




Weak signals in the mobility landscape: car sharing in ten European cities

Chiara Boldrini^{1*} , Raffaele Bruno¹ and Mohamed Haitam Laarabi¹

*Correspondence:
chiara.boldrini@iit.cnr.it
¹IIT-CNR, Pisa, Italy

Abstract

Car sharing is one of the pillars of a smart transportation infrastructure, as it is expected to reduce traffic congestion, parking demands and pollution in our cities. From the point of view of demand modelling, car sharing is a weak signal in the city landscape: only a small percentage of the population uses it, and thus it is difficult to study reliably with traditional techniques such as households travel diaries. In this work, we depart from these traditional approaches and we leverage web-based, digital records about vehicle availability in 10 European cities for one of the major active car sharing operators. We discuss which sociodemographic and urban activity indicators are associated with variations in car sharing demand, which forecasting approach (among the most popular in the related literature) is better suited to predict pickup and drop-off events, and how the spatio-temporal information about vehicle availability can be used to infer how different zones in a city are used by customers. We conclude the paper by presenting a direct application of the analysis of the dataset, aimed at identifying where to locate maintenance facilities within the car sharing operation area.

Keywords: Car sharing; Smart transportation; Urban computing; Data mining

1 Introduction

Automobile transportation has been one of the main drivers of the population growth and increasing wealth that have characterised the last two centuries [1]. Thanks to cars, people have had greater access to jobs, goods, services. However, these benefits have not come for free. The price paid for our increased mobility has been huge in terms of environmental pollution, city congestion and resulting health issues. We are now at a turning point for personal mobility systems: policy makers and citizens share the common idea that it is time to rethink the way we move. There are three main driving forces behind this personal mobility revolution: smart transportation, sharing economy, and green vehicles, all tightly intertwined. The departure from ownership mindset to usage mindset will make it possible to have significantly fewer vehicles in our cities. The implications are that we can save space (public parking space and private garage space) and use it for something with increased added value than to host idle cars for hours (a private car is used only 5% of its available time, corresponding to 72 minutes per 24 hrs [2]). This usage mindset will also allow people to rent the car size most appropriate to their daily needs, thus implementing the Mobility-as-a-Service concept. Since the average vehicle has only

around 1.5 occupants [3, 4], people can refrain from buying a car able to address the extreme case of personal mobility (e.g., moving a whole family for a vacation) and instead use two-seaters, which are more suitable for everyday commuting. On occasion, they will be able to rent larger vehicles if needed. The virtuous mobility cycle is completed with the switch to electric vehicles, which allow for a drastic reduction in the carbon footprint of personal mobility.

In this framework, car sharing is emerging as one the most promising examples of Mobility-as-a-Service [5]. The general idea of car sharing is that the members of a car sharing system can pick up a shared vehicle of the car sharing fleet when they need it. Different operators may implement different pickup/drop-off policies. In station-based systems, members can only pick up and drop off vehicles at designated locations called stations, as in the Autolib system in Paris. If the service is two-way (e.g., Zipcar, MODO), people are asked to bring back the vehicle to the station where they initially picked it up. Otherwise, the service is called one-way. One-way services are definitely the most popular among customers thanks to the flexibility they provide. Examples of one-way car sharing are Autolib, Ha:Mo ride, CITIZ. One-way services can drop altogether the concept of station: this is the case of so-called free floating car sharing—such as Car2go, DriveNow, Enjoy—whose customers can pick up and drop off vehicles anywhere within a predefined operation area.

Car sharing is a *weak signal* in the city landscape: the fraction of people relying on car sharing for their daily trips is rapidly increasing but it is still in the order of single digit percentage points in the best cases [6]. So far, car sharing has been mostly studied through surveys and direct interviews with its members [5, 7]. In addition, car sharing is typically not accounted for in households travel diaries periodically collected by city administrations. Even if it were, the limitations of travel surveys are widely acknowledged, and range from their inability to capture changes in the routine travel behaviour to their underestimation (because of underreporting from people) of short, non-commute trips [8]. Moreover, running a survey is very expensive if one wants to capture a statistically meaningful sample.

