Improving public transport in Valencia

Alejandro Méndez[1] and Kacper Moll[1†]

Contributing authors: almencar@inf.upv.es;
kacper.moll@gmail.com;
†These authors contributed equally to this work.

Abstract

Public transportation has a fundamental impact on the city's standard of living. For this reason, great importance is attached to improving the transportation system. However, solving the problem does not mean buying new vehicles. A key procedure to optimise the transport model is a comprehensive analysis of the existing system and implementation of the IoT devices. Our study focuses on the transportation issue in Valencia. Our goal is to optimize the transportation model without generating additional costs, that is, using existing infrastructure. In that paper, we proposed a data collection system so that the transport model can be optimized regularly to make it flexible.

Keywords: public transport, transport system improvements, metro

1 Introduction

Public transport is one of the most important factors in a Smart City. It affects the city in multiple dimensions, such as the environment or the economy. A growing population in big cities as well as demographic changes in peripheral areas have become a challenge for public transport to ensure a quality service for all citizens, which makes maintaining and improving the public transport system an imperative.

The current public transport system of Valencia makes it ideal to exploit its flexibility, which is a promising approach to improve the efficiency and performance of passenger transportation. Flexible transport systems provide passengers with flexibility in choosing routes, times, modes of transport, service provider or payment systems. The flexibility is based on the integration of different modes of transport, or possibly spanning multiple service providers.

This concept is especially suitable within suburban areas, which usually suffer from a lack of service availability and demand uncertainties.

With the advances in technology in the last decade, numerous countries have started investing more money in the modernisation and improvement of their public transport system, encouraging the development of a Smart Public Transport system. Automated Fare Collection has gained popularity worldwide, as it allows public transport managers to collect high-quality data via the transit smart cards system [1]. This data has various applications, such as origin-destination estimation, route choice modelling, transit performance indicators or trip purpose detection, among others [2].

In this paper we are going to analyse the performance of the public transport in Valencia, using smart card data. We will analyse how public transport is used nowadays in Valencia and the implications of its current design in the transit flow of the city, and conclude possible improvements to the actual system.

The rest of the paper is structured as follows: In Section 2 the existing work related to our project is exposed, and Section 3 concerns the data collection system. Sections 4 and 5 are focused to explain the methods used in this paper and the results obtained, respectively. A summary of our work can be found in Section 5, dedicated to the discussion of the results obtained, and finally, future work that could be a continuation of this paper and some conclusions are proposed.

2 Related work

Problems in public transport have existed since its beginning. There is a lot of investigation by scientists trying to improve its efficiency in some way. In the study [3] the authors focus on finding the relationship between population density and public transport accessibility. Through multivariate logistic regression analysis, the authors concluded that as urban population density and public transport availability increased, commuters were less likely to drive and more likely to use public transport. They also noted that drivers could be encouraged to use more sustainable transport if it were more accessible.

Integration of all transport nodes is the key to a successful transport plan. The introduction of new vehicles to the public transport network involves a reorganisation of the pre-existing network. Such change was described in a paper about public transport integration on the example of Thessaloniki [4]. The authors describe the reconstruction of bus lines concerning the underground project. In the end, they proposed a ready-made public transport system with clear roles and a structured hierarchy of all transport nodes. In the new system, parallel routes are avoided for the most part, which extends public transport to further districts. A similar mechanism must be applied in Valencia, where public transport has recently been supplemented by alternative vehicles (electric scooters, scooters and valenbisi). Rebuilding the system in the city centre could allow transport to be improved in the surrounding areas.

There would be no way to improve transportation without a deep analysis of it. Some projects around the world make such an analysis. One of them is the Moovit application. According to a paper about mobility as a service and public transport [5], Moovit attempts to help governments and public transport agencies with solutions for reducing congestion, and shifting car drivers to public transport and shared mobility. In our research, we should make a similar analysis for Valencia.

3 Framework

To carry out this project, two different sources of data were needed: the geographical location of the subway and bus stations, as well as their schedule; and transit and usage data of the public transport system.

The data collected are in no way linked to personal data. Indeed, the only information needed is the total number of passengers using a given transport at a given time.

This data can be obtained through smart cards, which are already in use in Valencia. However, only smart cards are not enough. To create a complete data collection system, Automated Fare Collectors in the subway or bus stations are also needed. The AFC system should be able to record the station entry and station exit of each smart card. This result can be easily achieved because each AFC device will have its location.

