ANALYSIS OF URBAN MOBILITY PATTERNS USING DATA FROM PUBLIC TRANSPORT TICKETING SYSTEM: IMPLICATIONS FOR DEVELOPING AUTONOMIC SYSTEMS

Tânia Fontes
Researcher
Faculty of Engineering of University of Porto

Vera Costa
Researcher
Faculty of Engineering of University of Porto

Pedro Mauricio Costa
Researcher
Faculty of Engineering of University of Porto

Teresa Galvão Dias
Professor
INESC-TEC, Faculty of Engineering of University of Porto

Abstract

Today, mobility patterns have an important role in our society. From urban planning to economy and environment, several applications can be defined. To identify such patterns, in the last few years, different methods have been used based on data collected from mobile devices or smart cards. In the case of public transports, smart cards can be a good way to collect information of those patterns. Unlike mobile devices, where the location of individuals is continuously collected, smart cards do not provide such information. Moreover, smart cards do not collect detailed information about the specific choices of the individual's daily activities, which represents an additional challenge. In fact, when smart cards are used to analyse traffic patterns, several millions of trips and/or smart cards are usually analysed, however, usually the period of time ranges from a few days until three months. Therefore, in order to produce feasible automatic and autonomic systems, this large variability raises an important question: which is the minimum time required to analyse the mobility patterns of a public transport user? How the travels destination can be predicted? What are the main variables which influence such predictions? To achieve this goal, in this work, we analysed the spatial-time mobility patterns of users along three months of data from a medium size Metropolitan European Area located in Portugal, Porto.

1. INTRODUCTION

Mobility patterns have an important role in our society since have many applications ranging from urban planning to economy and from environment to epidemiology. To estimate such patterns, usually mobility models are used. Traditionally, these models are calibrated and validated based on Origin/Destination (O/D) surveys which are applied in order to collect an overview of the itinerary of individual travels (e.g. Li et al. 2013; Abreu and Oliveira 2014). However, these surveys are time consuming, expensive and covering only a restrict number of individuals. To solve some of these problems, today part of this information can be automatically collected: (i) by using mobile devices as GPS and smart phones (e.g. Krumn, 2006; Simmonset al., 2006); or (ii) by using smart cards as credit card or public transport cards (e.g. Hasan et al., 2013; Ma et al. 2013; Tao et al. 2014). However, the information collected by this last group brings specific challenges to relate individual mobility to conventional models since is incomplete. In those systems, the location of individuals is not continuously collected and there is no detailed information about the specific choices of the individual’s daily activities (Bagchi and White, 2005; Pelletier et al., 2011).

In the last decades, several models have been developed and applied in order to predict the destination of a travel (e.g. Krumn, 2006; Simmonset al., 2006). The majority of those methods predict the destination in real-time based on data collected by GPS or mobile phones. To predict those estimates Bayes rules (e.g. Krump 2006), markov models (e.g. Simmons et al. 2006) or clustering analysis (e.g. Ma et al., 2013) have been used. However, when smart cards are used to predict the destination, some differences are found.

Table 1 shows an overview of some recent studies conducted by using smart cards from public transports. Bagchi and White (2005) show that through the use of rule-based processing, public transport smart card data can be used to infer turnover rates, trip rates and the proportion of linked trips. Hasan et al., (2013) analysed the spatial and temporal patterns of individual's mobility in London using the data from smart subway fare card transactions while Tao et al. (2014) examined the spatial–temporal dynamics of bus passenger travel behaviour using smart card data, and Roth et al., (2011) analysed the structure of urban movements. Yu et al., (2014) evaluate the Beijing urban master plan based on subway data, while Ma et al. (2013) developed a novel data mining procedure to extract individual passengers’ travel patterns and travel regularity. Costa et al. (2015a; 2015b) applied different data mining algorithms (J48, top-k and Naïve Bayes) to identify the travel destination of different groups of users. In these works the best accuracy rate obtained ranges between 45% and 65%.