Cities have been considered kaleidoscopes of information since a long time [9] but the extent to which this is true has reached new heights now that a myriad of electronic devices have weaved into its fabric. From the car sharing perspective, this means that we can now know exactly when and where cars are available, and we can observe shared vehicle flows *as they happen* in the city. This knowledge opens up a new avenue of research that goes in the direction of the new science of cities and urban computing: using data and electronic devices to extract knowledge and to improve urban solutions. Along these lines, the goal of this paper is to stimulate a discussion on how to apply urban computing ideas to the car sharing domain. To this aim, we exploit the availability of public, web-based data about free floating car sharing in 10 European cities (whose main characteristics are summarised in Table 1 for the convenience of the reader) and we carry out an analysis with the following objective in mind: to understand what mining this kind of data can bring to cities and to car sharing operators alike. The main contributions of this study can be summarised as follows:

- We perform an explanatory analysis of the car sharing demand as a function of the sociodemographic and urban fabric (i.e., number, heterogeneity, and category of Foursquare Points of Interests—PoI) indicators associated with the cities of Milan,

Table 1 General information on the 10 cities (area: km², population density: people/km²; education: persons aged 25-64 with ISCED level 5, 6, 7 or 8 as the highest level of education; mean daily temperature: minimum/maximum average daily temperature over the year). Source: Urban Audit Database [14] & Wikipedia

City	€GDP/capita	Population	Area	Pop. Density	Education	Mean daily C°
Amsterdam	46,952	853,312	165.76	5147.876	254,000	3.4/17.6
Berlin	35,627	3,520,031	891.68	3947.639	752,300	0.6/19.2
Florence	31,547	382,929	102.32	3742.465	59,627	6.5/24.6
Copenhagen	70,183	559,440	88.25	6339.263	152,817	0/17
Milan	87,786	1,368,590	181.67	7533.380	224,256	2.5/23.6
Munich	46,377	1,450,381	310.70	4668.107	371,200	0.3/19.4
Rome	55,385	2,874,529	1287.36	2232.887	415,766	7.5/24.5
Stockholm	81,395	939,238	187.16	5018.369	234,787	-1.7/18.8
Turin	74,725	886,837	130.17	6812.910	109,314	2.9/20.8
Vienna	58,140	1,867,582	414.87	4501.608	232,009	1.2/21.7

Rome, and Turin.^a While a single explanatory pattern does not emerge across the cities, they share indeed several similarities. In fact, their car sharing demand is positively associated with high educational attainment (all Italian cities) and negatively correlated with commuting outside of the municipality area (Milan, Rome). These findings confirm the conclusion of the most recent sociodemographic surveys about car sharing services [10–13], but at a much finer spatial granularity and without relying on expensive and time-consuming interviews/questionnaires. With regards to the urban fabric indicators, the only PoI category that seems to have a statistically significant effect on car sharing demand is that of nightlife-related activities, suggesting that leisure is the most typical trip purpose.

- We take into consideration several approaches to demand forecasting, and we evaluate which are the best performing when it comes to car sharing pickups/drop-offs forecasting. Our results show that Random Forest yields consistently better results than simple average-based forecasting, time series forecasting, vanilla neural networks, and a popular custom approach proposed in the literature. However, prediction quality is in general quite good, even with the simplest solutions.
- Four distinct car availability temporal patterns can be recognised in the cities considered in this study. We have labelled them *day*, *night*, *neutral*, and *high-intensity* behaviours, based on when they exhibit their peak availability and on the intensity of this peak. We also show that these patterns tend to be spatially autocorrelated, i.e., neighbouring cells are likely to feature the same behaviour.
- Motivated by the importance that customers place on the cleanliness of vehicles, we propose a simple approach to the effective deployment of car sharing maintenance facilities. We show that including the airport zone in the operation area and locating maintenance facilities there is a simple yet effective strategy to reduce the maintenance trips carried out by the car sharing workforce.

2 Related work

In the following we provide a brief overview of the most relevant works in the area of data science for car sharing, data science for transportation systems in general, and data-driven car sharing operation models.

