking into consideration plausible timings associated to UK standards in terms of work, education, shop or leisure.

In terms of the parameters associated to the various utility functions involved in the scoring of plans, we have adopted the default values based on (Vickrey, 1969) as suggested by the MATSim User Guide. This is relevant especially for our modal split in London. While the current approach might lead to plausible results, we understand this is not ideal as we should calibrate those parameters through a modal split model defined by the disutility of travelling:

$$V_{i}^{leg} = \beta_{tt,mode} \times t_{trav} + \beta_{m} \times m_{trav} + (\beta_{dist,mode} + \beta_{m} \times \gamma_{dist,mode}) \times d_{trav} + V_{transfer}$$

Where $V_{i}^{leg}$ is the utility associated to leg $i$ within an agent’s daily plan [utils], $\beta_{tt,mode}$ is an additional marginal utility of travelling [utils/hour], $t_{trav}$ is the time spent travelling during leg $i$ [hour], $\beta_{m}$ is the marginal utility of money in [utils/unit_of_money], $m_{trav}$ is the monetary position caused by travelling i.e. a toll or a fare [unit_of_money], $\beta_{dist,mode}$ is a marginal utility of distance in [utils/m], $\gamma_{dist,mode}$ is a distance cost rate in [unit_of_money/m], $d_{trav}$ is the distance of the leg and $V_{transfer}$ corresponds to transfer penalties incurred in public transport trips [utils].

---

[4] There is a note on the MATSim guide on the testing of the Vickrey scenario parameters with the parameters resulting from a modal choice model in Zurich, both leading to values of the same order of magnitude.
One of the features related to MATSim which we have enabled considering our current activity-based approach is the use of what is called in MATSim terms as “chain based” processes for car and bike modes. This setting recognizes the use of cars or bikes, and the dependence involved on the agent using them for their travel mode choices throughout their daily plans. In other words, if the agent used the bike to get to work, the bike will remain at the work location waiting to be used, if needed, for the rest of the trips of the agent’s daily plan.

![Score Statistics](image)

*Figure 7 Score statistics for the London case study run using MATSim*

Finally, these three MATSim modules have been applied to only a proportion of all agents in a simulated annealing fashion. Hence, the core iterative process based on the following sequential loop:

replanning \(\rightarrow\) execution (mobsim) \(\rightarrow\) scoring,

is applied only to a proportion of the agents at each iteration so that the process can slowly converge to a solution. In our case study, we have set the following probabilities for each module which define the chance of an agent being picked to modify their last plan setting: (route assignment, time assignment, modal preference) = (0.2, 0.1, 0.1). In other words, at each iteration, 20% of agents are selected to revisit their routes, 10% to revisit their plan timings and 10% to switch modes. Still on this issue, this percentages have been disabled from iteration 190 to 200 so that the best plans can be consistently selected without any more randomness or
exploration added. This is formally known as disabling the innovation phase. The effect of this can be seen on Figure 7.

With these settings, we run MATSim and it took almost 6 days with an average memory consumption of around 60GB to go through the 200 iterations on a server with the following specifications:

- Windows Server 2012 R2 Standard
- Intel(R) Xeon (R) 3.30GHz (16 nodes)
- RAM: 128GB
- 64bit operating system

At the time of the writing of this deliverable, only one full run could be performed and so the issue of the robustness of the current results taking into account the stochastic nature of MATSim remains to be further explored.

### 2.3.1 Using Oyster card data to calibrate modal preference

One of the main objectives in EUNOIA is the use of non-conventional datasets in the design part of travel demand models. For the case study in London, we have devised a methodology to refine the choice of travelling by tube within the current modal preference framework in MATSim using the Oyster card data.

The Oyster card data provided to us by TfL is the sample of all users between 16th of June and 9th of September 2012 (86 days in total). Through this sample, we only select those trips for which we have a tap in and the subsequent tap out. Therefore, all records relating usage of buses have been discarded. To put this data into context, 85% of all tube users in London use the Oyster card.

