ABM of Turin's public transport and users' behaviour regarding the choice to validate the travel document

Alessandro Greco
Emanuele Pepe
Mattia Furlan

University of Turin
A.A. 2016/2017

---

1. Introduction

The public transport system is undoubtedly an important resource for the community, since it permits to reduce air pollution and traffic congestion considerably. Moreover, it will be improved to such an extent that its predominant role in the field of the future green city planning can be imagined. City11 administrations must guarantee a good service, but making it more affordable will be the real challenge. A leading role in this process is surely played by the ticket inspection. The moral sense of citizens should be supported with the aid of an inspection and penalty strategy which should be as universal and efficient as possible and at the same time, it should instruct the citizen in respecting the travelling rules. The public transport system is a complex system and we believe that a ABM simulation approach can nowadays be possible.

Our work consists of constructing a public transport network concerning the city of Turin starting from the available GFTS data and a mobility model thanks to the data made available by Piedmont's mobility agency[1].

Subsequently, intelligent agents travelling through this network classify their own experience and learn from that, whereas the inspecting agents check the tickets with the aim to increase the risk perception of the travellers. The result is a system whose evolution we aim to examine in order to define and find the best strategies to check the ticket.

2. Network construction

The geographic region considered for the network realization is composed of the City of Turin and its suburbs. It is represented as an ensemble of 265 parts, smaller areas, according to the partition provided by the Mobility Agency.
We reconstruct a structure based on bus stops and itineraries exploiting the available GTFS data. We proceed associating the latter with both a list of bus stops and the average frequency of transits for each hour of the day; time unit for the aforementioned GTFS data. Moreover, each couple of bus stops are characterized by an average travel time.

The transit network is modelled as a network with the following features:

- multilink: There can be more edges linking two stops, corresponding to different routes linking them
- weighted: every link has a weight that depends on the frequency of transit for the bus (with a specified itinerary) linking two bus stops
- Time-dependent: The frequency of transit changes with the hour of the day, therefore so does the weight of links.

We chose to identify every bus stop with its area. This approximation retains the most relevant information of the network (waiting time, changing bus) while allowing us to reduce the computational load for our simulations.

The paths between O-D couples (Origin-Destination) have to be computed in order to minimize an objective function defined as the sum of: travelling time, waiting time at the bus stop and time required to move between two nodes of the network by foot. By construction, this procedure disadvantages the change of routes, or "itinerary" in our code, during a travel.
We could not find an algorithm suitable for our purpose and because of that we created a custom one.
The algorithm is made up of the following steps:

1. Definition of the *path* class. It is composed of: a unique identification code, a list of itineraries, a list of bus stops, travel time and schedule duration as attributes.

2. Computation of every possible *path* between every couple of bus stop, by aggregating every possible combination of subsequent edges.

3. Walking paths computation. The upper bound of the walking time is 15 minutes which corresponds to 1.25 km using a Preferred Walking Speed of 5 km/h[2].

4. Composition of every possible couple of *path* merging its end with the head of another *path*, in order to allow changing bus and walk before or after a trip on a bus. Loops are excluded. Between paths with same O-D we pick the three faster in travel time.

5. Reiteration of step 4, extending ensemble of paths to 3 changes, or a change and a walk. Again, we pick best three trips for every O-D couple.

For paths for which the waiting time exceeds 15' this has been reduced to 10' because in these cases users tend to know when to go to the bus stop[3]
The implicit assumption that three is the maximum number of possible routes changes allows the reduction of the number of paths which must be taken into account while reducing the computational time.

3. Modelling of O-D Matrixes

OD matrixes are a standard way of representing the demand for public transport with variable precision. In these matrixes, the *i-th* element of the *j-th* row represents how many users will move from the i-th area to the j-th one in the time resolution offered by the matrix.
We interpret a travel as a stochastic process, the random variable is the vector Origin(t), Destination(t))

The data we received from the Agency consists of marginalized probabilities, p(O) and p(D), concerning mobility of the public transport. Our aim is to obtain p(O,D) by making hypotheses about the conditional distribution of destinations given the origins.

$$p(D) = \sum_{o} p(D|O)p(O) = \sum_{o} p(O,D)$$

The hypothesis about the shape of p(D | O) is left to our imagination, using the data in our possession as a constraint.