2.1 Knowledge mining from survey data

Until recently, knowledge about car sharing systems has been mostly acquired through surveys, in which car sharing operators and members are interviewed. The main goal of these studies is to characterise the sociodemographic profile of car sharing users, as well as investigating the reasons behind their choices and the impact that car sharing has had on their mobility behaviour. In 2005, Millard-Ball [15] presented one of the first comprehensive sociodemographic analysis of station-based car sharing in North America, highlighting a few key demographics indicators that will constantly reappear also in analyses of more recent car sharing solutions. After interviewing 978 US and 362 Canadian car sharing members through a web-based survey, Millard-Ball reports that car sharing members are typically young (25–44 year old), with high income and well-educated. They live in small households, often with no private cars. This survey does not support the finding, often presented in the related literature, that car sharing members are typically male. Recreational trips, shopping-related trips, and personal business trips are by far the most popular trip purpose for the respondents. In 2010, these findings are substantially confirmed by [16] for Europe, with the interesting addendum that car sharing customers tend to have season tickets for public transport more than the general population.

Considering that free floating car sharing is a recent addition to the car sharing domain (e.g. Car2go was founded in 2008, and started a significant expansion only in 2011), in the following we overview recent surveys [10–13] focusing specifically on the free floating modality. Kopp et al. [10] recruited 204 males between 25 and 45 years of age living in the cities of Munich and Berlin, Germany. 109 were free floating car sharing members (DriveNow), 95 did not use car sharing. Respondents were asked to use a custom-built app to track their trips and to specify the trip purpose and the mode of transport. The findings of this study confirm previous results obtained for station-based car sharing: free floating car sharing members have higher levels of education, higher income, fewer private cars, and more public transport subscriptions with respect to non-members. The study also highlights that car sharing members typically live in denser neighbourhoods, and are more intermodal and multimodal in their mobility behaviour. No statistically significant difference in trip purpose was detected between members and non-members: most trips are work-home trips (57%), leisure (19%) and shopping/errands (13%). Giesel and Nobis [12] perform a similar study for DriveNow and Flinkster users in Munich and Berlin, reporting substantially the same findings.

Becker et al. [13] directly compare free floating and station-based car sharing members in the city of Basel, Switzerland. While the sociodemographic profile of car sharing is largely the same between station-based and free floating and substantially the same as that pictured in the previous literature, free floating car sharing members in Basel differentiate from their station-based counterpart in that they tend to use public transportation less. The authors remark that free floating car sharing may act as a complement to public transportation, filling the service gaps that their users might experience. The trip purposes of free floating car sharing members is quite diversified, but mostly involve visiting, shopping, and commuting, while station-based car sharing mostly covers leisure trips, goods transport, and shopping.

Wittwer and Hubrich [11] discuss the findings from a two-stage survey carried out in Hamburg, Germany, among Car2go members. The first stage of interviews took place in 2011, at the beginning of the Car2go service in the city, the second stage was run in 2016,

when the service had been in place for a few years. From the sociodemographic standpoint, the 2011 and 2016 cohorts substantially share the same profile: largely man, 24–49 years old, high income, low car access, often with public transport season tickets. 2016 active users overwhelmingly rely on car sharing for leisure trips (72%), but significant percentages also use it for shopping and errands (50%) and for work/education trips (42%).

Based on the above overview, we can conclude that survey findings are consistent as far as the sociodemographic profiles of car sharing users are concerned, while contrasting results have been obtained regarding car sharing trip purpose and relationships with public transportation. In Sect. 4 we will discuss our findings in light of the above results.

2.2 Knowledge mining from digital data

The understandings and advancements brought about by the works described in Sect. 2.1 are invaluable, but the collection of survey data is expensive, time consuming, and does not scale. Typically, travel surveys cover a relatively small sample of all the trips of interest (because the number of participants as well as the observation period are typically quite limited). Furthermore, it is a well-known problem that travel surveys often tend to underestimate the number of trips and to show a bias in the types of trips being reported [8]. For these reasons, in this work we depart from this approach and we exploit public, web-based, digital records, whose geotagged and time-stamped variety of data can be analysed with data mining techniques. These data can be collected for a possibly very long time with minimal effort, and can provide geographically diverse and almost continuous measurements of the systems under study.

