TOWARDS DATA-DRIVEN SIMULATIONS IN URBAN MOBILITY ANALYTICS

Frederic Roulland, Cesar De Souza, Luis Ulloa,
Arturo Mondragon, Michael Niemaz, Victor Ciriza
Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
+33 4 76 61 50 50
{ frederic.roulland, cesar.desouza, luis.ulloa, arturo.mondragon,
michael.niemaz, victor.ciriza }@xrce.xerox.com

ABSTRACT

In this work, we present recent advances on the creation of data-driven models to address the needs of transportation planners during the conception, understanding and maintenance of transport networks. More specifically, we present data-driven models to understand and analyse mobility and to simulate and predict the impact of changes in existing networks. This new modelling approach leverages the massive amount of data collected in the field from daily users’ transactions and sensors’ outputs, and proposes to use in a more extensive way machine learning techniques that have emerged over the last decade. First, we present how this new transportation modelling approach differs from traditional practices. We then illustrate this approach through some specific use cases where it has been applied and present the preliminary results we have obtained. We finally end up with a discussion highlighting both the main advantages and the high potential of adopting such an approach in the transportation planning domain and also the main obstacles to be overcome before a large adoption can happen.

Keywords: data-driven modelling, transport system simulation, machine learning, public transport, parking

INTRODUCTION

Improving transportation in urban areas is a constant challenge for any transportation authority. However, developing transportation infrastructure is a costly and heavy work. Like any other business, transportation networks have to be operated efficiently, with offered services and associated prices that actually match the current user demand. In order to setup those offers, it is therefore imperative to carefully study the mobility needs of the target area, as well as to consider multiple different alternatives—preferably through comparatively inexpensive simulation mechanisms—before a decision regarding an infrastructure change can be firmly taken.

As of today, transportation infrastructure planning has not yet fully embraced the ‘Big Data’ era. Traditional business decisions are typically made upon intuition, mostly relying on an experts’ understanding of the customer’s needs and the state of operations. Furthermore, simulations supporting business cases rely on parameters that are setup in an ad-hoc way by these experts. For the most sophisticated cases, such as when building new infrastructure items such as train lines or roads, casting a correct decision involves simulating several
options based on a custom study. This study generally consists of a period of manual data collection and surveys which are then aggregated into a standard city model by an expert analyst. This approach is expensive and has disadvantages such as reliance on partial data (only few thousands of people surveyed and often with poor time coverage) and/or outdated information (e.g. global census of 5 years ago). Furthermore, mapping this data into a model is such an enormous work that data are often aggregated into a high level of granularity, allowing only a macro-level simulation of the mobility dynamics within a city. Although tailored, often same standard models are assumed to hold for diverse locations such as Stockholm and Mexico City. We therefore believe that the cost/precision ratio of simulations built from these custom studies is not optimal.

On the other hand, the presence of more and more intelligent transportations systems on the field allows for the collection of millions of transactions every day. Coupled with machine learning techniques and stochastic models, these data can be used to propose a data-driven approach to urban mobility modeling where city models are learned automatically from the collected data. These models can then be applied within simulation systems with a much higher precision, increased reliability for the predicted impact of a change, and with a much reduced cost of implementation to achieve such level of precision.

In this paper, we show how data-driven approaches lead to better backed-up decisions in example problems related to transportation applications. More specifically, we present two use cases where we have applied such approaches. The first one proposes a micro-simulation of a public transit system where the demand and the user choices models are learnt from fare collection data. The second addresses the problems of understanding and modelling parking lot customers on private-owned parking lots. With these examples we show how those techniques can be used to drive business decisions on much more justifiable grounds, and how the experience on designing data-driven models leads to a virtuous circle of continuous process understanding and process improvement. These are however only preliminary results of a larger potential of these approaches to change the practices of transportation planning. As such we conclude this paper mentioning what are still the challenges that needs to be addressed and what could be a more global outcome of the generalization of such approaches.

**POSITIONING WITHIN EXISTING APPROACHES**

In this section we present how this work is positioned within the space of modeling and simulating transportation systems. We first briefly recall the foundational elements of the traditional modeling and data driven modeling approaches. We then introduce transport simulation systems and how they rely on some models. We finally describe previous works related to the simulation of our two use cases: usage of public transport network and parking lots.

