CAN DATA MINING HELP CAR SHARING?

Chiara Boldrini, Raffaele Bruno, Haitam Laarabi
CNR, Italy

1 INTRODUCTION

Mobility and congestion are critical concerns for every city, be it large or small, due to the economic and environmental challenges that they pose. Many analysts have advocated addressing these issues by encouraging new multimodal services and fostering the deployment of more efficient on-demand mobility services. In this work, we focus on car sharing, a mode of transportation that is gaining increasing popularity with its promise to reduce traffic congestion, parking demands and pollution in our cities. There are two main classes of car sharing services: station-based car sharing or free floating car-sharing. In the former, shared vehicles are picked up and dropped off at designated (and reserved) locations within the service area, called *stations*. In free floating car sharing, instead, cars can be picked up and dropped off anywhere within the service area, as long as parking is permitted at that location. The two approaches have each advantages and disadvantages. Free floating car sharing offers a lot of flexibility to customers, but in cities were finding a parking spot is troublesome, the reserved parking space offered by station-based car sharing may appeal more.

Managing a car sharing service is generally costly. For example, there are the costs for the infrastructure (high especially for station-based car sharing) and there are the costs for optimising the car sharing operations. There is an ongoing discussion within the car sharing community about the opportunity and convenience of vehicles redistribution. The problem that motivates redistribution is that the car sharing system tends to become unbalanced during the day, with cars that get stuck in so-called cold spot while they would be needed in hot spots. The solution would be to move these unused cars from cold spots to hot spots. Unfortunately, the cost of redistribution can be significant (two operators are needed for each redistributed vehicle) and it is crucial to perform it in an optimal way. For redistribution, and for the car sharing operations in general, we advocate the use of real car sharing data, in order to uncover, quantify, and model properties of these systems that could be used to perform this optimisation of the car sharing operations.

The goal of this work is to provide a spatiotemporal characterisation of the car sharing system, and to compare and contrast the usage patterns of two real-life car sharing systems offering a different type of car sharing service (station-based vs free floating) in two different cities. First, we discuss how the two systems can be compared on equal grounds, despite their intrinsic differences. Then, we provide an aggregate view of the temporal evolution of system usage during the day. Finally, exploiting a clustering technique, we show how the heterogeneous usage patterns (i.e., how customers pick up and drop off vehicles) at individual stations can be expressed in terms of only two classes of behaviours. This clear dichotomy is linked to the customers' typical daily life: thus, we find that there are stations that attract cars mostly in the

morning and stations attracting cars mostly in the evening, depending on the nature – residential or business – of the area. Being able to identify the class to which each station belong can be crucial for vehicle redistribution: cars should never be moved between stations of the same class. In fact, these stations tend to experience peak demand at roughly the same time of the day, hence it would be like moving a car from a hot spot to another hot spot, which is not desirable.

2 THE DATASETS

Two types of data were used for the analysis, which are described below.

2.1 Data on station-based car sharing

The dataset is composed of all the pickup and drop-off times of shared vehicles at the car sharing stations of a large European city and its surrounding area for the whole month of April 2015. Observations were taken every 2 minutes, for a total of 1,881,727 observations. Each observation records the number of available cars, the number of available parking spaces, and the status of the station (operational or in maintenance). Each station is associated with its corresponding address and GPS coordinates.

Stations that are not operational for at least 99% of the time have been filter out from the analysis, in order to avoid outliers due to stations in maintenance. In addition, for the sake of homogeneity with respect to the free floating dataset, we only consider those station that are within the main city (i.e., we discard all observations referring to the suburban area). After this initial processing, we have ~500 station in the dataset, and a total of 1,284,512 observations, that roughly cover 100km^2.

2.2 Data on free-floating car sharing

The free-floating dataset covers a time period of one and a half months between May 17, 2015 and June 30, 2015, again at a large European city. There are about 350 shared cars in a service area of more than 70km^2. Observations are taken every minute, for a total of 100,200 pick-ups and drop-offs. Here the focus is on cars rather than on stations. Thus, each observation reports the GPS position of every car that is picked up or dropped off, as well as its battery level.

3 THE ANALYSIS

In this section, we compare the station-based and free-floating services, both from the point of view of their spatiotemporal behaviour and the usage patterns observed in the two systems.