In the related literature, the works by Schmöller et al. [17] and Willing et al. [18] are mostly focused on the external factors that may influence car sharing demand. In particular, Schmöller et al. [17] highlight the role played by weather and demographics on the car sharing demand, while Willing et al. [18] tackle the problem of understanding if Points of Interest (PoI) in each city can be used as demand predictors. Differently from Willing et al. [18], in this work we study the effects of PoIs taking into account collinearity of predictor variables and selection bias in p-value computation, resulting in a much smaller effect of PoIs on the car sharing demand. The same considerations apply for Schmöller et al. [17]. Our work is also close to [19], which considers free-floating car sharing in multiple cities. However, Kortum et al. [19] focus on the growth rate of free floating car sharing rather than on the characterisation from the supply side point of view. Finally, in [20], we have presented an analysis of station-based car sharing in a single city. The analysis in [20] is more oriented to issues related to the presence of stations (their capacity, how their behaviour can be mathematically modelled using queueing theory, etc.) and suffers from the lack of vehicle identifiers in the dataset. The technique used in [20] for detecting station usage is adapted here to the free floating case, but the analysis presented here is richer, because richer is the dataset extracted from the free floating car sharing operator.

Several works in the literature also focus on the problem of demand forecasting, which we tackle in Sect. 5. This is typically done in conjunction with a proposal regarding vehicle relocation, which involves deciding how to proactively relocate shared vehicles in the operation area in order to meet the future demand. We can group forecasting proposals in three different classes, based on the approach they rely upon. There is a group of papers whose forecasting approach relies on techniques for *time series* forecasting. Wang et al. [21] leverage selective moving averages, Holt's model, Winter's model as well as Tabu

Search heuristics for forecasting the demand in a car sharing service in Singapore. No prediction evaluation is carried out in the paper. Müller and Bogenberger [22] focus on the city of Berlin and investigate how to predict future bookings using seasonal ARIMA model and exponential smoothing with Holt-Winters-Filter. The second class of forecasting methods are those coming from the *machine learning* domain. Cheu et al. [23], for example, compare the forecasting performance of a neural network approach against that of Support Vector Regression, and find that the former provides better predictions. Neural networks have been later used also in [24–26]. The third class of forecasting approaches relies on *custom solutions* specific for the problem at hand. Boyaci et al. [27], for example, compile origin-destination matrices by simply averaging the observations for different hours of the day, days of the week, and months of the year from real car sharing data. Weikl and Bogenberger [28] devise a prediction algorithm based on finding clusters of behaviours for daily timeslots. In all the above works, the evaluation of forecasting performance is carried out considering only a single city.

A preliminary analysis [29] of this dataset has been presented at KNOWMe'17, an ECML-PKDD workshop without copyrighted proceedings. In this extended version, we have added the sociodemographic study (Sect. 4) and the demand forecasting analysis (Sect. 5). In addition, we have added the analysis of the spatial autocorrelation of vehicle availability clusters (Sect. 6).

2.3 Knowledge mining for other transportation systems

From the methodology standpoint, this work is close to [30–33], in which bike-sharing, rather than car-sharing, systems have been analysed. Due to the different nature of the two systems, people use them differently, hence the results obtained for bike sharing systems cannot be applied directly to car sharing. However, similar methodologies can be exploited, e.g., to group stations based on how they are used by the customers.

This work is also orthogonal to the research efforts in the area of car pooling/ride sharing [34, 35]. The idea of car pooling/ride sharing is that people may share a vehicle (be it a private or public vehicle, e.g., a taxi cab) to perform their trips. Works in the area of car pooling typically focus on the amount of rides that can be shared, based on the historical or real-time trajectories of users, hence their focus is very different from that of this work.

2.4 Operation models for car sharing

As one of the pillars of a smart transportation system, car sharing has recently been the subject of extensive research from the operational standpoint. The research activity on this area has focused both on short and long term strategic decisions. The latter involves problems like planning the station/parking infrastructure [27, 36, 37] or planning the recharging infrastructure. The former is focused on decisions such as when and how to redistribute shared vehicles [38–41] or when and how to recharge them [42, 43].

To address the above problems, optimisation frameworks and operational decision tools for car sharing systems have been studied in the literature, but the proposed solutions have often been evaluated either on simulated scenarios [44, 45] or using as input the demand (in terms of origin/destination matrix) obtained from surveys [36, 46]. On the contrary, the availability of a statistical characterisation of the general properties of real car-sharing systems, as well as a precise understanding of their emerging trends, is essential to both researchers and operators in order to design more effective decision support tools, and for

the calibration and validation of simulations of car sharing systems. Thus, a data-driven analysis as that presented in this paper can be exploited to both drive and evaluate solutions for the supply-side of car sharing.