**Traditional modelling**

Traditional models are specified by specialists or a system architect well familiar with the process in question. Often, if a specialist is not available or is not applicable for the task in hand, traditional modeling techniques involve the conduction of customer surveys, which are tedious, cumbersome, potentially expensive, and which can also possible lead to user experience degradation if customers are always bothered to help the business on a timely fashion.
In public transport applications, one example of a traditional model approach can be given by the manual specification of customer choices when taking public transport. One could manually enumerate and specify conditions under when a particular customer would or would not take public transport by following a finite set of actions – an algorithm.

![Traditional Modeling Workflow](image1)

**Figure 1: traditional modeling workflow**

**Data-driven modelling**

Specifying those models inquire some need for knowledge representation. Data-driven models can be inferred (or learned) from big amounts of data. This idea is indeed appealing as it decreases the importance of surveys, which might need to be done less often (although it would be risky to assume they could have been totally unnecessary afterwards).

![Data Driven Modeling Workflow](image2)

**Figure 2: data driven modeling workflow**

Three of the main tasks in machine learning are: supervised classification, unsupervised classification, and regression. The trick is to be able to cast our business decision problem or our modelling approach as the answer of one of those three tasks. Afterwards, the process may resume to the task of gathering enough data to learn by picking up one of the diverse models existent for each task.

In a supervised learning problem, the goal is to, given a new input sample \( x \), be able to create a function \( f(x) \) such that the difference between expected outputs values \( y \) associated with each \( x \) and the output of \( f(x) \) diverges as less as possible. When the system being modelled outputs a finite set of discrete values, such as label numbers (e.g. 1, 2, and 3) then this problem can be seen as a supervised classification problem. When the system outputs continuous values, then this problem can be understood as a regression problem. Furthermore, when we cannot observe the system’s outputs for the labels, then we arrive at an unsupervised classification problem – where the task is to be able to group data into smaller subsets of values sharing some common characteristics between themselves.

The existence of comprehensive libraries for diverse platforms makes it straightforward to assess multiple machine learning models.
Transportation system usage simulation

A transportation system simulation can be described at high level like in

![Diagram](image)

Figure 3: Transportation system simulation overview

The demand model describes different transportation needs for people within a city. In the case of public transport or road networks it is generally represented as an ODT matrix where each row represents a user travel needs with three dimensions: Origin of the trip, destination of the trip, and time (either of departure or arrival). In the case of parking lot system this will be typically represented by a bi-temporal matrix where each row represents the stay of car either with the time of arrival/ time of leave or time of arrival/duration of stay. However, for simplicity or lack of information, demand is often only modelled in a more aggregated way with flows between origins and destinations zones over a period of time, and parking occupancy level over time. It is also important to note that more advanced modeling of demand may include the notion of profile associated to each user, including for example demographic information or any information useful to distinguish different segments of the user population. On the same way, a specific demand model can be attached to a specific context such as weather condition, day of week, etc, which is likely to influence the characteristics of the demand.

Once a demand model is available, the assignment model defines how these travel needs will be allocated within a given transportation infrastructure. A large variety of assignment models and approaches exist in the literature ranging from macro to micro level modelling. In macro simulation models, people are seen like a flow and each element (e.g. road, parking lot) of the transportation network is assigned a flow which corresponds to the sum of the percentages of the incoming upstream flows to this element. In a micro simulation model, each user is represented as an individual agent and what is being modelled is the choice that these users are going to make at each time they are confronted with options within the transportation network.

Data driven model approaches can be used both for modelling the demand and the assignment of this demand. The demand can be typically reconstructed from the transactional data. In
addition to the data collected by the operations, some more elements can be inferred using machine learning techniques. For example, user profiles can be built using unsupervised classification or correlation between the demand, and external factors can be learned by multiple observations of this demand across different conditions. On the same way, assignment models can be learnt by supervised classification or regression algorithms that will define the value of the parameters of the user choices or flow assignment functions that best reproduce the traffic observed in the past.

**Simulation of a public transport system**

Although being much more complex to model, the simulation of public transport systems has been less studied than those of road traffic systems. It has often followed the same trends, with tentative to adopt some of the findings in the road traffic models to the more specific case of public transport. The history of such simulation can be roughly summarized in three steps, detailed below.

The first systems developed in the 80’s, also known as strategy based approaches [9], proposed a macroscopic modelling of the public transit where only lines and flows on these lines were considered. This generated a whole family of assignment models all based on the concept of equilibrium (Equilibrium Transport Assignment Model) where all flows are governed by the maximization of a single utility function defined by an expert. In these models the temporal dimension of the demand and the effect of connection time between vehicles are not considered, which made them realistic only for high frequency services with no overcrowding effects.

