Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 162 (2014) 340 – 349

Procedia
Social and Behavioral Sciences

XVIII Congreso Panamericano de Ingeniería de Tránsito, Transporte y Logística (PANAM 2014)

Considerations about the analysis of ITS data of bicycle sharing systems

Maria Bordagaray[a]*, Achille Fonzone[b], Luigi dell'Olio[a], Angel Ibeas[a]

[a]*University of Cantabria, Av. De los Castros s-n, Santander 39005, Spain*
[b]*Transport Research Institute, Edinburgh Napier University, Merchiston Campus, 10 Colinton Road, Edinburgh EH10 5DT, United Kingdom*

Abstract

Handling and managing data automatically collected by Intelligent Transport Systems (ITS) is a major opportunity and challenge for transport professionals nowadays. This study guides the management of smartcard data from public bikes by providing criteria to detect travel patterns that describe the specific use of bike-share systems and which cannot be encountered in other transport modes. The guidelines have been put into practice with data from the TusBic system in Santander, Spain.

The major discovery that has resulted from the analysis of the data is the high number of records that describe very short trips that show the same terminal at origin and destination. An algorithm is proposed that assumes the users try and return the bike if this is not working properly, and pick a new one from the same terminal. In such cases, the records are joined to describe a unique trip by considering the origin as the pick-up instant of the first bike, and the destination, the instant at which the second bike has been returned.

The indications presented in this article should be considered in future studies of demand and level of service of public bicycle systems since not only they can make a big difference in the accuracy of the results, but also they can provide interesting information regarding the management and design of the system. Therefore, they are of interest for different stakeholders such as politicians and decision makers, service planners, and agencies responsible for the operations and direct management of public bicycle systems.

Peer-review under responsibility of PANAM 2014.

Keywords: Intelligent Transport Systems, smartcards, data mining, travel behaviour, demand analysis

* Corresponding author. Tel.: +34-942-201734
E-mail address: bordagaraym@unican.es

Peer-review under responsibility of PANAM 2014.
doi:10.1016/j.sbspro.2014.12.215

1. Introduction

Undoubtedly the huge amount of information supplied by ITS can improve the results of transport models and, hence, the conclusions of any study in which it is available. However, making the most of this data is difficult because of several reasons: incomplete knowledge about the potential use of these records, lack of awareness of the limits to the validity and reliability of ITS data, and complexity of managing raw data and developing efficient algorithms its application. This work has been motivated mainly by the need for studying and improving the validity of bike-share smartcard system data for travel demand analysis. Acknowledging the lack of standard analysis procedures, we also discuss the data preparation stage, as a memorandum for scholars and practitioners in new field of bike-share analysis using ITS data.

The interest on the demand for cycling has increased in the last decade. Numerous studies have focused on the factors affecting such demand (Martens, 2004; Moudon et al., 2005; Olio, Ibeas, Bordagaray, & Ortúzar, 2011; Ortúzar, Iacobelli, & Valeze, 2000). However, less numerous are the contributions dealing with bike sharing systems, and more specifically, those applying data from ITS such as smartcard transactions. Vogel, Greiser, and Mattfeld (2011) develop a data mining process to study the mobility patterns in order to understand the imbalances that occur in the system. They also identify the limited research that had already applied data mining algorithms to study the demand of public bike schemes by 2011. The spatial demand and its profile along the day, together with the prediction of available docks are the objective of the studies. Such results concern the design of the system and are needed for an optimal operation of the system (Romero, Ibeas, Moura, Benavente, & Alonso, 2012). Interestingly, O'Brien, Cheshire, and Batty (2014) compare the demand of various public bike systems, allowing the interpretation of the differences encountered in travel patterns.