In those studies, different methods have been applied to analyse the spatial temporal patterns of travellers as clustering analysis, GIS techniques, probabilities, and data mining. In these analyses, several millions of trips and/or smart cards are usually analysed. However, the period of time ranges from few days (e.g. Yu et al., 2014; Ma et al., 2013) until three months (e.g. Hasan et al., 2013). Moreover, the accuracy values achieved to identify the travel destination, is in some cases to low. These factors raises two important questions:

- Which is the minimum time required to analyse the mobility patterns of a public transport user?
- How the accuracy rate of the destination probability of a public transport user changes over time? What are the main factors which influence that prediction?

These questions are very relevant in order to produce autonomic public traffic systems. Therefore, since their pertinence, the main goal of this work is find an answer to this question.

Table 1. Studies developed based on data collected by smart cards from public transports.

Study	Domain	Method	Period of time	N.º of trips or users
Bagchi and White (2005)	Bradford (UK) Southport (UK)	Use of surveys to identify the porpoise	35 days	396,331 trips 90,062 trips
Hasan et al., (2013)	London (UK)	Probabilities	3 months	626 users
Ma et al., (2013)	Beijing, (China)	Clustering	1 week	3.8 million users
Roth, et al., (2011)	London (UK)	Clustering	7 days	11.02 million trips
Tao et al., (2014)	Brisbane, (Australia)	GIS techniques	-	5 million trips
Yu et al., (2014)	Beijing, (China)	Data mining	1 day	8.7 million users
Costa et al., (2015a)	Porto (Portugal)	J48, Top-K, NB	2 months	600 users
Costa et al., (2015b)	Porto (Portugal)	J48, Top-K	2 months	803,892 trips

NB: Naïve Bayes

This paper is organized as follows. Section 2 will present the study domain and the used methods while Section 3 will present an overview of the results achieved.

2. MATERIAL AND METHODS

The study was conducted in the Metropolitan Area of Porto (AMP), a medium size European Metropolitan Area located in Portugal. Section 2.1 presents the study domain and section 2.2 the used methods.

2.1 Study domain

The public transport network of the Metropolitan Area of Porto covers an area of 1,575 km^2 and serves 1.75 million of inhabitants (INE, 2013). The network is composed by 126 buses lines (urban and regional), 6 metro lines, 1 cable line, 3 trans lines, and 3 train lines (TIP, 2015). Such system is operated by 11 transport providers. The biggest transport providers are Metro do Porto and STCP.

In the Oporto network, users have available an intermodal and flexible ticket called Andante. Andante is a zonal system where the zones are defined based on known travel patterns and not concentrically defined as the most national and international multimodal ticket systems. Although this is a more fairly system than the traditional concentric, there are some drawbacks since is more complicated to use. Figure 1 exemplifies the functionality of Andante considering two different points of AMP. This system was introduced in 2005 but only since 2011 it was completed integrated and adopted by all transport providers.

With the Andante system, a validated occasional ticket allows for unlimited travel within a specified time period, currently 1 hour for the minimum 2-zone ticket, and longer as the number of valid zones increases. Andante holders can use different lines and transport modes in a single ticket. Tickets must be validated only before travel.

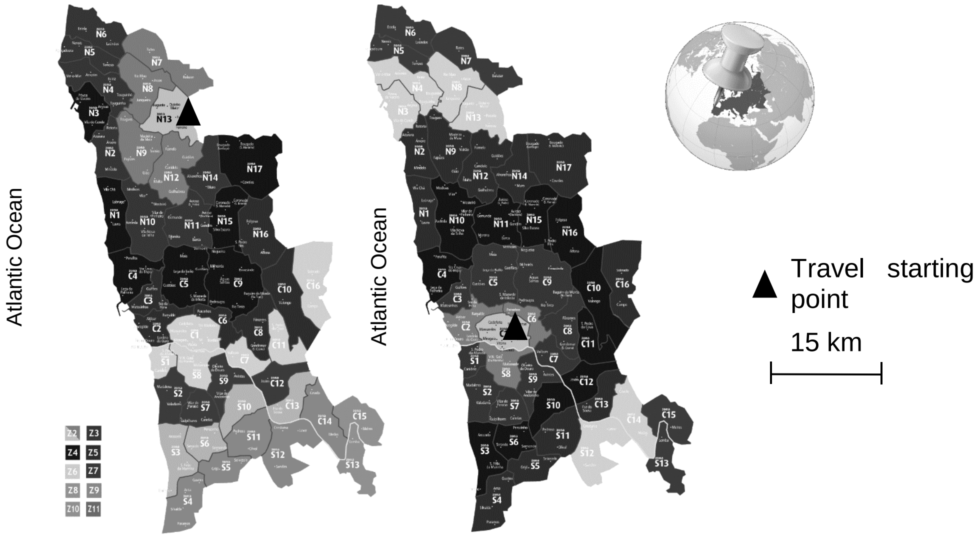

Figure 1. Porto zonal system (example for a travel starting in two different locals)