Looking closer at the matrixes, we distinguish different resolutions depending on the geographic area:

- Matrixes with origins and destinations for every hour in the city of Turin (translating in p(O) and p(D));
- Aggregate information about the surrounding area
- It is possible to obtain an OD matrix for the three macro-areas: City of Turin, Surrounding Area, remainder of the metropolitan area from data presented in the 2008 Mobility Report.

So, for example, for trips from a central area in the city of Turin to a suburban one, we only know p(O) precisely, whereas the information about the destination is known in an aggregated form and vice versa.

3.1. Null model

In this model, we have not introduced information other than the one contained in the above-mentioned data. In other words, we imposed the less biased assumption for the conditional probability, the uniform distribution:

In this relation, we consider the uniform distribution for p(D | O):

$$p(O,D) = \begin{cases} 0 & if\ O \equiv D \\ k \cdot p(O) & otherwise \end{cases}$$

We imposed a posteriori that the probability of going from a stop to the same stop in an hour is null.
We get the value of k through the following relation of marginalization:

$$p(D) = \sum_{O \neq D} p(O,D) = k \times \sum_{O \neq D} p(O) \Rightarrow k = \frac{p(D)}{\sum_{O \neq D} p(O)}.$$

It is possible to verify that the normalization of p(D) and p(O) is sufficient to obtain a normalized joint probability p(O | D):

$$\sum_D \sum_O p(O,D) = \sum_D \sum_{O \neq D} \frac{p(D) \cdot p(O)}{\sum_{O \neq D} p(O)} = \sum_D p(D) \cdot \frac{\sum_{O \neq D} p(O)}{\sum_{O \neq D} p(O)} = 1$$

If any of the two zones belongs to the surrounding area, let's say D, we cannot use the described procedure. The only known value is, indeed, $\sum_{surr.area} p(D)$. For this reason, we must use the following relation in order to find k:

$$\sum_{surr} p(D) = \sum_{surr} \sum p(O,D) = k \times \sum_{surr} \sum_O p(O)$$

$$\Rightarrow k = \frac{\sum_{surr} p(D)}{\sum_{surr} \sum_O p(O)}$$

We notice how k is independent from the destination, and $p(O,D)$ depends only on O.
Analogously for origins.

If both O and D do not belong to the city, we only know aggregation for both, thus we cannot write a relation like (1), and we use the less informative formula:

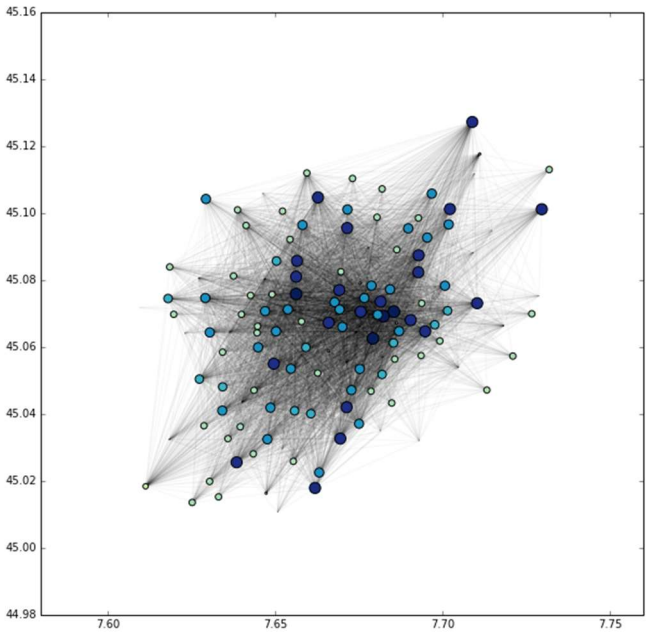
$$p(O,D) = k$$

where k is calculated as follows:

$$\sum p(D) = \sum_{area} \sum_{O} p(O,D) = k \cdot \sum_{area} \sum_{O} 1$$

$$\Rightarrow k = \frac{\sum p(D)}{\sum \sum_{O} 1}$$

In this way, it is possible to obtain a constant p(O,D) for each couple of nodes in the suburbs. This model has the advantage of using only the provided data, on the other hand it does not take into account the spatial relations among different geographical areas. In other words, a travel which crosses the city is as probable as a travel among close areas for equal probabilities p(O) and p(D). This is a clearly unrealistic limitation of this null model.



