Modeling Cyclists' Itinerary Choices: Evidence from a Docking Station-Based Bike-Sharing System

Un modello per gli itinerari dei ciclisti: risultati da un bike-sharing a stazioni fisse

S. T. Gaito - G. Manzi - G. Saibene - S. Salini - M. Zignani

Abstract This paper presents a bike itinerary choice model for a bike sharing system (BSS). Machine learning techniques are used to characterize the bike itineraries within a given day. We present a method to detect the 'most faithful' users with respect to itinerary types and a Bayesian network model which emphasizes the link between the conditions with which the itinerary is chosen and the itinerary type.

Abstract *Questo articolo presenta un modello di scelta di percorsi ciclabili per un sistema di bike sharing (SBS). Tecniche di machine learning vengono utilizzate per caratterizzare i percorsi in un dato giorno. In questo articolo viene presentato un metodo per rilevare gli utenti 'più fedeli' rispetto ai tipi di itinerario e un modello di rete bayesiana che enfatizza il collegamento tra le condizioni in cui si sceglie l'itinerario e il tipo di itinerario.*

Key words: Bike sharing, Bike itinerary choice, Machine learning, Bayesian Network

Sabrina Tiziana Gaito
Department of Computer Science "Giovanni degli Antoni" and Data Science Research Center, Università degli Studi di Milano, via Celoria, 18, 20133 Milan, Italy, e-mail: sabrina.gaito@unimi.it

Giancarlo Manzi
Department of Economics, Management and Quantitative Methods and Data Science Research Center, Università degli Studi di Milano, via Conservatorio, 7, 20122 Milan, Italy, e-mail: giancarlo.manzi@unimi.it

Giorgio Saibene
Hamburg Business School, University of Hamburg, Moorweidenstraße 18, 20148 Hamburg, e-mail: giorgio.saibene@studium.uni-hamburg.de

Silvia Salini
Department of Economics, Management and Quantitative Methods and Data Science Research Center, Università degli Studi di Milano, via Conservatorio, 7, 20122 MiIan, Italy, e-mail: silvia.salini@unimi.it

Matteo Zignani
Department of Computer Science "Giovanni degli Antoni", Università degli Studi di Milano, via Celoria, 18, 20133 Milan, Italy, e-mail: matteo.zignani@unimi.it

1 Introduction

Urban mobility is receiving increasing attention as it is considered one of the most important dimensions of the so-called *smart city* [1]. Recent developments in urban planning management have led BSSs to be a viable complement to traditional public transport systems. However, there are some important quandaries in organizing a successful BSS, for example when rightly predicting the behavior of users, and avoid an uneven distribution of the bikes across the city.

In this paper we are concerned with the optimization of fixed BSS, i.e. BSS with docking stations, and implement a decision framework to help policy makers to obtain an optimal prediction of cyclists itinerary choice in the case of a consolidated docking station-based BSS, the *BikeMi* system in Milan, Italy. It is operated by a private company having a service contract with the Municipality of Milan and receiving in return discounted advertisement spaces nearby the docking stations.

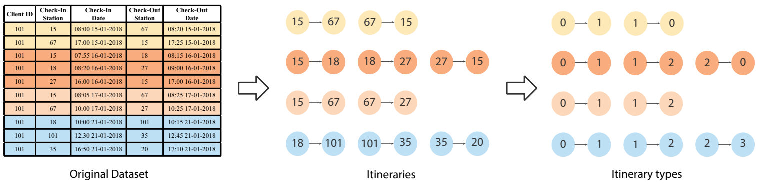
Obviously, all the actors in this kind of BSS (municipalities, private companies and citizens) can benefit from it if the service is well run. One of the most important obstacles to the well-functioning of the service lies in the relocation of bikes across crowded stations and the behavior of users. In this paper we address the former problem focusing on users' i tinerary choices. In particular, daily closed users' paths are particularly important because they are those paths that identify the 'faithful' users, i.e. users that tendentiously use the bike similarly day by day. We analyze in details these path choices using machine learning techniques.

2 Data

Data were daily collected from the BSS *BikeMi* on each bike trip from June 2015 to May 2018 forming an overall data set containing initially 11,771,185 records. The data set was formed by two categories of data: a group of variables related to the bike sharing process and renting, and a group of variables related to the atmospheric conditions (including pollution status). In the first group of variables we had the details about the bike sharing process and renting transaction, including client and bike ID, the type of bike (traditional or e-bike), information about check-out and check-in (station number, station slot number, date and time of rent, etc.), rental time in minutes, distance covered in metres, amount of CO2 avoided, calories consumed by cyclists, while meteorological variables (average daily temperature, average daily atmospheric pressure, precipitation condition - i.e. rain or no rain) as well as air condition indicators (daily average amount of PM10, PM25, NO2 and CO2) were included in the second group.