The selection of travelling by tube within the public transport travel mode is based on the use of the timetable travel times and these are currently fixed in our case study. However, more realistic costs between pairs of stations at specific time periods during the day can be obtained from the Oyster card data. We are planning to update these costs iteratively in order to slowly adjust the choices of our agents to the new updated costs. In other words, through this process, we are looking to account for the discrepancies between the current timetable-based travel times and the Oyster card ones. We are planning to do this through the following formulation:

\[
C_{q+1}(s_i, s_j, t_k) = C_q(s_i, s_j, t_k) + \Delta C_q(s_i, s_j, t_k)
\]

\[
\Delta C_q(s_i, s_j, t_k) = \frac{1}{\lambda} \log \left( \frac{N_{model}^q(s_i, s_j, t_k)}{N_{obs}^q(s_i, s_j, t_k)} \right)
\]

Where \(s_i\) is the id of the entry station, \(s_j\) the id of the exit station, \(t_k\) the time period during which the measures collected apply (using 15 minute periods), \(N_{obs}^q\) the number of agents using entry station \(s_i\) and exit station \(s_j\) during time period \(t_k\) using Oyster card data, \(N_{model}^q\) the number of agents using entry station \(s_i\) and exit station \(s_j\) during time period \(t_k\) at iteration \(q\) in MATSim, \(C\) is the associated travel costs using entry station \(s_i\) and exit station \(s_j\) during time period \(t_k\) and \(\lambda\) is the parameter guiding the level of adjustment of the costs.

The first equation defines how the overall transport costs between all pairs of stations can be modified if Oyster card occurrences are present in the data. The variation from the original transport costs (timetable based if \(q = 1\)) is calculated via the second equation.
In technical terms within MATSim, we have developed a new scoring function to run this after the mobility simulation module so that the updated costs are available for the next iteration’s replanning module.

This is the theoretical formulation we have devised to include the Oyster card data actively within the modal split calibration process. Tests within MATSim need to be done to evaluate the convergence of the process.

2.4 Case study validation

In the UK, the Department for Transport is in charge of an extensive document reporting a series of guidelines for the use of techniques related to urban transport models. It is called the *(Transport analysis guidance - WebTAG, 2014)*. This document suggests three tests in order to establish the quality of any urban model:

- Validation. This step consists in the comparison between a set of model outputs and observed datasets.
- Realism testing. The model needs to respond realistically to standard changes in the model inputs.
- Sensitivity testing. The model outputs should be robust to changes in the values of the parameters intervening in the model.

In this section, we will center the discussion around the validation after the traffic assignment step, in our case, MATSim. Even though the London case study includes both, private car assignment (also known as highway assignment in WebTAG) and public transport assignment, we will perform tests only on the private car travel mode.
In terms of the highway assignment, we have looked at the traffic count values at screenlines and cordons within the case study area. These can be seen in Figure 8 above. These counts correspond to 2012 weekdays and they report values for AM peak (8-9h), Inter-peak (average from 10-16h) and PM peak (17-18h). In this case we are using car-related counts.
As the results from the comparison in Figure 9 show, the current model travel flows underrepresent the observed count values. We believe these results are not surprising considering the fact that:

- we are only modelling travelers living and working within the case study area which means that in and out-commuters, which play an important role in London, are not included in the resulting model count data
- we have not yet refined the parameters involved in calibration of the modal share for London and
- the variables involved in the calibration of the destination choice, that is travel costs and attractiveness values, should be more realistic.

At this stage, subsequent changes in the identified modules should lead to improvements in the error values of traffic counts.
3. Conclusions

Work done in the London case study has led to a framework which enables the testing of many mobility-related issues. We have built this framework taking into account discussions held with TfL which showed an interest on activity-based models, dynamic traffic assignment and the use of the Oyster card data in models of this sort.

The subsequent refinement of the identified framework modules will be key to enable the testing of the current theoretical formulation to use Oyster card data in the mode choice process.