2.2 Methods

In this work three months of data were used, January, April and May of 2010. Initially, this database had near than 9 millions of validations, and represents about 30 % of the overall validations of the network. Data were provided by only one transport provider, the STCP, urban bus company.

Since this system is based in the validation only at the entrance, to identify the travel destination, an inference algorithm was used (Nunes et al., 2015). Due the absence of information, the results obtained with the application of this algorithm were restricted, since data from only one transport provider was available. Thus, in order to minimize the errors, in this work were only analysed the users of the network which have at least 80 % of destination inferred with this method, and at least 120 validations during the three months.

Based on this data, the spatial-temporal patterns of the public urban network were analysed. First, the data were used to analyse the individual spatio-temporal patterns of users along time. To perform this first analysis, the probability distribution of the most frequent travel destination of which user was analysed. The results were evaluated considering different groups of users according to the travel frequency: 120, 121-150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450 and > 300 travels by month.

In order to automatically classify the travel destination of a public transport user, a probability analysis was conducted. To perform such analysis two approaches were defined: i) one based in the past travels of the user; and another ii) based on the network knowledge (N). In the first approach three different criteria were analysed: i) the most probable travel origin of the user (O); ii) the most probable travel destination of the user, independently of the origin (D); iii) and the most probable travel destination of a user when is know their travel origin (DD). In the second approach, the data from all users were aggregated. For both previous approaches, the influence of the weekday (W) and hour of day (H) was analysed. To study the influence of the hour, a time window of 1, 2 and 3 hours was selected (H1H, H2H and H3H). Therefore, 11 combinations were defined considering the past travels of the users (O, OW, D, DW, DD, DDH, DDW, DDWH, DDWH1H, DDWH2H, DDWH3H) and 10 considering the network knowledge (N, NH, NW, NH1H, NH2H, NH3H, NWH, NWH1H, NWH2H, NWH3H). For each case, the first five more probable travels were analysed which totalize 105 predictions.

To understand the behaviour of the error trend along the time, each prediction was estimated considering a variance of training set. Thus, the predictions start using the first 7th days (one week) to train and one day, the 8th day, to test. We test one day each time, considering the first day after the training set. Thus along the time the training set increases, while the test remains the same. Such predictions were done for the two consecutive months of data available (April and May of 2010). For each simulation, 5 000 random travels were analysed.

3. RESULTS

Uncovering patterns of usage based on trips performed using public transport presents a range of challenges, as described in the previous section. In order to minimise them, the most frequent trip is used as a signpost or reference for individual travellers. Therefore, the analysis performed focuses on this reference to investigate the robustness of travelling behaviour over time.

Figure 1 presents the average and the standard deviation of the probability distribution of the most frequent travel done by a user, considering a period of three months. In this analysis, nine different groups of users were used, considering the number of travels done during the three months of data analysed: <120, 121-150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450 and >= 300 travels. Figure 1 shows the results for the first fourth groups. In this study near 7 000 users were analysed and near 1.3 million of records.

Preliminary results show that probability of distribution of the most frequent travel of users vary along the time, in particular when less than two weeks of data are analysed. Along the three months, the probability value ranges between 0.1 and 0.2, and as expected, when the frequency increases, this value tends to decrease. Moreover, the analysis of such trends shows differences across the groups analysed. The results suggest that, while for some groups of users, the probability stabilizes around two months (e.g. <120 travels), for other groups, these stabilization is only achieved when near three months of data are used (e.g. 351-400 travels). The difference between the groups is partly explained by inherent patterns of usage, resulting in different amounts of trips performed during the three months analysed. Amongst other factors, the lack of added context (e.g. location provided by personal devices), different purposes for the trip (work related vs. occasional events) and variability in time, these patterns tend to emerge only after extended usage.