3.1 Comparing the two car sharing systems

In the station-based dataset we only know about stations and we cannot track individual cars. Vice versa, in the free-floating dataset the information we have is about available shared vehicles (stations do not exist at all). In order to be

able to compare the two systems, we need to establish the analogue of the concept of “station” in the free-floating system. The simplest way for achieving this objective is to divide the service area using a grid with cells of a given size *l*, which we have fixed to 500m. Then, all trips starting or ending within a cell will be associated with that cell. The cell becomes thus the equivalent of a station in a station-based system.

While being the only viable solution for comparing these two different systems, the approach has clearly some limitations. In fact, stations have a built-in concept of capacity: the operator is allocated a certain number of parking spots, hence no more than an equivalent number of cars can be parked at that station. This limitation is not present in the free-floating system, since customers can park wherever they want (the only limitation here is the physical parking capacity of the roads). Despite these differences, the station/cell analogy is instrumental to capture the usage patterns of the two systems, as we will show below. In the following, when ambiguity does not arise, we will use the term *station* to generically denote both the stations in the station-based system and the cell in the free-floating system.

3.2 Availability of shared vehicles during the day

Depending on the structure and dynamics of the car sharing systems, each station or cell can see a different number of parked cars during the day. For example, there can be stations with at most one or two cars parked during the day, while others can have six, seven or more. In Figure 1 we plot the distribution of the maximum number of cars parked at stations in order to illustrate this heterogeneity. Clearly, the larger the cell size, the higher the maximum number of cars that can be parked simultaneously. In general, there is a higher variability in the free-floating system than in the station-based system. This is partly due to the fact that stations are more easily saturated due to the fixed upper bound on their maximum capacity (typically, never more than 7 parking spaces).

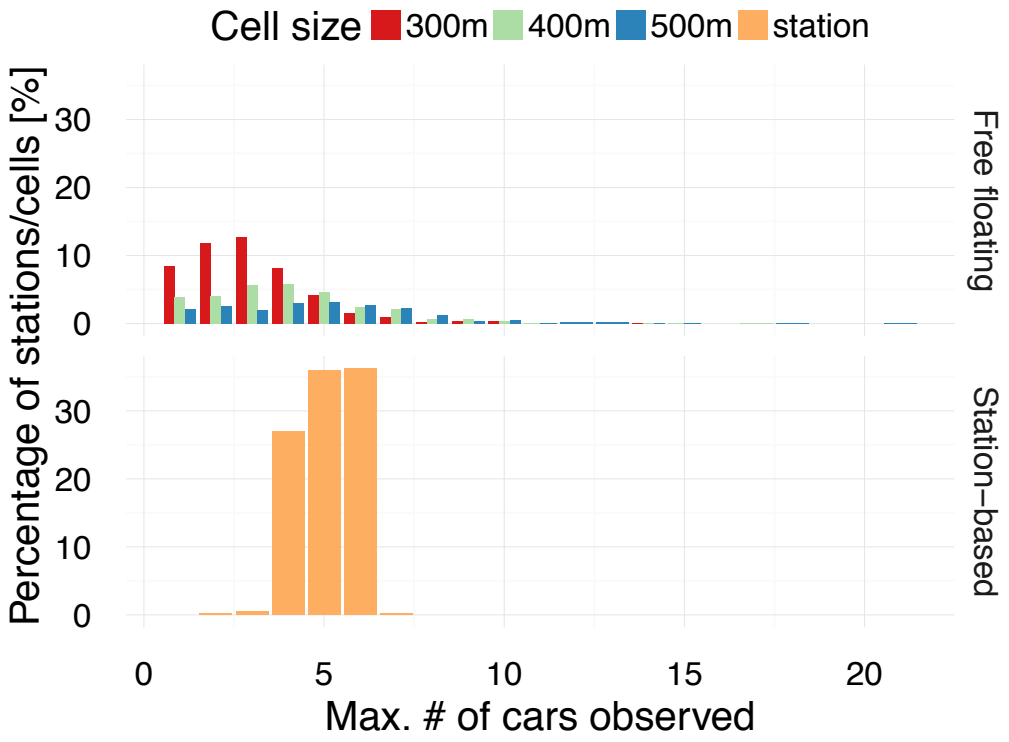

Figure 1: Maximum number of cars parked at stations/cells

In order to get rid of the effects due to the heterogeneity in the two systems, in the following we focus on the *availability* of shared vehicles at a station/cell, which we define, after focusing on a specific time t, as the ratio between the number of cars currently at the station/cell and the maximum number of parked cars registered at the station/cell during the observation period. The availability metric allows us to compare stations of different size more fairly than when looking at the absolute numbers of cars parked during the day.