The municipality of Santander (Spain) has recently introduced a bike-share system – TusBic – with the aim of fostering an increase of the modal share of bicycle. Data collected by the system are expected to provide insights on the potential for bicycle mobility in Santander. The initial analysis of data from TusBic has shown the existence of a particular user behavior pattern related to the presence of defective bicycles. Numerous occurrences have been detected of very short trips, starting and ending at the same station, followed by longer ones, departing from the same station within few minutes from the end of the previous one, and finishing at a different destination. This behavior should be carefully considered in demand analysis, because the two records may be part of the same journey and if this is not accounted for an unduly inflated demand would be measured. Furthermore, the frequency and the spatial distribution of this kind of trips could provide useful input to system design and management purposes. The research puts forward a procedure to detect and handle this type of behavior using smartcard data and shows the relevance of the phenomenon in the Santander case.

The paper is structured as follows: firstly, a brief summary on previous applications of ITS data is provided. Then the TusBic system in Santander is presented. An approach is proposed for the identification of pairs of records concerning the same trip, and the application to the Santander case is illustrated. Finally, general conclusions are drawn from the Santander case study.

2. The TusBic bike-sharing system in Santander

Santander is a medium-size coastal city with a population of about 200,000 inhabitants in the North of Spain. The public bike system TusBic was set up in 2009 and was aimed at stimulating the adoption of the bicycle as an alternative mode of transport among citizens. However, despite the introduction of TusBic, the modal share of the cycle mobility in the city is still scarce (less than 1% of the total number of trips correspond to cycling). Steep slopes and the rainy and humid weather are generally acknowledged as the main reasons of the still low demand for cycling. The analysis of the demand for public bicycles is expected to cast light on the attitude towards cycling of the inhabitants of Santander.

TusBic envisage the use of a smartcard to rent the bicycles. For this research smartcard data of 2011 is available. In 2011 the system consisted of 14 terminals, about 200 bikes and more than 300 docks. The system can be used only by registered users. Annual, weekly and daily subscriptions can be bought for €10, 5 and 1 respectively. The subscription covers trips lasting less than an hour. To discourage prolonged use, annual subscribers are charged €0.30 for every extra half hour. The charge is €0.50 for weekly subscribers, €0.60 euro for daily ones.

The analysed dataset includes 26,290 records, collected from the 1st of July to the 31st of August 2011. The study is limited to the summer period because the usage of the system in the rest of the year is too scarce due to the inclement weather. For each rental, the system records:

- Bicycle pick-up and drop-off docking station
- Identification number of the stand occupied by the bicycle at the origin station
- Time (date and time – hour, minute, second) of collection and return of the bicycle
- User's type of subscription (annual, weekly or daily)
- Subscriber identification number
- Bicycle identification number

3. Bike-trial trips in Santander

3.1. Raw database preparation

Normally preliminary elaboration of raw databases is required to make records ready to be analyzed by automated routines. The case of Santander allows illustrating some of the most common necessary passages.

[1] Data interpretation: A clear understanding of the meaning of the variables that have been measured in each record is essential. This kind of information sometimes is not available or clear, and collaboration with the system operator is needed. Problems may arise when the subject interested in data analysis (e.g. a city council) is not the owner of the data (which often belongs to the system operator).

[2] Database organization: The very first stage is to organize the files supplied by the system so that they can be easily analysed. The TusBic system provides a file for each day of operation. All the records have been collected in a single file to make data analysis easier. However, it should be highlighted that this was possible in our study because we have analysed only two months of a small system. In general databases can include huge numbers of records, and this can makes data management burdensome.

[3] Data cleaning: Before the analysis data needs to be cleaned so the analysed database does not include
- Repeated entries: The database in this research contained 882 duplicated records of two specific days during the period of study
- Missing data. In automated systems, missing data is often identified by special entries. Our of the 26,290 rentals recorded in July and August 2011 in the TusBic system, 46 showed a label "not referenced" in the field of the destination docking station. These records have been omitted in the analysed database. When the missing data is not used in the analysis, other information included in the record can still be used.

At the end of the cleaning process, 26,244 rental records, made by 5,615 users are left for the analysis of demand. Fig. 1 reports the usage of the system per kind of subscription.