Since nearby stations will have the ability to share data, it will be possible to predict traffic congestion at each station, as well as in each vehicle. Such an estimate would be useful to both passengers and traffic managers. For instance, if such information were displayed at stations, a passenger could decide to take another mode of transportation to avoid crowding. On the other hand, people who manage the system would get the alert to increase the number of vehicles. To make this immediate response a reality, high costs are required to maintain a flexible transport system. While this is possible, however, our project focuses on passive improvements.

Since the data transfer and computation between stations is done by using fog in real-time. That already compiled data can be uploaded then to the cloud. As a result in the long term, the data collected in the cloud will provide information on where changes need to be made.

The following picture presents the system described above.

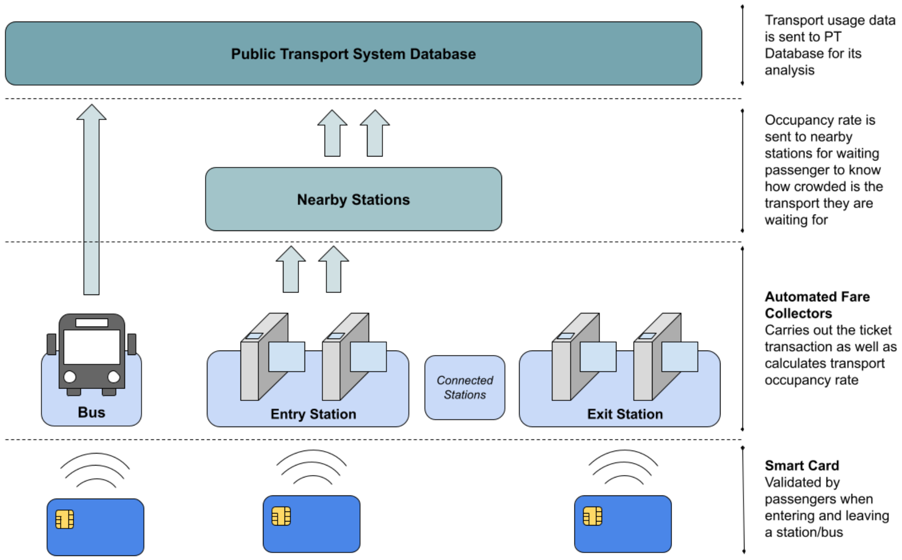

Fig. 1 Framework diagram

4 Methods

Many variables have an impact on the final public transportation system. For this reason, different types of data are required to achieve the study's intended goals of reducing waiting time, which is closely related to reducing commuting time and encouraging people to travel by public transportation. However, some data are necessary to perform sufficient analysis of the public transportation system, these data include:

- Timetable/schedule
- The location of each station
- The number of people using a given transport at a given time

The schedule and timetable, as well as the station locations, are publicly available, but not in a format we can work with. For this reason, web scraping techniques were employed to extract this data in a format useful for analysis. It is not possible to make improvements to a system without past data on the number of people using that system. Since the Valencia government does not provide passenger flow data as open data, the data has been simulated and experiments will be performed on the data so generated. Although the results will not allow us to draw meaningful conclusions, we will try to develop our methods in such a way that they can be repeated on real data. Data simulation, which is a trip simulation, was performed according to the system whose operation is described in the "Framework" section. As a result, after the implementation of the proposed system, the analysis that will be performed on the generated data will also be able to be performed on real data in the future. Due to the lack of access to actual data, the data generated is only for the metro Valencia transportation network to ensure transparency of the research. The simulation was conducted using Python as the programming language, and public transit journeys were generated using stochastic models, and simple estimates of trip metadata such as waiting time, travel time etc. Below is a general demonstration of how the trip generation algorithm works:

1. Two stations were chosen, with the more busy stations having a larger probability of being picked up.
2. Waiting time and travel time are estimated, using a probabilistic approach based on the distance from the station to the city centre.

The number of passengers each hour is estimated using a normal distribution, varying the mean and standard deviation depending if the hour is a peak hour or not. Weekdays and weekends are also taken into account in the simulation.

5 Results

After generating the travel data, an exploratory analysis of this data was performed. Although analysing generated data does not seem useful, may this serve as a proposal of what can be done with real data, where the results obtained may be much more fruitful.

First, an analysis of the travels per hour was performed. Fig. 2 shows the travels registered each hour for the first 7 days (where day 1 corresponds to a Monday). As it can be seen, this is a stationary process, with a repetitive pattern each day of the week, except at weekends, when the travels decrease as fewer people use the subway to commute to work.