OD matrixes can be easily interpreted as adjacency matrixes for a weighted temporal network. This network differs from the onr we described above because represents different data, but helps in the visualization of the public transport demand. In the picture, a representation for the OD matrix at 4 p.m.

3.2. Poissonian model

In order to take into account the probability to take a bus in order to reach a destination which is not close it is possible to define a poissonian joint probability distribution p(O,D)

dependent on the distance d(O,D) (this parameter can be both seen as spatial distance and temporal distance). This distribution has several advantages, firstly it depends only on one parameter and secondly it penalizes both near and very far destinations.

p(O,D) becomes:

$$p\,(O,D) \;=\; e^{-\lambda}\frac{\lambda^{d(O,D)}}{d(O,D)!}\cdot p(O)$$

λ is the parameter to be defined through p(D)

$$p\,(D) \;=\; \sum_{O} p\,(O,D) \;=\sum_{O} e^{-\lambda}\cdot\frac{\lambda^{d(O,D)}}{d(O,D)!}\cdot p(O)$$

This equation can be inverted and the resulting transcendental equation can be numerically solved obtaining p(O,D) which is again unique for each OD couple. The previous considerations can be repeated for areas of the suburbs, about which only generic information is available.

Both the hypotheses and the constraints we considered are still valid if we use, if we use the actual count of travellers on each OD couples provided by the Agency instead of probabilities. Consequently, non-normalized p(O,D) represent the number of travellers moving during an hour.

3.3. Corrections

Travels crossing the borders of the network

The provided matrices contain the information associated with trips that start or end outside the borders of the city of Turin. The percentage of such travels is not negligible and for this reason they must be taken into account. More specifically, we have decided to redistribute these fluxes of passengers on the border areas on the city (identified through QGIS) proportionally to the probability associated to each specific area; p(O) or p(D) according to the nature of the travels under analysis.

Correction for the network topology

In order to improve the movement logic of the agents, a first modification of the null model, regards the introduction of the information associated with the topology of the network. The aforementioned matrices obtained from the data are the only constraints that have to be introduced in our models and they are only related to the demand for public transport.

For this reason, they must not consider:

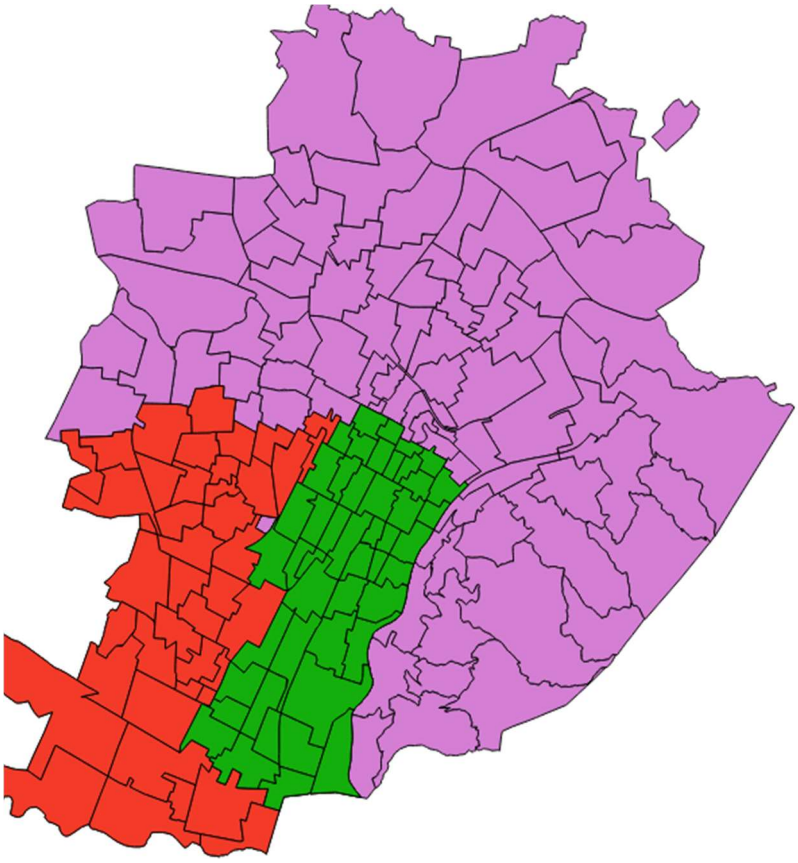
- destinations easily reachable by walking,
- destinations not reachable through the public transportation system (according to our network)