[4] Data elaboration: New quantities may be calculated from the automatically registered fields, which provide information with more directly linked with the objective of the analysis. For instance in the case of Santander, information has been extracted as to:
- Day of the week and type of day (weekday, Saturday, Sunday; normal day/holiday) from the date of the rental origin.
- Duration of rental as the time elapsed between the collection and the return of the bicycle.

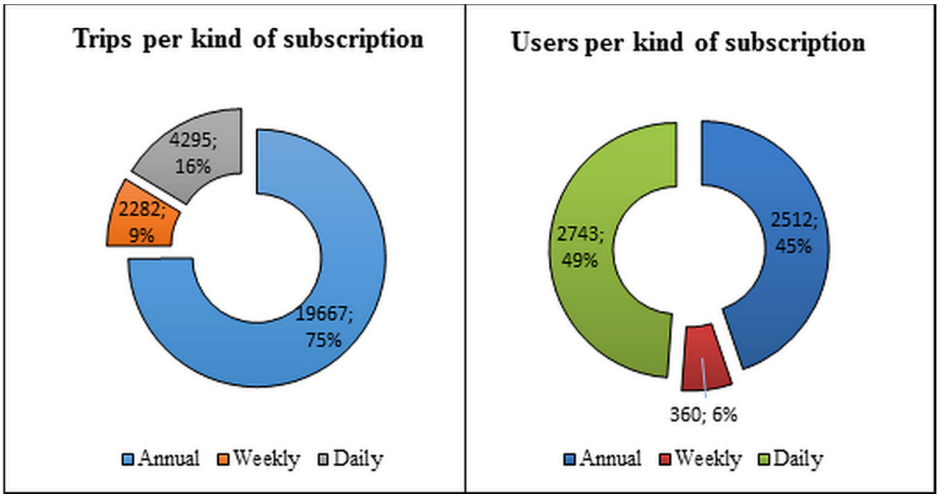

Fig. 1. Sample description regarding the type of subscription: (a) share of trips; (b) share of users.

3.2. Initial analysis of bike-share demand

When standard procedures are not available, data mining has to start from expert knowledge. Also in our case the familiarity with the city and the service has been fundamental to figure out systematic behavioral patterns, whose existence has then been confirmed by ITS data. In particular, different behaviours were expected for journeys in which the origin and the destination coincide (hereafter COD trips) and those with different docking stations at origin and destination (DOD trips). The distribution of rentals turns up to be as follows:

- COD rentals: 6,335 cases (24.2%)
- DOD rentals: 19,909 cases (75.8%)

Fig. 2 represents the frequency distributions of rental duration for these two kinds of journey.

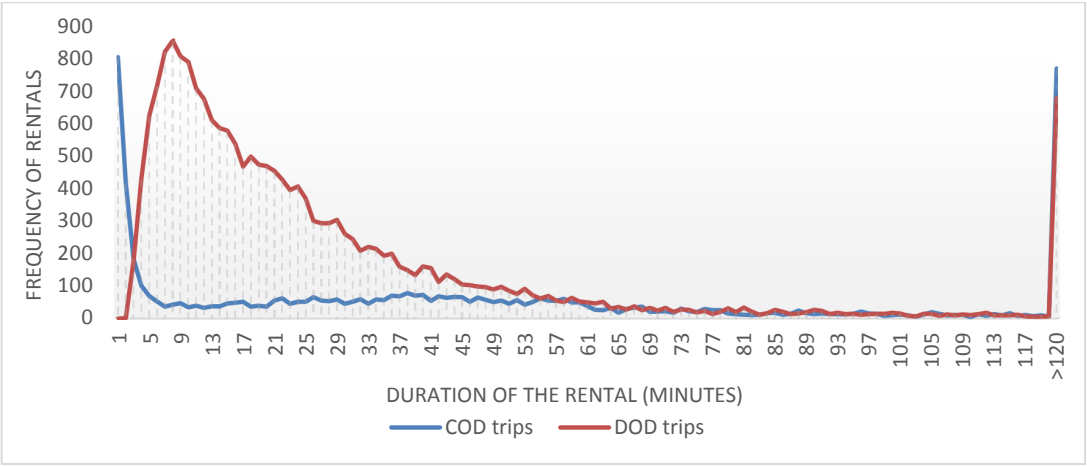

Fig. 2. Histogram of rentals where the origin and destination are the same (O=D) or different (O ≠ D), according to the duration of the rental.