The current framework allows the testing of various mobility-related policies on a short, mid and long-term basis. In order to run long-term scenarios, the joining of this framework to a Land Use model is key to reproduce the reciprocity between the land use configuration and the transport infrastructure. This will involve the building of linkages to the SIMULACRA land use model at CASA (Batty et al., 2013).
Annex I. References


Annex II. Abbreviations and Acronyms

CART: Classification And Regression Tree algorithms.

COBA: COst Benefit Analysis.

HAM: Highway Assignment Model.

HSAR: Household Sample of Anonymised Records.

LoRD: London Regional Demand model.

LTDS: London Travel Demand Survey.

LTS: London Transportation Studies model.

MSOA: Middle Super Output Areas.


NPTDR: National Public Transport Data Repository.

NS-SeC: National Statistics Socio-economic Classification.

TfL: Transport for London.
Annex III. Capacity values for road links in the London case study

The values used to quantify the maximum throughput for each link is an important issue which has a big impact on the trip assignment step performed by MATSim. In order to calculate those values we have used the information available from the *(The COBA Manual, 2002)* in the following manner. Part 5 (Speed on Links) of the manual classifies roads into 11 categories and, for most of them, a margin of capacity values are provided based on variables which might affect that value in one way or another such as the percentage of heavy vehicles crossing that road links of that class. In this annex, we will go through each of the 11 road classes, we will provide its corresponding capacity equation (when there is one) and we will list the variables involved to get to the value. We will finish with a diagram which summarises the decision process to set the capacity values for the links in the case study in London. Some classes have been grouped together based on them using either the same equations or the same typical capacity values.

The 11 road classes from the manual are:

1. **Rural single carriageway**
   a. Capacity equation:
      \[
      Q_c = \frac{2400 \times (CWID - 3.65)}{CWID} \times \frac{(92 - PHV)}{80}
      \]
   b. Variables involved:
      i. CWID: Average carriageway width between white lines edge markings, excluding any painted out portion (m); (min, max) = (6, 11)
      ii. PHV: Percentage of heavy vehicles; (min, max) = (2, 30)
      iii. Qc: Capacity flag defined as the maximum realistic value of Q (vehs/h/dir); (min, max) = (900, 1600)

2. **Rural all-purpose dual 2-lane carriageway**
3. **Rural all-purpose dual 3 or more lane carriageway**
   a. Capacity equation:
      \[
      Q_c = \frac{2100}{(1 + 0.015 \times PHV)}
      \]
   b. Variables involved:
      i. PHV: Percentage of heavy vehicles; (min, max) = (2, 30)
      ii. Qc: Capacity flag defined as the maximum realistic value of Q (vehs/h/dir); (min, max) = (1400, 2250)

4. **Motorway, dual 2-lanes**
5. **Motorway, dual 2-lanes**
6. **Motorway, dual 4 or more lanes**
   a. Capacity equation:
      \[
      Q_c = \frac{2330}{(1 + 0.015 \times PHV)}
      \]
   b. Variables involved:
      i. PHV: Percentage of heavy vehicles; (min, max) = (2, 30)
ii. \(Q_c\): Capacity flag defined as the maximum realistic value of \(Q\) (vehs/h/dir); (min, max) = (1400, 2250)

7. Urban, non central
8. Urban, central
   a. \(Q_c\): Capacity flag defined as the maximum realistic value of \(Q\) (vehs/h/3.65m lane); the typical value is set to 800.
9. Small town (population < 70000)
   a. \(Q_c\): Capacity flag defined as the maximum realistic value of \(Q\) (vehs/h/3.65m lane); the typical value is set to 1200.
10. Suburban\(^5\) single carriageway
11. Suburban double carriageway
   a. Equation:
   \[
   Q_c = \frac{1500 \times (92 - PHV)}{80}
   \]
   b. Variables involved:
      i. PHV: Percentage of heavy vehicles; (min, max) = (2, 20)
      ii. \(Q_c\): Capacity flag defined as the maximum realistic value of \(Q\) (vehs/h/3.65m lane); (min, max) = (1350, 1700)

The diagram which summarises the 11 eleven classes above can be seen in Figure 10 below.