A second generation of simulations, known as scheduled based approaches, was developed in the 90’s. Its main difference from previous approaches it was to propose a mesoscopic modelling of public transit where not only lines but each vehicle on a line and their corresponding schedules were represented [10]. Whilst these models capture much better the dynamics of public transportation, the behaviours of vehicles and users were still governed by global utility function maximization. The capability of this utility function to model the actual complexity of phenomena observed in the reality is still very approximate. Such systems are however the most widely used in the current practices [4, 5] because of the fact that they need an achievable amount of effort to be configured and calibrated from manually collected data.

With the increasing computational capabilities, some micro-simulation approaches emerged in the 2000’ with activity based assignment models where each vehicle and traveller is represented as an individual agent that can make his/her own choice at any moment of decision [6, 7]. The user choice model allows the freedom to simulate very complex behaviours of users taking account many parameters and representing the full dynamics of the public transport system. However in order to be setup, such microscopic simulation needs very detailed data, up to the level of each individual origin and destination of travellers and are potentially difficult to calibrate. These models have not so far fully benefited from the usage of operations data and machine learning techniques, demand is often using census data and parameters of the user choice are calibrated in order to match with some partial manual counting at some location of the network. Their use has been often limited to small network’s sub-section simulation or research experiments.
Simulation of a parking lot system

There have been several attempts to simulate parking usage but always without a data driven model approach.

Highly specialized parametric approaches are the most common way to model parking user demand. The SUSTAPARK \([3]\) model is an agent-based system that can simulate parking and traffic situations under different parking management conditions in the context of an entire city. In this system, the number of agents operating inside the city was roughly estimated from a travel survey, and was further tuned by running the model several times after gradually reducing the number of agents at each time. The authors specified the parking demand for different places by considering different attraction factors per trip motive (recreation, work, shopping) for each kind of building (restaurant, residential or office). Afterwards, the calibration process was performed by asking field experts of the city administration to assess results based on their knowledge.

Another related work is given in \([8]\). In this study, the author has used MISIM (a microscopic traffic simulator developed at MIT) to simulate off-street parking and investigate how to create a user-choice model for this task. However, in such work, the author chose to define behaviour groups \(a\ priori\), dividing users into fixed, guided and unguided. Moreover, a manually-crafted algorithm is used to determine whether a user would chose to park or queue for a new option. In our system, the groups are determined automatically by an unsupervised algorithm. Furthermore, in our system the choice behaviours of each group are learned using a supervised algorithm from the historical data, instead of being manually and rigidly programmed into a computer.

The work detailed in \([1]\) proposes an approach for designing a parking search model based on utility maximization theory. However, this approach is centred towards on-street parking or parking lots spread throughout a city. It is neither based on learning the user behaviour nor on user profiles.

The work of \([11]\) developed a simulation model for parking systems. However, when confronted to the specific problem of simulating duration time (the time a car stays parked in the parking lot), the author chose to adopt an average duration time for all vehicles within a certain period.

Finally, some more prior work addresses the specific modelling of parking demand. Demand strategies include the regression model, which attempts to estimate parking demand as a parametric formula whose general form has to be specified manually or picked among candidates. One very famous source for parking demand models is \textit{Parking Generation}, a document prepared by the Institute of Transportation Engineers, now on its 4\textsuperscript{th} edition (2004). This publication presents parking generation rates, equations and data plots for parking areas in the United States. Those formulas are created and measured using distinct sources such as surveys and data reports by individuals and organizations.

Existing parking business applications do include some user profiling capacities. But they are often limited to predefined and well known user behaviours frequently observed in reality. For example, parking operators often assume a recurrent profile of users who arrive Friday afternoon and leave Sunday night. Those would correspond to people parking their car for a weekend, so parking operators can target these users’ by offering them special discount
packages or targeted advertising. The above case is when user profiles are known \textit{a priori}. Instead we expect that others, latent profiles could be discovered in some other way. Furthermore, when those new profiles are discovered by a clustering algorithm, there will still be the problem of their easy interpretation. The discovery process should be transparent to the human decision maker, it should help her understand why those users have been grouped together, and entail an appropriate offer or proposal.