The curve of DOD rentals shows an increasing, quasi linear trend for durations up to 7-8 minutes, where the frequency reaches a maximum of 856 rentals. The DOD case in Santander results in an interquartile range of 10-33 min (Table 1). The COD journeys present a surprisingly different distribution; the maximum number of rentals happens for a duration of less than a minute and 25% of records show duration shorter than 5 minutes.

Table 1. Quartiles of the duration of the rentals

	COD rentals	DOD rentals
1st quartile	5 mins	10 mins
Median	37 mins	18 mins
3rd quartile	67 mins	33 mins

The very high number of COD journeys with low durations is interpreted as the cases in which the bicycle is tested and then returned to the terminal without actually making the trip, presumably because it is faulty and so the users decided to replace it before heading out for their destination. We refer to this behaviour as “bike trial”. If our interpretation is correct, most short COD trips should be immediately followed by a new rental from the same terminal. To the aim of demand analysis, it seems natural to consider the bike trial rental and the following one as part of the same “journey with bike substitution”. Clearly failing in acknowledging the existence of this kind of trips would result in biased results of the demand analysis, regarding both the number of trips generated and attracted by each docking station (and so by different city zones), and the average duration of trips by bicycle (i.e. regarding the mobility segment for which bicycle is competitive). However, to the authors' knowledge, the existence and the relevance of bike trials and of journeys with bike substitution have not been considered in the literature, probably because it is typical of bike sharing schemes and the operations of such schemes is still largely to be explored.

3.3. Detection of bike trial

Fig. 3 describes the flow chart of an algorithm to deal with bike trials and journeys with bike substitution.