The first condition can be analytically expressed as:
$p\,(O,D) = \gamma_{OD} \times k\,(D) \times p\,(O)$, with $\gamma_{OD} = 0$ for (O,D) couples linked through a journey which can be covered only by foot or not linked at all and $\gamma_{OD} = 1$ otherwise. This relation keeps the normalization of the joint probability.

4. Network reduction

An initial simulation of the network composed of bus stops and itineraries has highlighted the computational difficulties associated with the management of such an enormous structure. For this reason, we have decided to take into account only a portion of the network by progressively reducing the area under analysis; we preferred this approach rather than reducing the resolution of the entire network, in order to exploit its features when our computational resources grow.

The cutting procedure is designed in such a manner that the passengers flux is not altered.



Red area quarters: Pozzo Strada, Cenisia, Borgo S.Paolo, Mirafiori nord, Mirafiori sud, S.Rita.
Green area quarters: Centro, Crocetta, S. Salvario, Lingotto, Nizza.

5. Agent based model

5.1 House distribution

The aforementioned matrices, obtained from the data analysis, can be easily interpreted as transition matrices of a finite stochastic process (with some necessary adjustments, for example defining the diagonal elements that are excluded from the ODs because they do not cause any movement on the network, but they represent the probability of remaining in the same place). However, describing the mobility of the agents merely as a random process would be unrealistic. For this reason, it is necessary to construct a spatial distribution of fixed origins, domiciles, and destinations, variable locations defined according to the information provided by the mobility agency about the purposes of the mobility.

The OD matrixes share some similarities with transition matrixes of a Markov process and can be interpreted as such with some necessary adjustments (normalization and probability of staying in the same state, not addressed in the OD matrixes). Following this interpretation, the compatibility of the OD matrixes with a model with fixed houses for every agent can be verified.
By multiplying the 24 different OD matrixes we obtain the transition matrix over 24 hours from one state to another. We expect this matrix to be close to the identity, because every agent should start its daily routine from the same location (its house). The closeness of these two matrixes is a necessary condition for the creation of a model with fixed origins for agents.
The frobenius norm of the difference between the product of the normalized OD matrixes and the identity is compared to the norm of the first.

$$\Delta = \frac{|\mathbb{1} - \prod_i OD_i|}{|\prod_i OD_i|}$$

The analysis showed small values for Δ ($\sim 10^{-2}$), so a fixed origin model is viable.

5.1.1 Heuristic model

This model is based on the idea that the OD matrixes, can address the distribution of/ among homes on the region, together with the hourly distribution of purposes. Of course, it is not strictly justifiable because it makes implicit assumptions of homogeneity in the trips with the purpose of returning home.
The home distribution is modelled as follows:

$$p_{HOME}(A) = \sum_{h=0}^{23} p_D^h(A) \times p^h{}_{return\ home}$$

This equation induces a small systematic error, in fact a traveller could go home several times in a single day, over-representing the destination in the probability distribution. As another consequence of multiple trips to an agent's home, this formulation of p(home) leads to overestimate the number of passengers which benefit from the public transport system.