User profiling using a clustering algorithm is not a novel idea, as this is, roughly, the objective of basically any clustering algorithm. However, the process of applying this technique in a transportation-related analysis system, where profiles can be created, edited and explored, plus where automatic tags and automatic textual descriptions can help the operator better understand his data was not found in current parking management systems. Most systems rely on the manual or static definition of a user profile, and don’t offer much flexibility around those definitions. Parking (or public transportation user) profiles are usually tuned using data surveys, which can be costly and tricky to be performed in an adequate and timely fashion.

\textbf{CASES STUDIES}

\textit{Business case: Learning public transport user choice and demand from fare collection data}

Planning new infrastructure, in particular for public transit, is one important dimension where transportation authorities can act. However development of public transport infrastructure is costly and heavy work for a city. Such decision required therefore a very careful study of the mobility needs and simulation of different alternatives. As we described in the previous section, there have been in the recent years a constant evolution toward more microscopic and precise simulation models for public transportation. However the use of this model is limited because of the growing cost it implies in order to collect and inject all the data required for such simulation.

We use fare collection data to build both the demand model and to learn optimal user choice and demand model parameters. Fare collection data systems provide the full list of validations made by passengers through their use of the transportation systems. In most of the transportation systems people will validate when they enter a vehicle (check-in). Few systems will also ask to validate when they leave the vehicle (check-in/checkout). In many systems regular users will use a transportation medium such as a smart card that will provide a unique identifier of the user for all the validations she performs. Fare collection data constitutes a source of information that covers the whole population of users and the whole history of interactions of these users with the transportation systems.

At the core of our system, there is a learning module that is able to model the city mobility without any human intervention just by using the data previously collected by the fare collection system. This module relies however heavily in two sub-components:

- \textbf{The passenger journey reconstruction module}, whose role is to reconstruct all the journeys for each individual public transit service user, based on validation logs of the cards and tickets used within the network. It relies on the work presented in [2] which is based on the assumption that travellers that follow a symmetrical pattern and use the
same ticket/card are good representatives of the demand. This module generates in
essence what will constitute, together with the schedules and lines description of the
current service, the training set for the learning module.

- **The micro-simulation module**, that is able to generate a set of traces that represents
all the simulated journeys of each individual public transit passenger for a given
period of time. It uses automatas simulating each individual traveller and each public
transit vehicle within the city network and their interaction. This micro-simulator takes
as input the schedules definition and the lines of the public transit service. This is used
to populate all the vehicles and to simulate their behaviours. This is used as an initial
point to learn the probability of schedule adherence. It also takes as an input the city
mobility model that is being learned by the learning module. This model is composed
basically of two parts. One part models the demand to be learned, i.e. the origins and
destinations at any time of the travellers. The micro-simulator uses this part to
populate all the travellers and their journeys in the simulation. The second part models
the travellers’ behaviour to be learned, i.e. what will drive the choice of a particular
path given the origin, the destination, the context and the profile of a traveller. This
user choice model relies heavily on a trip planner which has been designed
specifically for computing very quickly a large set of alternatives for a given origin
and destination at a given time. It is called during a simulation each time a traveller
automata needs to simulate the choice of a traveller starting a journey or having to
reroute his journey because of an unexpected change.

In training mode we use the demand model generated from the fare collection data of one past
period. The assignment model forces the public transportation vehicles to follow the exact
same behaviours that were observed for this period. The simulation outputs a set of trips with
simulated boarding and alighting events that can be compared directly with the individual
trips reconstructed from fare collection data when building the demand model. The system
iterates with different parameter values and ends once the simulation output and the original
reconstructed trip are judged to be close enough. Likewise passenger flows in vehicles can be
forced to learn vehicle’s parameters. In this phase the period used for building the demand
model may vary from one instance to the other in order to avoid finding parameters which are
only optimal for the training period and to allow cross validating results.

In prediction mode we build a demand model from a past period which is considered to be
representative of the demand for which we want to simulate a transit in a new transportation
infrastructure. This time, both vehicles and users behaviours are simulated. User choices are
based on the assignment model parameterized with the best values obtained in training phase.

So far very few approaches have attempted to use a micro-simulation in order to simulate an
entire public transit network. We believe this is mainly due to two constraints. The main one
is the labour cost that would be associated to collect all the data and to build manually the
mobility model of the city. The second one is associated to the computational complexity
of running the simulation. Whilst our approach of automated learning of the model solves the
first issue we have as well implemented custom algorithms to allow micro-simulation and trip
planning in a distributed environment.