The results that we report below refer to a shorter period of time. In particular, we limited the investigation to the last available year - 2018 - for a total of 151 days, 50858 users and 1,457,609 transactions. Moreover, since we are interested in the set of 'faithful' users only, we pre-processed the data set by filtering out sporadic users. Specifically, we denote as faithful user a person who was active for at least

50 days (on average at least 10 days per month) only; we assume that the user was active on a specific day if he/she performed at least one transaction on that day. A further manipulation of the data set concerned the creation of the users' daily bike itineraries, since the originally released data set is centered on the single transaction enriched by identification and context variables. To this aim, we first reconstructed the transaction sequence of each client by leveraging the unique client ID field associated to each transaction. Operationally, we grouped the check-in and check-out transactions by the client ID and we ordered them chronologically. Then, to remove the circadian rhythm, we subdivided users' sequences into daily itinerary; thus, for each user, we obtain as many sequences as the number of days he/she was active. So, an itinerary travelled by a user u at day d is a sequence $< s_1, s_2, \ldots, s_{n-1}, s_n >$ of docking stations s_i, where even indexes indicate check-out stations and odd indexes denote check-in dock stations. Finally, we re-scaled the sequences so that they started from 0. For instance, the itinerary $< 10, 2, 2, 3, 67, 10 >$ was re-scaled to $< 0, 1, 1, 2, 3, 0 >$. In this way, we were able to highlight the typical behaviors of the users on riding, rather than focusing on the used docking stations. For this reason, we denoted the scaled itinerary as *itinerary type*. The entire process has been represented in Fig. 1 and resulted in a data set made up of 418,591 itineraries.



Client ID	Check-In Station	Check-In Date	Check-Out Station	Check-Out Date
101	15	08:00 15-01-2018	67	08:20 15-01-2018
101	67	17:00 15-01-2018	15	17:25 15-01-2018
101	15	07:55 16-01-2018	18	08:15 16-01-2018
101	18	08:20 16-01-2018	27	09:00 16-01-2018
101	27	16:00 16-01-2018	15	17:00 16-01-2018
101	15	08:05 17-01-2018	67	08:25 17-01-2018
101	67	10:00 17-01-2018	27	10:25 17-01-2018
101	18	10:00 21-01-2018	101	10:15 21-01-2018
101	101	12:30 21-01-2018	35	12:45 21-01-2018
101	35	16:50 21-01-2018	20	17:10 21-01-2018

Fig. 1 Process for extracting daily itineraries of the 'faithful' users from the original data set.

We associated to each itinerary the context information inherited from the corresponding transactions, i.e. the first check-in and last check-out time slots (from 1 to 24), the weekday of the rent (from 1 to 7), the client ID, the average/minimum/maximum temperature of the day, the length and the duration of the whole itinerary, the number of visited dock stations, and the rain condition. Finally, we also introduced a boolean variable which indicates whether or not the user comes back to the first dock station of the itinerary. In the first case we indicate the itinerary as "closed", while in the latter the itinerary is "open".

3 Results

In this preliminary work we first tried to study the types of itineraries on the basis of the heterogeneity and the frequency with which they occur. Table 1 shows the corresponding Gini coefficients for the first 10 itineraries in terms of frequency. For each itinerary the Gini heterogeneity coefficient (calculated using the users as categories) is obtained on the basis of the number of times each user chooses that itinerary at least once. It will be noted that heterogeneity is high; this means that the particular itinerary does not discriminate for selecting 'faithful' users. However, considering for each user the itineraries he/she chooses daily and considering a Gini heterogeneity (calculated using the itineraries as categories) threshold of 0.8 and choosing always the same itinerary type the number of users 'faithful' for a particular itinerary can be obtained (Table 2). So, for example, 2331 users can be considered preferring the itinerary type (0, 1), 455 users can be considered preferring the itinerary type (0, 1, 1, 0), and so on. Fig. 2 displays the CDF with respect to the Gini heterogeneity coefficient computed in this way and the red bar is the chosen threshold. Finally Fig. 3 shows the map of the first 20 docking stations from which most of open (red circles) and closed (blue circles) itineraries originate.

Table 1 The Gini heterogeneity index measured for the ten most common itinerary types.

Itinerary type	Gini Heterogeneity	Number of users
(0, 1)	0.999957	5603
(0, 1, 1, 0)	0.999813	5248
(0, 1, 1, 2)	0.999788	5237
(0, 1, 2, 3)	0.999798	4859
(0, 1, 2, 0)	0.999757	4817
(0, 1, 1, 2, 2, 3)	0.999551	1672
(0, 1, 1, 2, 3, 4)	0.999630	1494
(0, 1, 2, 3, 3, 4)	0.999583	1484
(0, 0)	0.998336	1426
(0, 1, 1, 2, 2, 0)	0.999004	1418

Table 2 Distinguishing itinerary types.