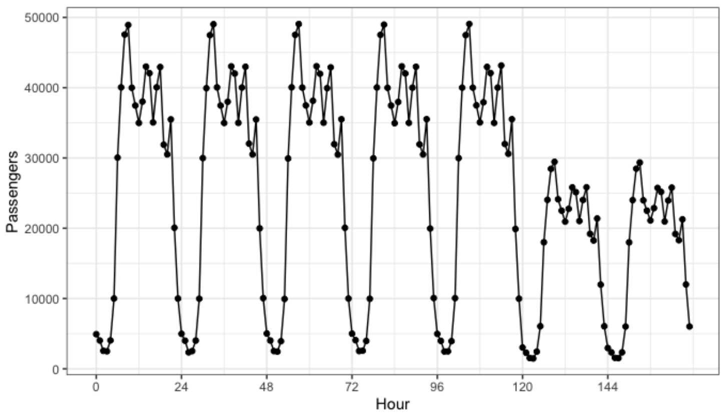

Fig. 2 Travels per hour

To investigate the impact of distance from the city centre on travel, a similar study was conducted grouping the source stations into five different categories, according to their distance to the city centre. Fig. 3 shows the results obtained, where can be clearly seen that the nearest stations to the city are the ones receiving the greater number of passengers. All groups are balanced (except the "< 1" group), so the number of passengers is comparable between them.

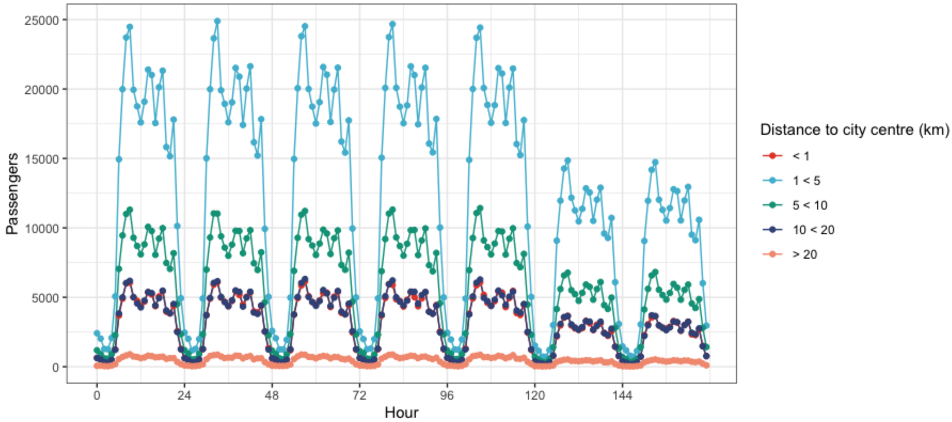

Fig. 3 Travels per hour, divided by groups

Another important milestone would be studying the sections of the subway that tend to be the most used by the passengers. Fig. 4 shows a map with the occupancy rate of the sections in the central area of the city. As it is expected, the most central lines are the ones with the highest occupancy rates. This makes sense, not only because the central stations are the busiest ones, but also because the radial infrastructure of the subway makes it almost impossible to avoid these sections when travelling from one station to another, especially when these stations are in different areas of the city.

This problem could be solved by using ring roads, already implemented in other transportation systems, like the bus, so it makes more important the merging between both public transport modes.

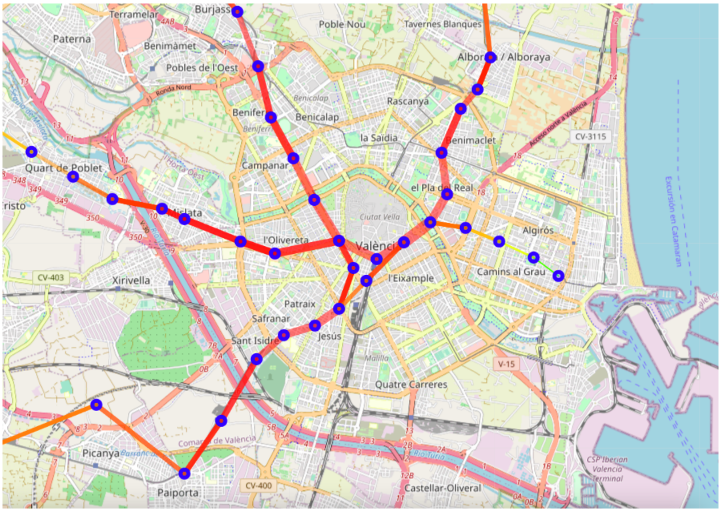

Fig. 4 Occupancy per sections

Another important feature that would be important to analyse is the waiting time. In this case, waiting time was generated, although it is far from realistic as the waiting times are very complex to simulate. In real life, there would be two types of passengers that can be depicted by this example: a student that leaves home knowing exactly when the subway arrives at the station, so he will not have to wait too much in the station; but then, in his way back home, he arrives at the station whenever his classes end, so the time he has to wait in the station is undetermined. The latter one is the interesting case of study: knowing how much a passenger has to wait in a station when arriving at an arbitrary hour, which is highly correlated with the number of transports in the line.