**The system has been implemented and is currently tested with fare collection data
provided by the system deployed in the city of Nancy, France. This city operates a
network with usage in the order of 100 000 journeys per day. The results of the**
simulation can be analysed using the data mining tool developed on top of this fare collection system. For illustration purposes, we can show for example in Figure 4 what is the aggregated load of the public transit network over a day of activity in Nancy and what is, in contrast, the result of a simulation where we tested what would happen if the tram line acting as a backbone in the network would be removed. We can see, from the two different visualisations, how the removal of the line is partly reducing the overall use of the public transit and how in some places some parallel routes are experiencing major load increases.

Figure 4: Heatmaps for Nancy public transit network daily load (left) and load simulation without tram line (right)

Business case: identifying patterns in parking customers’ actions

The profiling solution presented here is part of Xerox’s Mobility Analytics Platform (MAP). This project is both a visualization engine for visualizing parking and traffic data; and also a framework for creating, testing and exploring data processing techniques in geo-spatial data with ease. The data available in MAP’s processing engine come from several sources and reflect different areas over the world. Some of them come from standalone parking lots, malls and airports; others come from networks of parking lots covering entire cities or agglomerations. Given the amount of available data, applying machine learning algorithms to such information helps us understand better the behaviour of parking users.

A major concern in business applications is that models should be interpretable where such understanding is achieved by identifying user groups presenting a similar behaviour. This similarity of user behaviour can be captured in many ways, such as similar recurrent arrival and leave weekdays, making similar journeys and choosing same parking zones. If user profile is identified correctly, making an appropriate offer or proposing a service will have higher chances to be accepted by the user. Furthermore, when those new profiles are discovered by a clustering algorithm, there will still be the problem of their easy interpretation. The discovery process should be transparent to the human decision maker, it should help her understand why those users have been grouped together, and entail an appropriate offer or proposal. We have therefore proposed a profiling solution which is semi-supervised so that it allows the system to discover patterns from the data and then propose an interface for an expert to interpret the results.

The first step to explore different user profiles in a parking database is to attempt to let the system automatically extract some profiles from the data. For this, the user is presented to the Profile Explorer interface (see Figure 5), where it is possible to create, remove, edit and publish user profiles. As in most machine learning tasks, profiling user activities requires
extracting features that describe the parking actions and processes in the most complete way. Feature extraction step can be customized by the user, and depending on the user’s choice, features can be included or excluded before a clustering algorithm runs with the data. This reflected in the Profile Explorer interface where the user can select to create a new clustering algorithm: either letting the system guess initial parameters for his data; or by manually overriding those parameters and attempting manual settings.

After, profiling results can be decomposed and visualized by different grouping criteria, such as profile, day of the week, and day of the week by profile. Different characteristics can be readily seen in the profile editor, such as number of parking events, generated revenue and arrival times. Profiles can also be labelled or renamed manually if necessary.

![Figure 5. Interface for exploring different aspects of the clustered data.](image)

Afterwards, the multiple clustering views and user profiles can be permanently stored to the database so they can be seen in a geographical / parking data analyser, such as MAP. The profiles can be integrated and individually seen in any of the previously existing visualizations, such as MAP’s Average Occupancy view.

**Business case: understanding and learning how off-street parking customers’ behave**

In this use case, we address off-street car parking business applications, where modelling customer behaviours is achieved by first identifying user groups presenting a similar behaviour like if it was described in the previous use case. If user profile is identified correctly, making an appropriate offer or proposing a service will have higher chances to be accepted by the user.

We created a parking lot simulation system to aid in this task. As other simulators, this system attempts to simulate different user behaviours by assuming some standard groups of users will behave similarly. However, in this system the groups of users, henceforth user profiles, are learned automatically rather than specified by hand. Those profiles are then used as a key element for automatically learning the decision function of parking users, automatically learning one decision function per profile.
Furthermore, this simulation system works by casting the user-decision problem as a machine learning problem. Such approach enables us to extract inter-zone price elasticity estimations right out from the learning models, as well as detect which features a certain user profile may be more concerned of, which may be crucial in better determining how to tune and optimize both service and revenue for different segments of the parking lot consumers. No previous study or patent had been found detailing such an approach for parking simulation, specifically presenting the same proposed data workflow and system architecture. Given the amount of available data, applying learning algorithms to such information helps us understand better the behaviour of parking users.