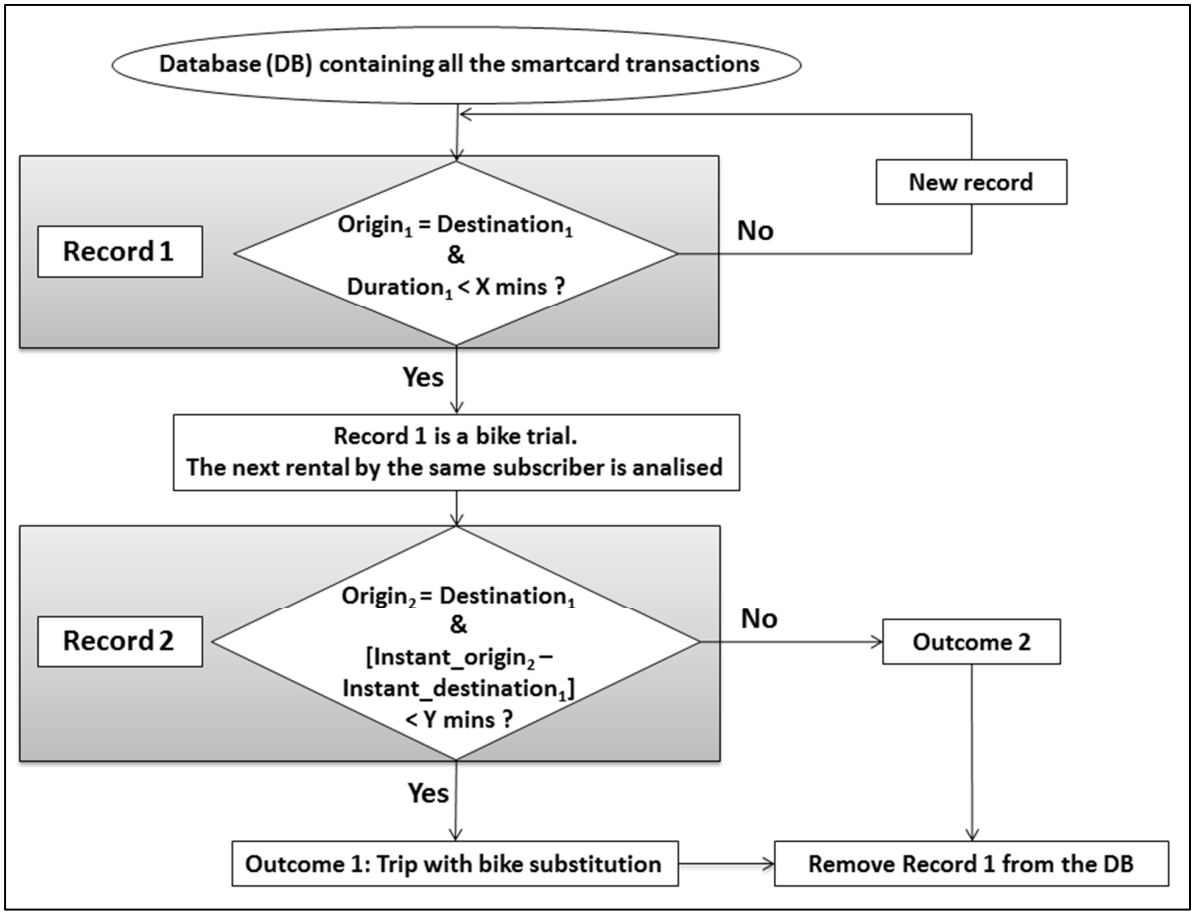

Fig. 3. Data mining algorithm to test the bike trial and bike substitution hypotheses.

The proposed procedure involves two steps. In the first, every rental in the database is evaluated (the record considered in this stage is called Record 1), and two characteristics are checked: 1) whether the destination docking station ($Destination_1$) coincides with the origin ($Origin_1$); 2) if the rental duration ($Duration_1$) is lower than *X* minutes. *X* such as *Y* below are two arbitrarily thresholds. If either condition is not verified, then the rental is interpreted as a journey in itself and it remains in the database for demand analysis. On the contrary, if both conditions hold for Record 1, the rental is considered a bike trial. In the latter case, the following rental made by the same user (Record 2) is evaluated to determine: 1) whether the same user picks a new bike at the same docking station where the previous one was returned; 2) if the interval between the return of the first bicycle and the rental of the second is smaller than *Y* minutes. Two cases can arise:

- A Record 2 exists which satisfies both conditions (Outcome 1): Record 1 and Record 2 are considered part of the same journey with bike substitution.
- No Record 2 is detected fulfilling both conditions (Outcome 2): If the second rental takes place at the same docking station of the first one, depending on the interval between the two, several scenarios may be hypothesized: the user does not intend to go anywhere but just to test a bicycle, or he would like to make a trip but cannot find a suitable bicycle, or he started his planned journey but was then he decided to abort it.

Record 1 is not be considered in the database for the analysis of demand in both cases. Nevertheless, distinguishing the case of Outcome 1 from these of Outcome 2 is interesting, because the former can be linked to malfunctioning of bicycle and so to the quality of the service offered by the system.