We can use the availability to capture the typical daily usage of the systems as a whole, i.e., to understand at which time of the day the number of rented shared vehicles is higher/lower. This is shown in Figure 2, where we have also distinguished between weekdays and weekends. In both datasets, during weekends the renting tends to start later in the morning (since people typically sleep more). The Mon-Fri and Sat-Sun usage is markedly different. In weekdays, the availability roughly plateaus towards the higher/middle end during the working hours in a weekday. This might be due to customers using less the car sharing service (e.g., because they are confined to their offices) or to the injection of vehicles from the suburban areas (the latter only holds for the station-based system). Vice versa, during weekends there is a plateaux of low availability in the afternoon, meaning that many vehicles are in use.

Among the many similarities, a notable difference between the free-floating car sharing system and the station-based one is that, in the former, the availability of cars never rises significantly during the day. This means that

once the vehicles start being used, they keep being used at about the same rate during the day.

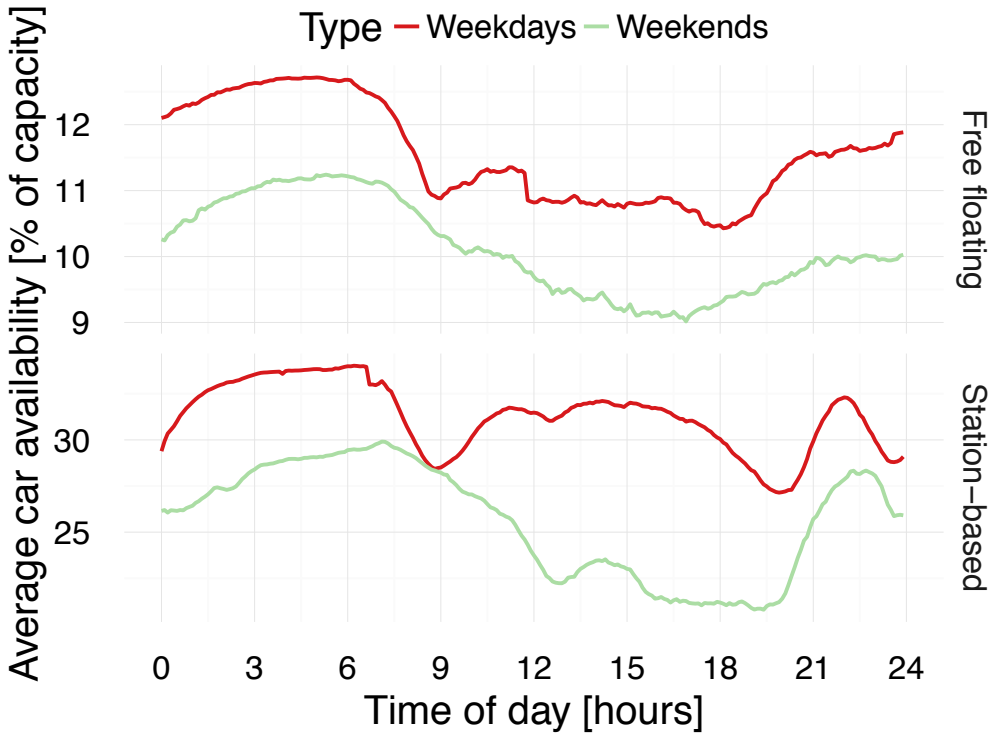

Figure 2: Average daily availability

3.3 Parking times

The average car availability is a by-product of the average in- and out-flow at each station. This means that we can have a high or low availability regardless of whether individual cars remain parked at the station for a long or a short time. For example, we can have an average availability of one car either when a single car remains parked for days at the same location or when for every newly picked up car, another one is dropped off. From the car sharing operator's point of view, cars that remain parked for a long time are money drains: they are not used where they are and are not where they might be used (e.g., at hot spots). Let us take a look at the time vehicles spend parked at stations/cells in our datasets, in order to understand whether the two systems under study might suffer from the above-mentioned problem.