After making a deep exploratory analysis, we wanted to apply machine learning models to predict key metrics for developing a flexible transportation model. Certainly one of those metrics is the travel time from station A to station B. It is already proved [6] [7] that Supporting Vector Regression (SVR) is a suitable model to predict travel time. Since it has higher than other models generalization ability and guarantees global minima for the training data. That is why by implementing that model we expected it to perform well in time series analysis. To apply the SVR model, we took the following steps:

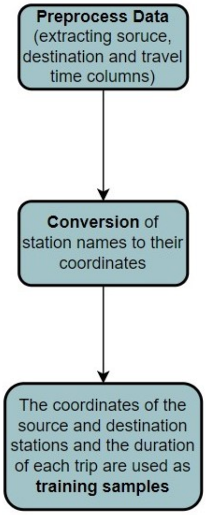

Fig. 5 Implementation of SVR

Nevertheless, after estimating the validation of the model using cross-validation with 5 partitions, the results were not very good. The MSE was of almost 15 minutes, which is far from optimal. Other models, such as Stochastic Gradient Descent or Multilayer Perceptron were also tested, but the results did not improve. Probably, this was caused by the unbalanced nature of the data: the farthest stations were the ones with fewer travels, so the data had a lot of short time travels but just a few long travels, predicting a travel time for all travels much smaller in some cases than expected.

In real life, the subway time travels are the easiest ones to predict, as they are not normally affected by traffic or climate. Other transportation modes are more complex to predict, as they are affected by these factors. Good use of AFC and Smart Card data would be to implement a travel time predictor that recommends to passengers the best route combining all public transport modes available in the city. That is why in the model developed in this paper geographical coordinates were used as predictor values.

Lastly, with the data generated, a prototype of an application for occupancy control was developed. This application shows the occupancy levels for each of the sections of the subway infrastructure. This kind of application could be useful to both public transport managers and citizens who would like to use public transport. The prototype is available on this link. The application shows the stations displayed on a map, as well as the lines, where their width and colour are characterized by the occupancy level of the section. Fig. X shows a screenshot of the application. In a real-world scenario, some additional features should be added to the application in order to obtain reliable estimates of the occupancy levels, as they can only be exactly measured *a posteriori*. For example, the estimations could be obtained via an ARIMA model using the travel time series, or an Origin-Estimation prediction could be performed the moment a passenger arrives at a station, obtaining the sections the passenger is going to need to reach the predicted destination. In the latter case, a probabilistic data structure, such as Count-Min Sketch, could be used in order to reduce the amount of memory needed and be able to compute these predictions in memory-restrained devices.

6 Conclusion and future work

In this paper, we developed a proposal for data analysis and prediction using a simulated dataset of travels using the subway of Valencia. An exploratory analysis was conducted, and a predictive model using Support Vector Machine was performed, obtaining poor results in the latter due to the nature of the data used to train the model. Lastly, a prototype for an application for occupancy control was described.

This work had some major limitations caused by the unavailability of real data, and the consequent use of simulated data. Future work could be performed by implementing some of the proposals described in this paper with the actual Metrovalencia data, as the hardware infrastructure for obtaining Automated Fare Collection and Smart Card data is already deployed.

References

[1] Ma, X.-l., Wang, Y.-h., Chen, F., Liu, J.-f.: Transit smart card data mining for passenger origin information extraction. Journal of Zhejiang University Science C **13**(10), 750–760 (2012)

[2] Faroqi, H., Mesbah, M., Kim, J.: Applications of transit smart cards beyond a fare collection tool: a literature review. Advances in Transportation Studies **45** (2018)

[3] Chng, S.: Commute decisions: The role of urban density, travel distance and public transport accessibility. Travel Distance and Public Transport Accessibility (May 31, 2017) (2017)

[4] Saliara, K.: Public transport integration: the case study of thessaloniki, greece. Transportation Research Procedia **4**, 535–552 (2014)

[5] Santos, G., Nikolaev, N.: Mobility as a service and public transport: a rapid literature review and the case of moovit. Sustainability **13**(7), 3666 (2021)

[6] Wu, C.-H., Ho, J.-M., Lee, D.-T.: Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems **5**(4), 276–281 (2004)

[7] Gomes, D., Nabil, R.H., Nur, K.: Banking queue waiting time prediction based on predicted service time using support vector regression. In: 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 145–149 (2020). IEEE