<table>
<thead>
<tr>
<th>COBA Class</th>
<th>4, 5, 6</th>
<th>2, 3</th>
<th>1</th>
<th>10, 11</th>
<th>7, 8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_c) min</td>
<td>1400</td>
<td>1400</td>
<td>900</td>
<td>1350</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(Q_c) max</td>
<td>2250</td>
<td>2250</td>
<td>1600</td>
<td>1700</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(Q_c) typical</td>
<td>1902</td>
<td>1714</td>
<td>1155</td>
<td>1443.75</td>
<td>800</td>
<td>1200</td>
</tr>
</tbody>
</table>

\(^5\) The COBA Manual defines suburban in the following manner: “the suburban speed relationships apply to the major suburban routes in towns and cities where the speed limit is generally 40mph (64kph)”
Through the diagram in Figure 10, and road classifications in terms of road type (‘motorway’, ‘A road’, ‘B road’, ‘minor road’ and ‘local street’) and road nature (‘single carriageway’, ‘dual carriageway’, ‘slip road’, ‘roundabout’, ‘enclosed traffic area link’, ‘traffic island link’ and ‘traffic island link at junction’), we have set the capacity values associated to each road segment.
Annex IV. The land-use configuration for the London case study

The plots below shows the land use configuration resulting from assigning the land-use information for each activity locations to each of the 850 wards in the case study based on spatial containment. The values in the plots correspond to number of addresses normalized by the area of each ward. The Natural Jenks categorization has been followed in the following plots.

Figure 11 Land use configuration corresponding to residential activities in the left and work-related activities in the right
Figure 12 Land use configuration corresponding to retail activities in the left and education-related activities in the right.
Figure 13 Land use configuration corresponding to leisure activities in the left and other-related activities in the right.
Annex V. MATSim, an overview

Philosophy
MATSim is a software framework, aiming at representing interactions between transport supply and mobility behaviour with a high level of details. It is based on the convergence of three established streams of research, namely:

- activity-based mobility behaviour analysis, that explicitly looks at travel as a consequence of the desire of individuals to perform activities, to represent demand,
- microscopic traffic simulation methods, allowing to represent complex but important effects in traffic flows, such as queue spillback, as well as other effects of individuals’ interactions, such as crowding of public transport vehicles,
- game theory and co-evolutionary computation methods, that provide a theoretical ground for modelling the interactions of supply and demand, as well as practical and efficient ways to find approximate solutions for the resulting theoretical model.

The software was implemented in the Java programming language, which is multi-platform and relatively easy to learn for modellers with little programming background. It is free and open-source, meaning any interested researcher or modeller can analyse, use or extend it for his or her use (see www.matsim.org). The software is designed to allow usage in large-scale simulations, in particular by adopting a minimal yet powerful traffic flow model.

Modelling Principles: Game Theory and Co-evolutionary Computation
Given the philosophy above, the aim of MATSim can be summarized as modelling the influence of individuals’ interactions in the transport network on their activity and trip planning decisions. Modelling the interaction of individuals with possibly conflicting objectives has been the subject of game theory for decades, making this theoretical framework particularly well suited for the problem at hand.

A game theoretic view of transportation systems has indeed been popular for more than half a century, since the seminal work of Wardrop (Wardrop, 1952). The essential idea behind it is to see the transportation system as a set of shared resources (road space, public transport vehicle seats...) for which individuals compete and individuals in the population trying to maximize their own satisfaction given the resources left available by others. Game theory studies solution concepts for such strategic interactions. A game theoretic solution concept is a definition of which states are equilibria, that is, stable under assumption of rationality — a state being considered stable if no agent/player has an incentive to change its behaviour. The static, trip-based approach of Wardrop has been refined and extended with time. In particular, the equilibrium idea can be pretty naturally transferred to the activity based framework: individuals do not just try to optimize their trips, but their whole day. The implementation of this idea in a practical, usable software framework, was the aim of the MATSim software framework from the beginning (Axhausen, 2006; Nagel & Flotterod, 2009). MATSim simulates agents having daily activity plans, consisting of activities located in space, and the trips between those activities.