This model only returns a probability distribution over the bus stops, that is combined with data from the 2008 Mobility Report: the mobile population is composed of 826000 units for the entire city of Turin, considering a percentage of it (26.4%) that uses the public transport, we estimate 218604 units for the OD matrixes. The area selection for the network reduces this number to 131766 for the smallest area.
The probabilistic description of the destination extraction prevents our approach from guaranteeing the return for every agent.
Assignation of a daily routine to every agent is the first phase of the model in Gama.
For this part, we decided to ensure closed routines for every agent. Starting from a house distribution outputted by the heuristic model, we ensure a closed routine for every agent by considering different routines for the agents that do not return home at the end of the day after changing their initial position (house).

The definition of mobile population is of people that move at least once a day. This excludes a relevant portion of the moving population, that is the "occasional" travellers, which travel on average less that once a day. This portion of the population must be considered but cannot be inferred from data. We plan to introduce in the future an occasional number of people moving randomly on the network that choose whether to buy a ticket or not.

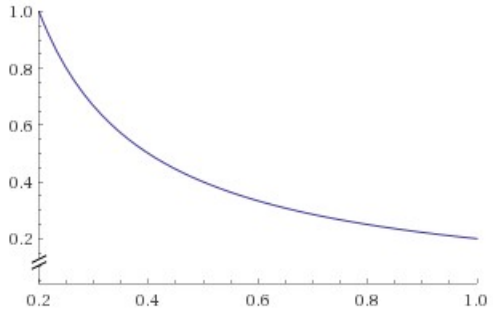
5.2 Path choice

We computed three different paths linking the origin and the destination of a trip. An agent chooses the option that maximizes a score function computed on the path. This function takes the saturation of the path in the first segment and the time duration in input and evaluates the score as follows:

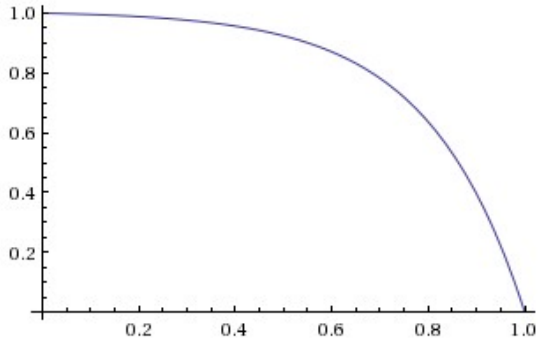
$$score = \frac{duration_{min}}{duration} \cdot \frac{1}{1 - e^{-5}} \cdot \left(1 - e^{(5 \cdot (x-1))}\right)$$

In this formula, x describes the saturation of the mean of transport and is computed by dividing the number of travellers on the trait joining two stops for the capacity of the route (deriving from the number of transit per hour times the capacity of the bus/tram/tube car).
$duration_{min}$ is the minimum duration through the paths between O and D.

The first factor rewards the choice of the shortest path, the second one resembles the exponential saturation of a capacitor shifted and rotated, it goes rapidly to zero as the saturation approaches the unity.



Time - dependent factor of the score function



Saturation - dependent factor of the score function

5.3 Inspectors movement

In this section the network structure of the infrastructure emerges, although it is not explicitly coded in an edge list or an adjacency matrix.
Inspector agents are responsible for the verification of the ticket of traveller agents, by choosing the link with the highest score connecting its stop with an adjacent one. The score can be computed in two different fashions:

- Counting how many people flow through an edge in their path. This score assignation supports the idea according to which ticket inspection instructs users to buy the users to buy the ticket by increasing their perceived risk;
- Rewarding the edges with a high score when users are fined while passing them. This modelling choice gives credit to the economic advantage of fining irregular users.
- Of course, a superposition of the previous approaches gives relative credit to every point of view.

We chose to initially score the links with the first approach.
Each ticket *inspector* saves the list of controlled stops, each with their respective itineraries. This list is then read by each traveller who decides whether to buy a ticket or not, based on the risk of being inspected.

5.4 Risk evaluation and ticket decision

The decision to make or not the ticket is taken independently by every agent according to its own experience. The algorithm to perform this task has to be computationally light (it has to run on hundreds of thousands of agents) and learn continuously with new experiencing while giving less relevance to outdated information (online learning).