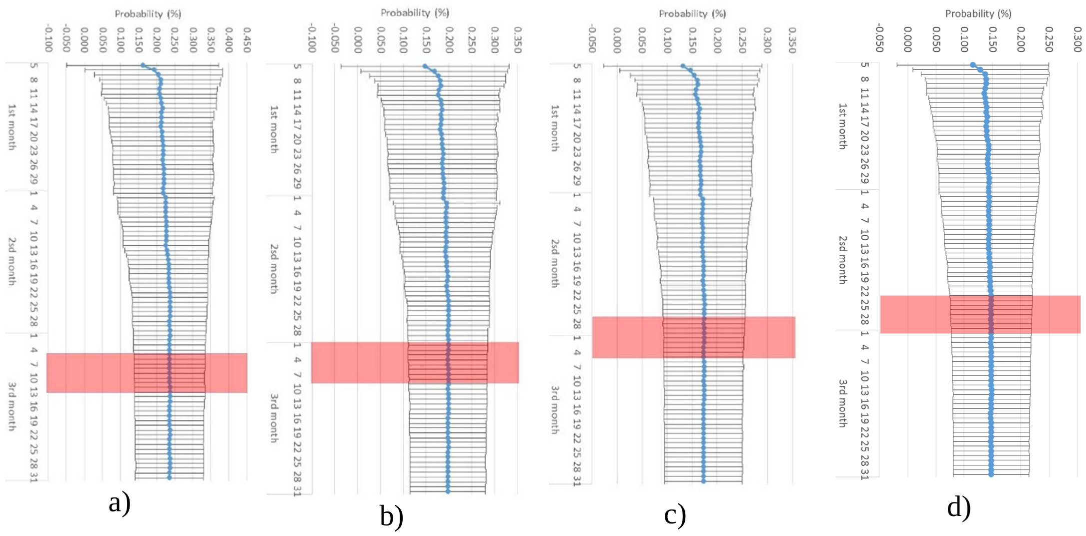

Figure 2. Average probability distribution of the most frequent travel of a user (and standard deviation) considering the travel frequency along a period of three months. Examples to: a) <120 travels, b) 121-150 travels, c) 151-200 travels, d) 201-250 travels.

Figure 3 shows examples of the average accuracy obtained to predict the journey destination for a weekday and for a weekend day. The results show that while during the weekdays the maximum accuracy values are of about 60%, during the weekend days lowest values are found (50% approximately). When the set of past data each user increases, the probability to adivinhar also increases. During the two consecutive months analysed (April and May of 2010), the values ranged from 57 % to 65 % and from 45 % to 49% during the weekdays and weekends respectively. The maximum accuracy values are achieved when the past data of the users are used, particularly the day of the week and the hour of the day where the travel starts. Nevertheless is interesting to note, that where the accuracy value is computed analysing the three most frequent travels, the probability increases significantly. In this case, the maximum accuracy values achieve 80 %, and 60 % during the weekdays the weekends respectively. This suggest that the predictions presented in some previous works can be improved (Costa et al. (2015a; 2015b)). Nonetheless, is curious to note, that when are analysed the three most frequent destinations of a travel, the network knowledge can allow very good predictions. This occur when is only used the occurrence hour of the travel in the predictions.

Preliminary results suggest that a large amount of data needs to be used to analyse individual travel patterns using smart cards from public transport systems. Nevertheless, when no data is available, the knowledge of the network can be used to provide such prediction with high accuracy bridging an important gap recorded in previous works. Thus with this method new intelligent transport services can be developed namely by distributing directional promotions able to influence the daily consumption of different services.

AKNOWLEDGEMENTS

This work was financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within the project “SI IDT - 38957/2012 F3”. The authors also acknowledges to the transport provider of Oporto, STCP, which provide travel data for the project and also to our partners in the project, NOVABASE and OPT companies.