Co-evolutionary algorithms are particularly well suited for searching for states approximately satisfying game theoretic solution concepts, when the search space is too big for exhaustive search (Ficici, 2004; Popovici, Bucci, Wiegand, & De Jong, 2012). In this kinds of algorithm, an evolutionary process is performed on sub-solutions that are evaluated by interaction. In the case of MATSim, those sub-solutions are individual daily plans which are evaluated by interaction in a mobility simulation which represents congestion. This process can be seen as
an emulation of a learning process (Nagel & Marchal, 2003). The steps of this process, represented on Figure 1, are the following:

1. **Initial demand.** All agents have an initial daily plan, which will serve as a starting point for the iterative improvement process. Some characteristics of the plans are left untouched during the simulation, and should therefore come from data or an external model. This is typically the case of long term decisions, such as home and work locations, or decisions involving a larger time frame than a single day e.g. do the weekly shopping or not.

2. **Mobility simulation.** Plans of all agents are executed concurrently, to allow estimating the influence of the plans of the agents on each other.

3. **Scoring.** The information from the simulation is used to estimate the score of each individual plan. This information typically takes the form of travel times and time spent performing activities; experiments also include information such as facility crowding (Horni, Scott, Balmer, & Axhausen, 2009). The functional form is the one used by (Charypar & Nagel, 2005). It uses a linear disutility of travel time, and a logarithmic utility of time passed performing activities. Different parameters can be defined for each mode/activity type.

4. **Replanning.** This step actually groups two of the important components of co-evolutionary algorithms: (a) selection of the interactions for evaluation, and (b) application of the evolutionary operators (selection and mutation).

   To do so, part of the agents select a past plan based on the experienced score, following a Logit selection probability. This will have two consequences: (a) the state of the transport system, used for evaluation, will only evolve slowly from iteration to iteration, giving the time to the agents to adapt, and (b) those plans will be re-evaluated, given the new plans of the other agents. The other agents copy and mutate one of their past plans. If the number of plans in an agent’s memory exceeds a predefined threshold (usually 4 or 5), the worst plan is deleted, pushing the evolution towards plans with higher scores. Steps 2 to 4 are then iterated until the system reaches a stable state.

   What kind of mutation is performed determines which alternative plans will be tried out by the agent. Typical replanning strategies include least cost rerouting using travel time estimates from the previous iteration, departure time mutation, and mode mutation at the subtour level, considering mode chaining constraints. A tour is a sequence of consecutive trips starting and ending at the same location, named anchor point. A subtour is a tour, possibly without other tours it contains. Vehicular modes can only be performed for whole subtours, which must be anchored at home or in subtours of the same mode. Experiments included secondary activity location choice (Horni et al., 2009) and activity sequence (Feil, 2010).

Those steps are iterated until a stationary state is reached, and the state of the system in this stationary state is taken as a result.

**The Physical World in MATSim**

With the scoring function, an important element is the *mobility simulation* which is used to evaluate the influence of agents on each other’s satisfaction.

The standard mobility simulation uses a step-based queue model to simulate road infrastructure, which agents might use with a car, or by boarding a public transport vehicle. Other modes are assumed not to suffer from
congestion, and are simply “teleported” with a pre-computed travel time. This simulation framework was found to allow fast large-scale simulation, while still representing the essential dynamics of congestion.

To Sum Up
MATSim is a freely available open-source simulation software which has already been used in Switzerland (Meister et al., 2010), Berlin, or Singapore (Erath et al., 2012), to cite only a few examples.

It searches for an equilibrium for daily plans in a synthetic population, where individuals cannot improve the satisfaction they get from their day, given the state of the transport system, that depends on how individuals use it.