In Figure 3 we plot the distribution of parking duration for the station-based and free-floating systems. The difference in parking duration is noteworthy. In the free-floating system, the duration is significantly longer than in the station-based system. This may be due to cars that get stuck in peripheral areas, in which the demand is low and where thus nobody picks them up.

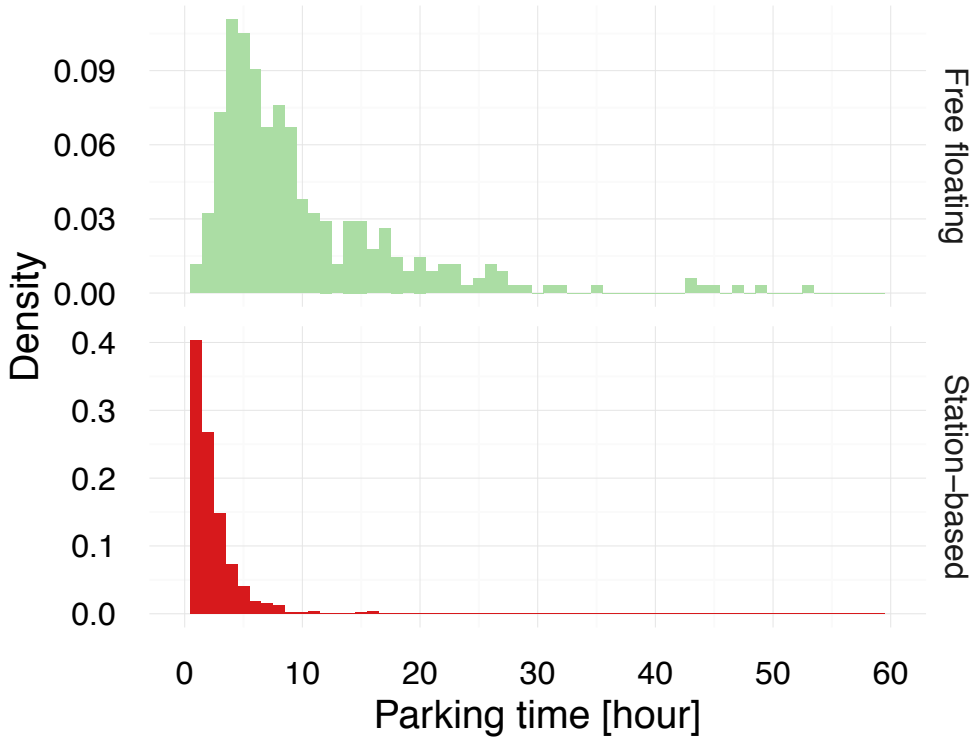

Figure 3: Time cars remain parked at stations/cells

3.4 Usage patterns: the invariants

In Section 3.2 we have observed pretty clear common daily trends in the aggregate behaviour of availability at stations, but we have also discussed in Sec. 3.1 and 3.2 that stations can be very heterogeneous, especially if we only consider absolute metrics. In the systems we are studying there are a lot of stations (~500) or cells in the free floating (~350) and it may be difficult to make sense of their heterogeneity while still capturing hallmarks in their individual behaviours. In order to address this point, in the following we describe a way for classifying stations and cells based on how they are used by the customers of the car sharing service.

The data mining approach we use is the following. First, we obtain a time series similar to the one in Figure 2 for each individual station and we normalise it dividing by the average availability at the station. Then, for each station pair, we measure how "close" their two times series are using the Dynamic Time Warping (DTW) [Esling and Agon 2012] technique. The DTW technique is able to ignore the effects of minor shifts in the time series. After running the DTW algorithms, we have, for each station pair, a measure of their distance. Then, we feed these distances to the Partitioning Around Medoids (PAM) clustering algorithm [Kaufman and Rousseeuw 2009] in order to create *k* groups of stations with a similar behaviour (i.e., with small DTW distance between them). We test different values of *k*, then we rely on the Silhouette method [Kaufman and Rousseeuw 2009] for selecting the most informative *k*.