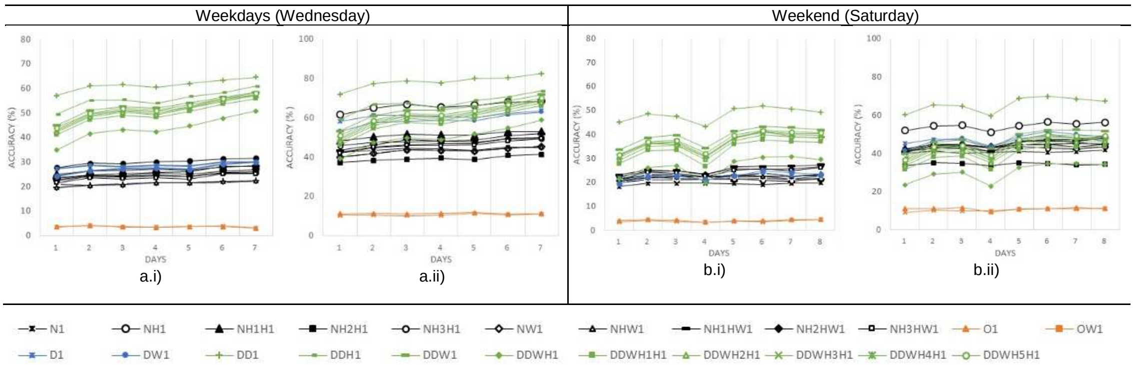

Figure 3. Average accuracy rate of journey destination prediction obtained for a weekday (a) and a weekend day (b). Values were computed based in the prediction of the most frequent destination (i) and based in the three most frequent destinations (ii).

REFERENCES

Abreu, BRA, Oliveira, LK. (2014). The Potential of Response Rate in Online Transportation Surveys. Procedia - Social and Behavioral Sciences, 162, 34–41.

Bagchi M, White PR (2005). The potential of public transport smart card data. Transport Policy 12 (2005) 464-474.

Costa C., V. Fontes, T., Costa, P.M., Dias, T.G., (2015a). Prediction of journey destination in urban public transport. EPIA 2015 – XVII Portuguese Conference on Artificial Intelligence, Coimbra, 8-11/09.

Costa, V., Fontes, T., Costa, P.M., Dias, T.G. (2015b). How to predict journey destination for supportingcontextual intelligent information services?, IEEE 18th International Conference on Intelligent Transportation Systems (ITSC 2015), Las Palmas, Submited.

Hasan S, Schneider CM, Ukkusuri, SV, González MC (2013). Spatiotemporal Patterns of Urban Human Mobility. Journal of Statistical Physics, 151:304–318.

INE, (2013). Anuário Estatístico da Região Norte – Statistical Yearbook of Norte Region. Instituto Nacional de Estatística I.P., Portugal.

Krumn J (2006). Real Time Destination Prediction Based On Efficient Routes. Society of Automotive Engineers (SAE) 2006 World Congress

Li W, Wang W, Zhao Y, (2013). Research and Development on Survey and Statistical Analysis Software of Resident Travel OD Based on B/S Mode. Intelligent and Integrated Sustainable Multimodal Transportation Systems Proceedings from the 13th COTA International Conference of Transportation Professionals (CICTP2013).

Ma X, Wu Y-J, Wang Y., Chen, F, Liu J (2013). Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies 36, 1–12.

Nunes et al., (2015). Temporary User-Centred Networks for transport systems. Submitted.

Pelletier MP, Trépanier M, Morency C, (2011). Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4), 557–568.

Roth C, Kang SM, Batty M, Barthélemy M (2011). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. PLoS ONE 6(1): e15923.

Simmons R, Browning B, Zhang Y, Sadekar, V (2006). Learning to Predict Driver Route and Destination Intent. Intelligent Transportation Systems Conference, IEEE, Toronto, 127 – 132.

Tao S, Rohde D, Corcoran J (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. Journal of Transport Geography, 41, 21–36.

TIP (2015). http://www.linhandante.com/. Transportes Intermodais do Porto.

Yu W, Mao, M, Wang B, Liu X, (2014). Implementation Evaluation of Beijing Urban Master Plan Based On Subway Transit Smart Card Data. 22nd International Conference on Geoinformatics, IEEE, 1 – 6.