The results of the algorithm depend on the value chosen for X and Y. The former is the minimum rental duration for a record to be considered a real trip and not a bike trial or equivalently the maximum duration of a bike trial. The latter is the maximum interval between returning a bike and collecting a new one in the case of journey with bike substitution. A possible approach to choosing *X* and *Y* is illustrated in the TusBic case. Fig. 2 shows that COD trips present a peak of frequency for rental durations up to 5 minutes, whereas the distribution is rather uniform for longer rentals. It can be hypothesized that the two ranges of duration are typical of different types of rentals: the former of bike trial, the latter of proper trips centered on the same docking station. Therefore a value of 5 minutes is assumed for *X*. With this threshold, 1580 rentals are classified as bike trials, where 77.9% correspond to annual subscribers, 8.5% to weekly and 13.6 to daily subscribers. This share is similar to the distribution of trips in the sample (Fig. 1a), which means that the percentage of bike trials compared to the total number of trips made by each of the three types of subscribers remains constant for all of them and results in the 5-6%.

On the other hand, Fig. 4 shows the sensitivity of the number of Outcomes 1 (and, therefore of Outcomes 2) against the magnitude *Y* for *X*=5 min. *Y* is the time to choose the new bike and to book it using the automatic interface at the terminal. The number of additional Outcomes 1 (trips with bike substitution) decreases sharply as *Y* increases so as to detect more than 95% of the cases for *Y*=5. However, as it can be seen in Fig. 4, it is for *Y* equal to 14 minutes that the number of Outcomes 1 and 2 establishes, so Y=13 minutes is assumed to be the threshold for Santander. Although such interval may appear too long for a bike change in the case of commuting trips – when users can be in a hurry – or rentals made by experienced system users, it is not unreasonable for recreational trips and above all when users are not familiar with the system, this is, when more time may be needed to choose the new bike and deal with the interface. Furthermore, sometimes the registration of daily or weekly subscribers may take longer than average because the system fails in detecting the credit card where the fee is charged. This may cause queues to access the machine to register and delays in booking bicycles also for commuting trips and experienced users. In general, the decision about the value of *Y* should be taken considering the average time for registration and the percentage of regular users.

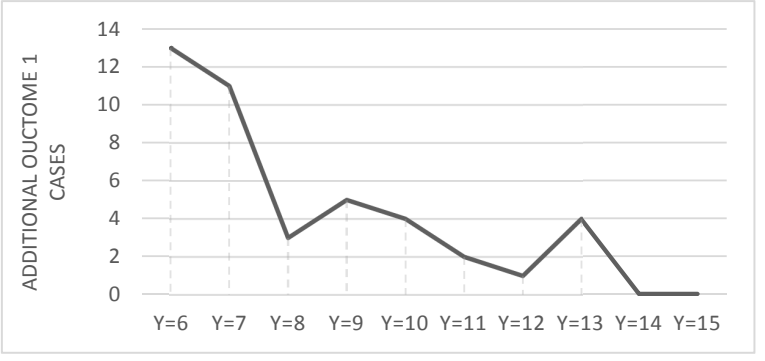

Fig. 4. Sensitivity of the number of cases classified as trip with bike substitution.

Table 2 informs on the distribution of Outcomes 1 and 2 detected distinguishing the subscription.

Table 2. Number of cases Outcomes 1 and 2 regarding the type of subscription

	Annual	Weekly	Daily	Total
Outcome 1	994 (80,7%)	118 (88,1%)	185 (85.6%)	1,297 (82.1%)
Outcome 2	237 (19,3%)	16 (11,9%)	30 (14.4%)	283 (17.9%)
Total bike trials	1,231	134	215	1,580

The first insight relates to the fact that, in the case of Santander, most of the bike trials (at least the 80%) detected with X=5 min can be considered first legs of trips with bike substitution. Furthermore, although the difference is not dramatic, it can be seen that Outcome 1 cases are more frequent for less regular users (weekly and daily subscribers). Such result can be explained by the fact that experienced users are more able to distinguish bicycles in good conditions. However further analysis is needed because the result may also indicate that these users place less importance on the bike itself, or it may also be linked to the purpose of the journey.

3.4. Impact of considering bike trials and bike substitution

The results obtained from the pattern described as bike trials provide interesting information, such as the bikes that are being returned the most. Since the system records the identification number of the bike that is rented, once the trips interpreted as bike rentals have been identified, the number of times that each bicycle has been substituted could serve as an indicator of the bike conditions. Undoubtedly, the agency responsible for the maintenance of the bikes would benefit from such information.