The system works closely with the Xerox’s Mobility Analytics Platform (MAP) analysis tool, being able to produce a simulated parking environment that can be visualized and explored through the standard dashboard interface in a web-browser. Besides the overall framework, this system also presents a few user interfaces which can be used to edit price grids, invoke and test simulations and produce basic outcome reports comparing the simulated world against the real-world.

The application framework is divided into the following modules:

- The Database module;
- The User Demand module;
- The User Choice module;
- The Price Engine module;
- The Simulation orchestrator alongside its virtual parking model state.

The database contains historical data that is used to build the user demand model and provide the basic infrastructure information for the simulator to build a virtual parking. A virtual parking is a model of the real parking presenting the same real-world constraints on the capacity of each parking zone and their distances from the central objective building of the parking (i.e. the airport).

The user demand model is an entity that can guess when a new user is going to arrive in the parking lot. It is built from historical records about previous users of the parking, and, in the case of airports, should also be able to use information about the current departing and arriving flights. Afterwards, the user demand module can send this information to the simulator. For each predicted arrival and duration times obtained from the demand model, the simulator will ask the current choice model to guess to which zone this user might want to go.

The user choice model is using a machine learning model to learn the user behaviour. One component that certainly will be part of most user choice models is the price engine, which computes how much the user would have to pay in case he opts for a particular tariff choice.

The usefulness of the described techniques has been tested in the context of an airport parking lot. Currently, the database encompasses data from the parking systems from the Blagnac-Toulouse Airport and flight data from this same place.

The Toulouse-Blagnac Airport is located in southern France and is the 5th airport in the country, having served in 2012 around 7,559,350 passengers. The airport has 7 public and 2 reserved parking zones which are able to accommodate roughly 9000 cars.
The available data corresponds to the period of February 2th to July 24th, 2013. In all of our experiments, the data had been further divided into two disjoint subsets, reflecting past and future behaviour of the parking lot. A near 4 month period from 02/02 to 24/06 has been used for training and calibrating our system; whereas the last month from 24/06 to 24/07 has been used to validate our findings. In all experiments, the objective was to simulate user’s behaviour in this last month (generating new users and using the automatic models to cast their decision) and compare how far the obtained results were from the real world.

The simulations have been created assuming a 7-day (1 week) granularity. The system also supports periods of one month or a year. Please note that utilizing periods of one year would be ideal to capture seasonal effects; however, not enough data was available to accomplish this feat. Each simulation ran for a period of at least 24 weeks until results could be gathered. Each experiment has been repeated a total of 25 times, and results have been averaged.

The idea to take user profiles as the base for learning specialized user choice models plays a fundamental role in this system. Table 1 shows how different aspects of the parking data are affected as inspected using various visualizations.

One interesting characteristics of the simulation done without considering the profiles is the presence of artefacts in the qualitative data, which are particularly visible in the day of the week duration visualization for parking zones P5 and P6 (second row of table 4). As it can be seen, the data not using the profiles shows an inexistent strong band characterizing short-duration stays for those zones, which are clearly not the case in the real world and are not present in the simulation using the profiles.
Table 1. Qualitative comparisons between real-world and simulation

<table>
<thead>
<tr>
<th></th>
<th><strong>Real World</strong></th>
<th><strong>Profile-based simulation</strong></th>
<th><strong>Direct simulation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day of the week duration, by day, P1 P2, 95% filter</strong></td>
<td>![Graph 1]</td>
<td>![Graph 2]</td>
<td>![Graph 3]</td>
</tr>
<tr>
<td><strong>Day of the week duration, by day, P5 P6, 95% filter</strong></td>
<td>![Graph 4]</td>
<td>![Graph 5]</td>
<td>![Graph 6]</td>
</tr>
<tr>
<td><strong>Time Period Revenue, 24h, P1 P2</strong></td>
<td>![Graph 7]</td>
<td>![Graph 8]</td>
<td>![Graph 9]</td>
</tr>
<tr>
<td><strong>Time period revenue, 24h, P5 P6</strong></td>
<td>![Graph 10]</td>
<td>![Graph 11]</td>
<td>![Graph 12]</td>
</tr>
</tbody>
</table>
ADVANTAGES AND REMAINING CHALLENGES

Through these few use cases we can start to see how a data driven modelling approach can lead to much more efficient practices for planning and understanding mobility. As such, we can try to summarise the main advantages they bring to the state of the art in following paragraphs.