After performing all the above steps, we find that the "best" clustering is obtained with *k*=2. This means that the algorithms identify two main usage patterns for the stations and cells in the two car sharing systems. In order to understand what is the typical behaviour in these two clusters, we plot the availability at each station (which we have studied in Section 3.2) normalised by its daily average. In this way, even stations/cells with very different average availability can be compared on equal ground. Then, we take all the stations/cells belonging to the same cluster in each dataset, and we average their normalised availability value in each temporal bin. The resulting typical cluster availability (i.e., an aggregate normalised availability observed within the stations/cells of the clusters) is shown in Figure 4 and Figure 5. Due to their clear inverse patterns with either maximum availability at night or during the day, we have renamed the two clusters Night Peak and Mid Of Day Peak clusters. The similarity of the behaviour of the two clusters in the two datasets is striking. We observe the two opposite patterns in very different datasets (station-based vs free-floating) and in two different cities. Please note that the analysis has been carried out separately on the two datasets, so no cross influences were possible.

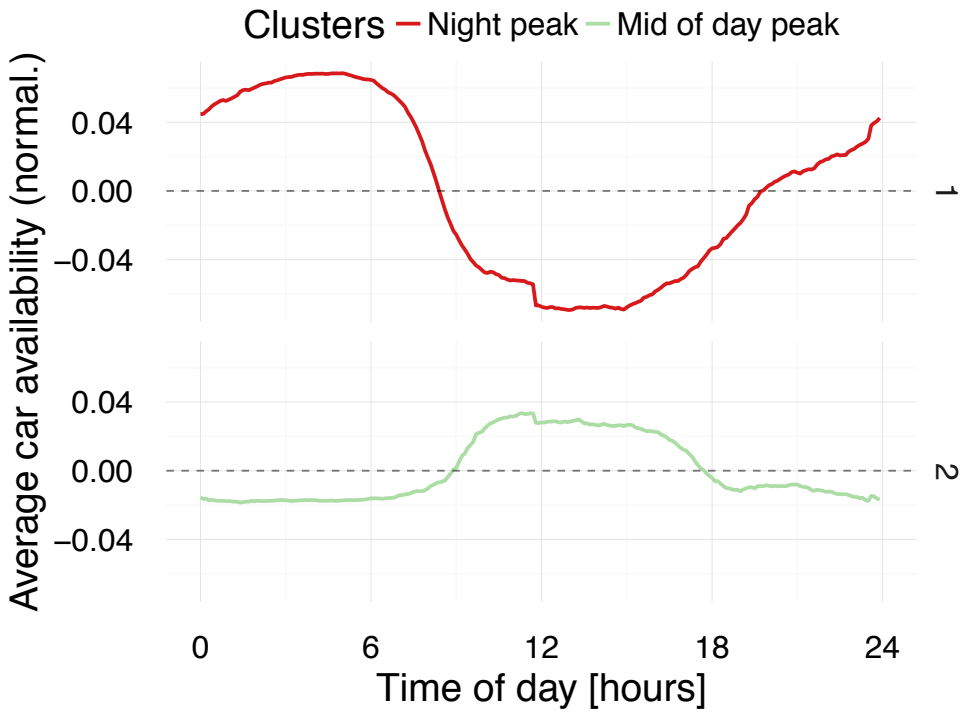

Figure 4: The two usage patterns in the free-floating system

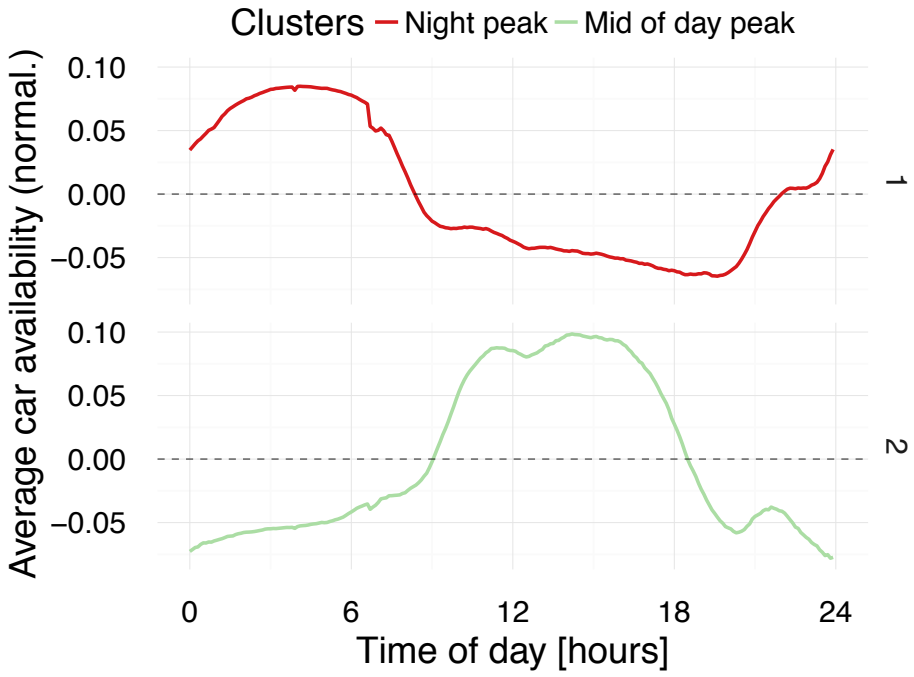

Figure 5: The two usage patterns in the station-based system

It is interesting to also have a look at the geographical distributions of stations focusing on the cluster they belong to. For the station-based car sharing (Figure 6) we observe a high degree of colocation among stations featuring the same usage pattern, with a strong separation between areas showing a night peak and areas showing a mid-of-day peak. Vice versa, in the free-floating system, the situation is blotchier, with small geographical clusters of stations with the same behaviour alongside other small geographical