Since the bike trials in the case of Santander correspond to the 5% of the total records, removing such rentals leads to an impact in the results of the demand analysis. For instance, the graphic showed in Fig. 2 is represented in Fig. 5 where the bike trials have been removed. The trips described in the latter figure are the ones that would undergo the demand analyses.

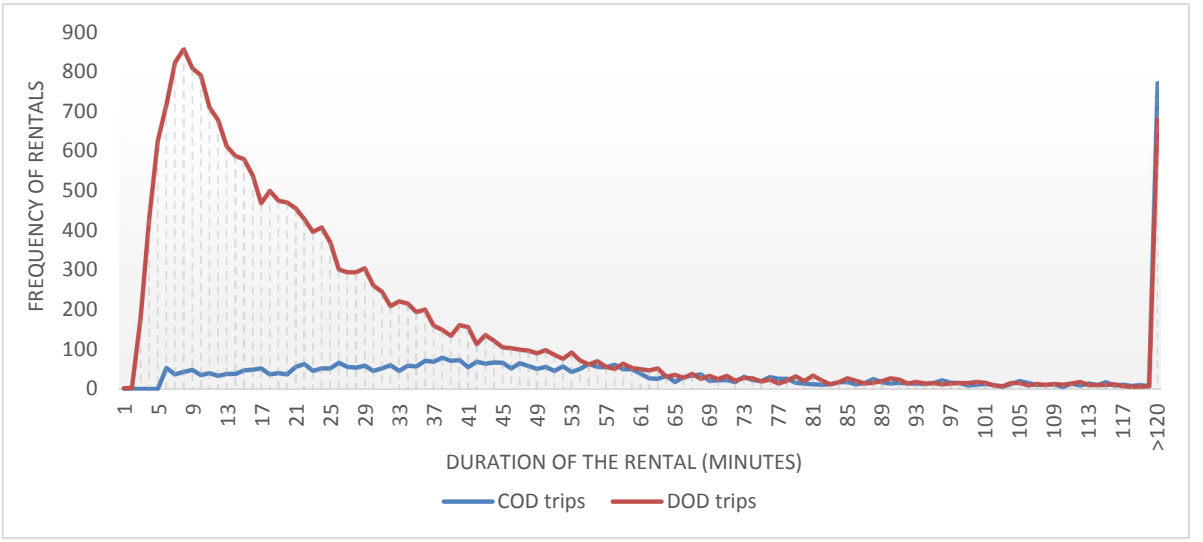

Fig. 5. Histogram of the trips after removing the rentals considered bike trials.

Needless to say, by removing the bike trials from the sample, the number of trips where the origin and the destination coincide decrease. Consequently, the OD patterns turn out to be rather different than in the case that this phenomenon is not considered. Fig. 6 shows the frequency of bike trials that have been registered in each terminal of the system. This graph gives a clue on the OD pairs where the demand decreases the most when considering the specific pattern described in the algorithm presented in this research.

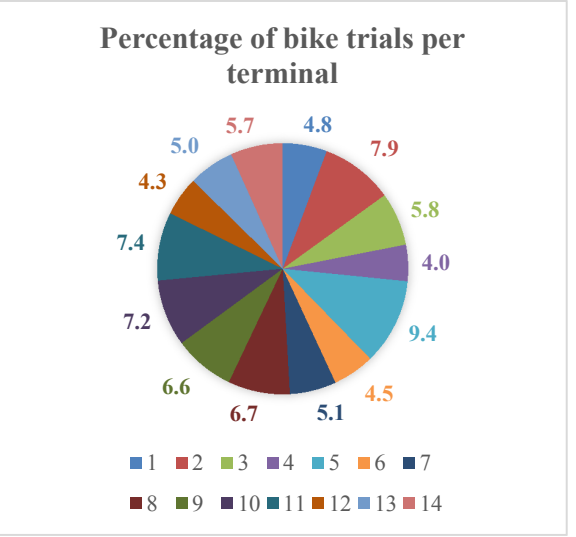

Fig. 6. Occurrence of bike trials compared to the total number of rentals registered at each terminal.