Itinerary Type	Number of users
(0, 1)	2231
(0, 1, 1, 0)	455
(0, 1, 1, 2)	119
(0, 1, 2, 0)	54
(0, 1, 2, 3)	31
(0, 0)	6
(0, 1, 1, 0, 0, 1, 1, 0)	1
(0, 1, 1, 2, 2, 1, 1, 0)	1

We also performed a Bayesian network (BN) analysis having the itinerary type as target variable. The question we wanted to address was the following: can the itinerary type, and perhaps its final station, be predicted knowing the weather conditions, the weekday, the starting station, the starting time, etc.? BNs implements a graphical model structure known as a directed acyclic graph (DAG), enabling an effective representation of the joint probability distribution (JPD) over a set of random variables. The structure of a DAG is defined by a set of nodes and a set of

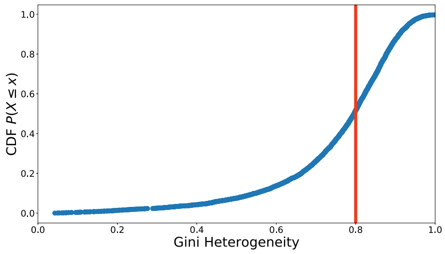


Fig. 2 CDF of the normalized Gini impurity on the sequence types of each bikers. The red vertical line indicates the threshold used to identify bikers having a distinguishing itinerary type.

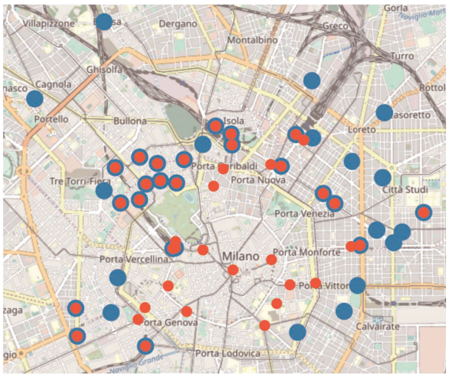


Fig. 3 Map displaying the first 20 docking stations from which most of open (red circles) and closed (blue circles) itineraries originate.

directed edges. The nodes represent random variables, whereas the edges represent direct dependencies among variables and are represented by arrows between nodes.

Fig. 4 shows a preliminary example of a BN applied to our data. A lot of arcs (edges) exist between sequence characteristics and target nodes. Using the JPD it is possible for example to predict the more probable sequence given the weather conditions (rain), the week day, the start section, the start hour, or to predict the more probable final station given the expected normalized sequence. One of the main benefit of BNs is that they provide an opportunity to conduct what if sensitivity scenarios.

4 Conclusion

This paper sought to evaluate the impact of closed itinerary choices of bike-sharing users in the overall functioning of the system. The high heterogeneity detected in the itinerary type analysis highlights the complexity of traffic forecasting. On the other hand, the first attempt to model JPD encourages the use of such models, that allow to simulate what-if sensitivity scenario. We limited the analysis to the last available year - 2018 - and only a subset of variables; further analysis and robustness tests are therefore needed. Future work will consider adding other data sources. Further information on users are available from the annual customer satisfaction survey.

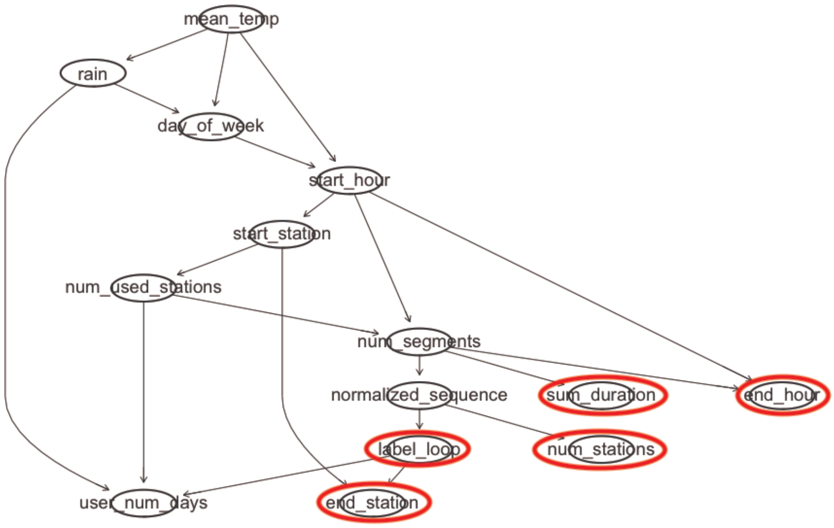


Fig. 4 BN obtained using score-based algorithms Hill-Climbing greedy search. Data-driven approach is used, no prior knowledge is imposed. Possible target nodes are highlighted in red.

We plan also to analyze neighboring stations together and the characteristics of the public transport network and of the commercial endowment of the area in which they are located. Other machine learning techniques, already applied in the BSS field [3] [4] will be taken into consideration.

References

[1] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers E. (2007). Smart cities. Ranking of European medium-sized cities, Final Report, Centre of Regional Science, Vienna UT. Available at: http://www.smart-cities.eu/download/smart_cities_final_report.pdf.

[2] Lathia, N., Ahmed, S., Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research, Part C*, 22, pp. 88- 102.

[3] Yang, Z., Hu, J., Shu, Y., Chang, P., Chen, J., Moscibroda, T. (2016). Mobility Modelling and Prediction in Bike-Sharing Systems. *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 16)*, Singapore, pp. 165-178.

[4] Manzi, G., Salini, S., Villa, C. (2019). Predicting Cycling Usage for Improving Bike-Sharing Systems. *Proceedings of the Second international conference on data science and social research (DSSR2019)*, Milan.