clusters featuring the opposite patterns. The size and location of these geographical clusters are linked to the nature (e.g., commercial or residential) of the areas of the city, hence it is not something the car sharing operator can control. However, being able to correctly identify and localise stations with similar usage pattern is crucial for the optimising car sharing operations. Take for example vehicle redistribution. Redistribution is a last-resort measure the car sharing operator has to implement in order to smartly balance the available cars in the car sharing system. In fact, it is a well-known problem, also confirmed by our clustering analysis, that stations tend to synchronise in groups. For example, referring to the station-based car sharing in Figure 6, it is clear that customers located at the very centre of the map may find it very hard to rent a shared car late in the evening (a time with lower-than-average availability), regardless of whether they are willing to walk a few hundreds of meters or not. In fact, all nearby stations are green in Figure 6, hence they will all show a low availability. This lack of vehicles at certain times of the day in a given area may majorly affect the car sharing operator (both in terms of loss of revenue and loss of customer satisfaction). Hence, the operator is motivated to redistribute vehicles (this is typically done by several pairs of employees, one driving the other one to where there is an abundance of cars to be picked up and moved). This process is costly for the car sharing operator and there is also a big question mark related to the station pairs that have to be the target

for redistribution. Luckily, the proposed clustering technique can be exploited to aid the redistribution process. For example, a simple approach to redistribution is to focus on stations belonging to different clusters, in order to be sure that they experience availability peaks at opposite times of the day, hence the surplus cars at one station can be used to feed the other one, and vice versa, depending on the time.