We modelled the evaluation of the risk for the travellers as a series of linear regressions that combine and compare with a threshold.
The coordinates on which the learning problem takes place are the geographic location of the bus
stops and the time, for every link constituting the path

$$\vec{x} = (x, y, hour)$$

The linear regressor combines linearly this input with a weight vector in order to predict the risk, that is the probability of meeting an inspector on the considered edge.

$$ris_{edge} = \vec{w} \cdot \vec{x}$$

The probability of meeting at least one inspector during the path is the complement to one of the probability of meeting no inspectors, which is the joint probability of the atomic event "not meeting an inspector on the edge". In formulas:

$$P(no\ vet) = 1 - P(vets \geq 1) = 1 - \prod_{edges} (1 - risk)$$

The decision to buy or not the ticket is based on the comparison of this risk to a threshold (also subject to learning). If the calculated risk exceeds the threshold, the agent buys the ticket.

Weight and threshold update

The update rule for the weight follow the Perceptron Learning Algorithm (PLA), with some substantial differences. In PLA, the dataset is static, and the algorithm iterates over the points until it reaches a good set of weights. The convergence of PLA requires the dataset to be linearly separable; but the convergence concept itself changes in an online learning problem, and the linear separability of the dataset is not required.
The update rule is performed on every link of the path, by comparing the classification of the point of the linear regressor (classification is obtained by applying a threshold of 0.5 to the regression) with the boolean (coded with -1,+1 values to ensure the effectiveness of the learning algorithm) event "meeting an inspector". The update rule is computed on this error:

$$error = outcome - \Theta(\vec{w} \cdot \vec{x})$$

$$\vec{w} = \vec{w} + \vec{x} \cdot error \cdot \eta$$

We assigned different learning rates (η) to different events, weighting the encounter of an inspector much more than the complementary event; the ratio given by the ratio between the fine and the cost of a ticket.

To give fines ulterior importance, we introduced an effect of fines on the threshold after which the final choice of buying a ticket is taken. Every received fine lowers the threshold by a quantity, increasing the probability of buying the ticket in the future. This effect is only temporary since the threshold returns to its initial value with an exponential fashion.

6. Further development

The next steps will move in two directions:

1. Refine the model, by increasing the intelligence of traveller agents and adding more strategy and a learning model also for the inspector agents;
2. Perform experiments on the model, investigating the effectiveness or randomization of controls and comparing the "fine oriented scoring" of edges with the "visibility score"

7. Conclusions

Our work is not concluded so conclusions cannot be definitive. Although, we believe that our work marks good results in modelling a realistic system of motion on a network of stops, creating a mobility pattern that allows to build interesting dynamics on the top of it. The system under study is huge, it contains hundreds of thousands of intelligent agents performing complicated operations with large amounts of data. Our means are limited and every programming step has been taken by considering the computational cost of its realization; we hope to be able to run our simulations on powerful calculators in order to enhance our results production. We believe that we will be capable of making quantitative statements on the best strategies to adopt to fight evasion and will provide a valid model for testing policies in silico.

8. Acknowledgments

We want to thank the members of the institutions that guided us in the modelling and gave us the data we used. In sparse order:

- GTT, in particular Mr. Savarino, Mr. Rabino, Mr. Fantini, and Mrs Aracu, who helped us giving their first-hand knowledge of the problem of evasion;
- SiTI, in particular Mr. Inaudi and Mr. Arnone for the guidance in our modelling, especially for the construction of the network;

- Piedmont Mobility Agency, in particular Mr. Bason, who gave us the data we used in modelling the flow of passengers on the network.

We would also like to thank Mr. Carrara for introducing us to the meetings that allowed us to start this project.
A heartfelt thanks to professor Pietro Terna, for introducing us to the principles of agent-based models, guiding us in every step of the modelling, and being extremely active in responding to our needs and coordinating meetings with the aforementioned organizations.

References:

1. Piedmont Mobility Agency. *Imq 2008*. (2008).
2. Levine, R. V. & Norenzayan, a. The Pace of Life in 31 Countries. *J. Cross. Cult. Psychol.* **30,** 178–205 (1999).
3. Noekel, K., Viti, F., Rodriguez, A. & Hernandez, S. *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*. **1,** (2016).