4. Conclusions

Extracting information from data provided by ITS is an art which requires both data mining skills and knowledge of the analysed system. Clearly, the kind of information which can be obtained depends on the characteristics of the recorded data. In the case of TusBic bike share system in Santander, the analysis can look at overall rentals, rentals by subscriber, rentals by bicycle, or rentals by docking station. The paper has looked at a dataset of two months, highlighting that raw data has to be carefully processed and examined before being used to generate information regarding the demand for and the supply of public bicycles.

In particular the existence has be shown of “trips with bike substitution”, made up of two legs: an initial short rental in which a first bike is tested and then returned to the docking station where it was probably because not working properly, followed by the rental of a new bicycle which allows carrying out the initially planned journey. Clearly considering the two rentals as different journeys would distort the study of demand. We presented an algorithm to detect this kind of trips, finding that they represent more than 5% of the recorded transactions in the case of Santander.

The detection of trips with bike substitution can provide valuable information regarding:

- User satisfaction: in fact, it is easy to see that users are disappointed if they have to interrupt their trip to change bicycle. This number of trips with bike substitution would enrich the study of quality perceived by users, which was approached by Bordagaray, Ibeas and dell’Olio (2012) for the TusBic bike-sharing service in Santander.
- Management and maintenance of the fleet of bikes: since the ITS system records the bike of each rental, it is possible to identify the bikes that have been returned more frequently and so to arrange for physical checks and reparation more effectively.

Acknowledgements

The authors would like to acknowledge the financial support provided by the Spanish Ministry of Economía y Competitividad in the projects TRA2010-18068 and TRA2012-39466-C02-02. Furthermore, it is the authors' desire to thank the City Council of Santander and JCDecaux for providing the data that has allowed validating this research.

References

Bordagaray, M., Ibeas, A., & dell'Olio, L. (2012). Modeling User Perception of Public Bicycle Services. *Procedia - Social and Behavioral Sciences*, *54*, 1308–1316. doi:10.1016/j.sbspro.2012.09.845

Martens, K. (2004). The bicycle as a feedering mode: Experiences from three European countries. *Transportation Research Part D: Transport and Environment*, *9*, 281–294. doi:10.1016/j.trd.2004.02.005

Moudon, A. V., Lee, C., Cheadle, A. D., Collier, C. W., Johnson, D., Schmid, T. L., & Weather, R. D. (2005). Cycling and the built environment, a US perspective. *Transportation Research Part D: Transport and Environment*, *10*, 245–261. doi:10.1016/j.trd.2005.04.001

O'Brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, *34*, 262–273. doi:10.1016/j.jtrangeo.2013.06.007

Olio, L., Ibeas, A., Bordagaray, M., & Ortúzar, J. D. D. (2011). Modeling the Effects of Pro Bicycle Infrastructure and Policies Toward Sustainable Urban Mobility, 1–8. doi:10.1061/(ASCE)UP.1943-5444.0000190.

Ortúzar, J. D. D., Iacobelli, A., & Valeze, C. (2000). Estimating demand for a cycle-way network. *Transportation Research Part A: Policy and Practice*, *34*, 353–373. doi:10.1016/S0965-8564(99)00040-3

Romero, J. P., Ibeas, A., Moura, J. L., Benavente, J., & Alonso, B. (2012). A Simulation-optimization Approach to Design Efficient Systems of Bike-sharing. *Procedia - Social and Behavioral Sciences*, *54*, 646–655. doi:10.1016/j.sbspro.2012.09.782

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, *20*, 514–523. doi:10.1016/j.sbspro.2011.08.058