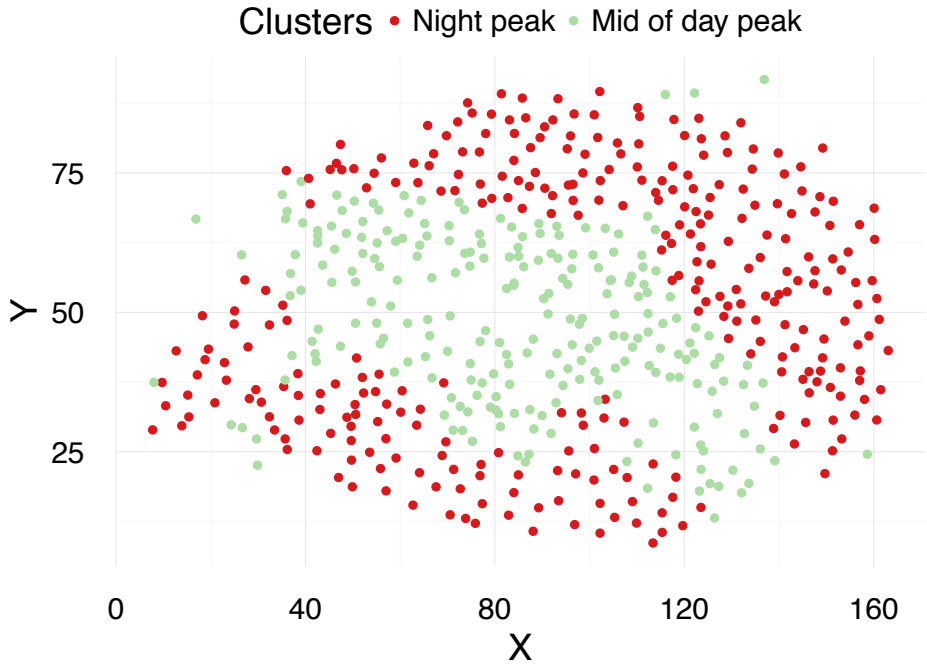

Figure 6: Clusters on the map (station-based)

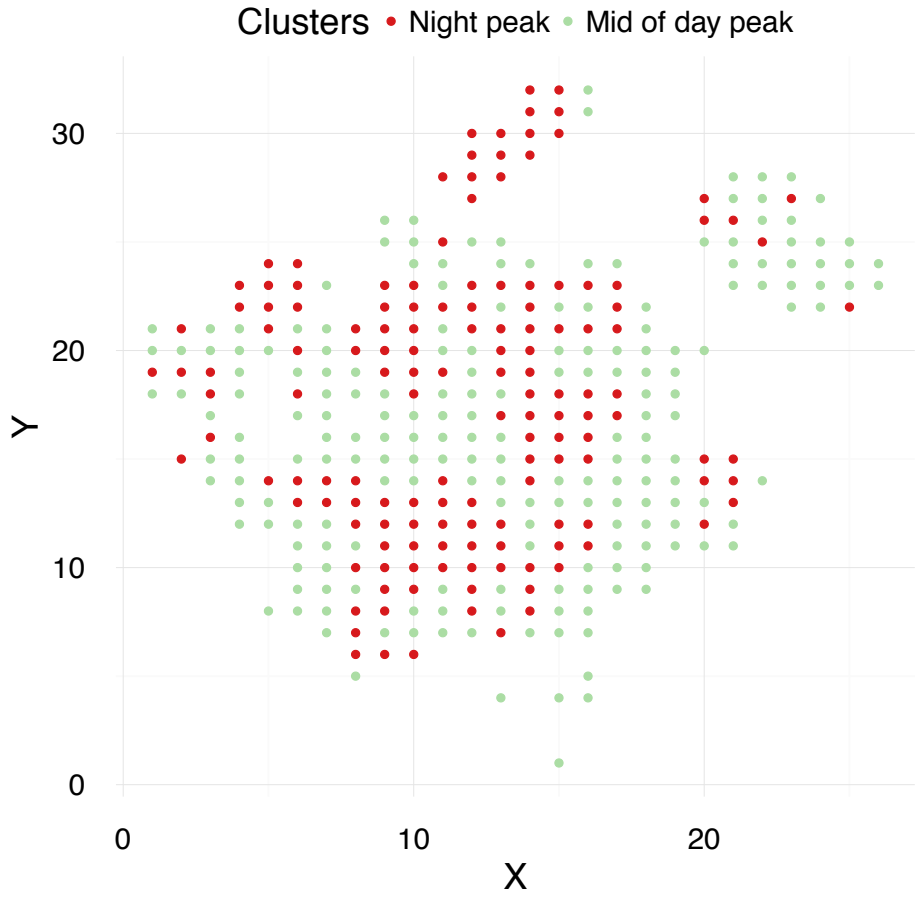

Figure 7: Clusters on the map (free-floating)

4 CAR SHARING OF THE FUTURE

We have often discussed the key role played by a smart vehicle redistribution in the overall quality of service provided to the customers. However, despite the optimisation and improvements (possible, e.g., using the data-driven approach discussed in the paper) moving vehicles one at a time from cold spots to hot spots in a city remains an expensive and inefficient operation (two employees needed per relocated car). It has been suggested in the literature (Weikl and Bogenberger 2013) that the impact of relocation could be mitigated by employing a user-based relocation, providing incentives (in the form of discounts or free rides) for customers willing to modify their route to pick up a car in a cold spot and bring it to a hot spot.