Advantages

The biggest advantage is maybe that these method do not need any manual data collection phase in order to be able to produce an analysis or a ‘what if’ study. Data gathering represents about one third of the budget of planning departments of transportation authorities. In addition to a potentially high cost reduction this also helps to have a much faster process. A data collection campaign may last in the order of a year or more depending on its scale.

Another important advantage is the fact that since transactional data are collected daily, the models can benefit from the most up-to-date representation of the demand. Similarly user behaviours can be calibrated dynamically taking into consideration the most recent evolutions in change of behaviour. This is a big advantage in order to have a more accurate study output.

Similarly, for all regular users we have the history of their use of the transportation systems compared to a survey which constitutes a snapshot of mobility. This information can be used in order to have more sophisticated model and here again increase the accuracy of the model output. It should be noted that the number of identified users with a payment history increases as payment technologies become cheaper and integrated into existing mediums (such as smart phones or credit cards).

Finally, another advantage is that transactional data covers the whole population of transport systems users. This is particularly important for the training phase. In state of the art approaches calibration of the parameters is done using a very limited amount of samples with a high risk of overfitting issues. Whilst overfitting of our model is still a risk, the fact that we have complete coverage allows for decomposition and cross-validation on time, space and population which reduces a lot the uncertainty of generalisation and again has some positive effect on the simulation accuracy.

When applied in a more holistic way to a whole urban area level, combining data from all the different transport modalities with external data which are correlated with mobility patterns, we will be able to produce models which reconstruct the whole mobility experience of people and all the observed dynamics at the level of the city. These models will enable to continuously understand the city and to accurately predict the effect of changes. However, what we have shown so far are just isolated examples of applications which are isolated from each other and not yet fully deployed in the field. There are still several barriers to the massive adoption of these new technologies in the domain.

Disadvantages

The first one is maybe the relatively newness of the technologies. The big data technologies emerged in the 2000’ mainly within the internet business and are led by the e-commerce and social media companies. Although it has started to generate a lot of discussions in the Intelligent Transportation System community, there are so far only few concrete applications and they require often still a lot of research and hardening before reaching their potential
outcome. Therefore the technology risks are still high in this immature market and the adoption in the field may stay very low until there is enough pilots which generate well documented outcome and convincing results.

In our view, the main obstacle to a massive adoption of data driven planning approaches is that despite the existence of massive amount of data it is very often not easily accessible and stored into silos with very specific needs and constraints. Two different evolutions are required in order to unleash a more integrated use of this data. A normalization and standardisation of the way data are stored and exchange will simplify the interoperability between the operation systems and the planning systems and will facilitate the integration of data coming from different sources such as road traffic, public transit, parking, road tolling, connected vehicles and even external to the transport such as demographics, social networks and mobile phones. At the organizational level, some agreement must happen between the different entities owning these data and maybe some organizations must be created to regulate the exchange of this data in order to guarantee that the business sensitivity of some of these data and the users’ privacy is properly protected. We believe these two evolutions will happen with the increasing common interest raised by the potential applications. The pace of this evolution, however, may be not as fast as the technology progress.

Finally, assuming the two first obstacles are removed, the adoption of these new technologies will as well require some extensive training of the transport planning practitioners so that they can take the full benefit out of them. Although the approaches propose to automatically learn from data, the user of his technology are still require to understand the main principle of it in order to be able to train and configure them and to be able to interpret the results of such model. This constitute a paradigm shift for the community where we will have to put next to the classical expertise and understanding of the transportation phenomena a significant amount of expertise in predictive data analytics and machine learning principles.

**CONCLUSION**

In this paper, we have tried to describe the emergence of a new approach required in order to make a more automated use of the mass of data collected in the transportation operations for understanding mobility needs and simulating effect of changes. Although we have described our own experience in applying such approach in different scenarios of application around parking and public transport, the aim was not to describe all the details of the implementation of such approach but rather to show at a conceptual level what are the types of problems that can be tackled with it, the advantages it provides and also what it requires for a more global adoption.

Although it may take some time, we believe this will change quite radically the way transportation systems are being planned. Our first next step will be to continue our ongoing experiments in order two build some more assessment of the output of these models. We will then work in parallel to define some data driven models for other transportation system type and to facilitate the integration and accessibility of the operations data for such model. We believe these two combined efforts are needed to move beyond a global modelling and understanding of urban mobility.
REFERENCES