The ESPRIT project, started in 2015, aims at improving the vehicle redistribution process in an innovative way. Specifically, it calls for a drastic redesign of shared vehicles (Figure 8) such that these vehicles can be driven in a train of up to 8 cars (by an employee of the car sharing service) or up to 2 (by the customers). Redistributing ESPRIT vehicles becomes thus more efficient and cheaper for the operator, and the customers experience a better

quality of service. ESPRIT vehicles will be electric and customers with a standard driving license will be able to drive them.

Figure 8: ESPRIT stackable vehicles

5 RELATED WORK

Current knowledge on car sharing systems derives mainly from surveys (Shaheen and Cohen 2015, Schwieger et al. 2015). In surveys, car sharing operators and members are interviewed directly, providing an invaluable understanding of the underlying car sharing systems. However, carrying out surveys is very expensive and the activity does not scale very well. A complementary alternative is to use timestamped and geotagged digital records, that can be analysed on a large scale using data mining techniques.

Something similar has been already done for bike sharing, providing extremely useful insights into the usage patterns of the system, as in O'Brien et al. 2014 and Sarkar et al. 2015. Clearly, car sharing and bike sharing are two different sharing systems, and the findings emerged from one cannot be directly applied to the other. For example, not all trips are suitable to biking, as well as not all trips are convenient for cars. The different motivation behind customers' choice may significantly impact the way the two systems are used. Thus, the goal of this work has been to carry out an analysis similar to that in O'Brien et al. 2014 and Sarkar et al. 2015 but for car sharing systems.

6 CONCLUSIONS

In this work, we have investigated how mining car sharing datasets can provide useful insights into the dynamics of the car sharing system. To this aim we have focused on two datasets. The first one covers the pickups and drop-offs of shared vehicles over a month in a station-based car sharing system. The second one contains the pickups and drop-offs of shared vehicles in a free floating system. By mining these datasets, we have characterised some spatiotemporal trends in the systems. First, we have discovered common daily peaks in availability (e.g., at night or when people are at work) but we have also observed a general heterogeneity in individual stations/cells as far as parking times and numbers of cars that are parked on average. Motivated by this results, we set out to uncover invariants, if any, in the ways the two systems are used. Exploiting standard data mining

techniques for the analysis of time series, we have found that, despite their heterogeneity, stations and cells can be classified in two simple groups based on the way they are used by the customers. Specifically, we have identified a cluster of stations that feature an above-average availability during the day (typically, commercial and business areas) and a cluster of stations with above-average availability at night (typically, residential areas). Interestingly, we have found the exact same cluster with the exact same behaviour in the two different datasets. Being able to attaching a “smart” label to each station using this technique can be very useful for optimising the operations (in particular, vehicle redistribution) of the car sharing service.

7 ACKNOWLEDGMENT

This work was partially funded by the ESPRIT project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653395.

BIBLIOGRAPHY

Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 12.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

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.

Sarkar, A., Lathia, N., and Mascolo, C. (2015) Comparing cities’ cycling patterns using online shared bicycle maps, Transportation, pp. 1–19.

Schwieger, B., Victorero-Solares, P., & Brook, D. (2015). Global carsharing operators report 2015. Team Red, Tech. Rep.

Shaheen, S., & Cohen, A. (2015). Mobility and the Sharing Economy: Impacts Synopsis – Spring 2015. Transportation Sustainability Research Center, University of California, Berkeley, Tech. Rep.

Weikl, Simone, and Klaus Bogenberger. "Relocation strategies and algorithms for free-floating car sharing systems." IEEE Intelligent Transportation Systems Magazine 5, no. 4 (2013): 100-111.