3 The dataset

The dataset comprises pickup and drop-off times of vehicles in 10 European cities for one of the major free-floating car sharing operator (Table 2). For nine of these cities, data has been collected between May 17, 2015 and June 30, 2015. For Munich, data covers the period from March 11, 2016 to May 12, 2016. The data has been collected every 1 minute using the available public API, which yields responses in the form of JSON files. Errors in the data collection process are due to technical problems on the booking website, in which cases corrupted entries have been discarded from the dataset. Each entry in the dataset describes the longitude-latitude position of available shared vehicles in the car sharing system, plus additional information. Each entry in the dataset has the following structure:

$$\langle \text{vin}, \text{date_time}, \text{lon}, \text{lat}, \text{fuel}, \text{interior}, \text{exterior}, \text{engine} \rangle, \tag{1}$$

where *vin* is the unique identifier of a vehicle, *date_time* contains the date and the time at which the available vehicle has been observed, $\langle \text{lon}, \text{lat} \rangle$ are the geographical coordinates, $\langle \text{interior}, \text{exterior} \rangle$ refer to the cleanliness of the vehicle, *engine* specifies where the vehicle is electric or not. Due to faulty GPS systems, the reported coordinates may be inaccurate. For this reason the dataset has been preprocessed and coordinates that are manifestly invalid (e.g., cars available in different countries) have been discarded. Data preprocessing and analysis has been carried out in R.

Given the nature of our dataset, movements of cars have to be inferred from their unavailability during a certain time frame. Thus, when a car disappears from location A to later reappear at location B, we assume that the car has been picked up for a trip. We have no explicit way for distinguishing between regular customer trips and maintenance trips (e.g., cars that have been picked up by the car sharing operator for cleaning or repairing), as we simply observe a car disappearing from the map.

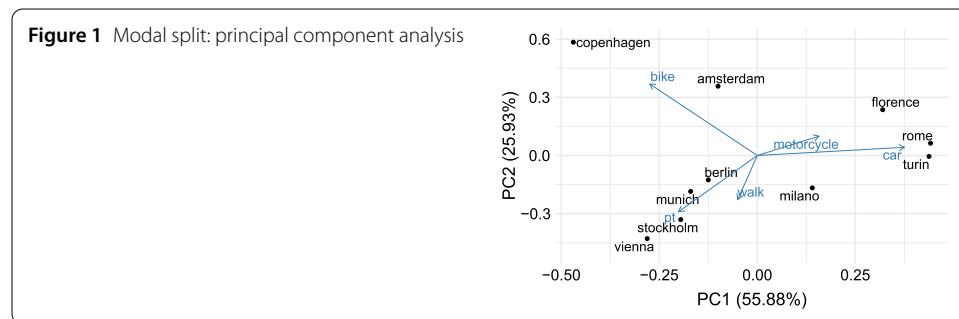
In order to understand the main characteristics, in terms of mobility, of the ten cities in which the car sharing system under study is operating, we have extracted information (summarised in Table 3) from the Eurostat’s City Urban Audit database [14]. Figure 1 summarises the main transportation mode in each city as resulting from the Principal Component Analysis applied to the reported modal share. We can identify three main classes

Table 2 Summary of dataset

City	# Trips	# Cars	Op. Area [km ²]	Cars/km ²	Starts	Ends	Duration
Amsterdam	49,901	349	59	5.880	2015-05-17	2015-06-30	45 days
Berlin	223,044	981	160	6.115	2015-05-17	2015-06-30	45 days
Florence	18,944	198	61	3.268	2015-05-17	2015-06-30	45 days
Copenhagen	12,168	194	41	4.712	2015-05-17	2015-06-30	45 days
Milan	156,080	686	120	5.737	2015-05-17	2015-06-30	45 days
Munich	81,862	499	89	5.592	2016-03-11	2016-05-12	63 days
Rome	99,515	584	90	6.480	2015-05-17	2015-06-30	45 days
Stockholm	15,612	250	36	6.871	2015-05-17	2015-06-30	45 days
Turin	25,091	299	53	5.646	2015-05-17	2015-06-30	45 days
Vienna	144,474	829	110	7.569	2015-05-17	2015-06-30	45 days

Table 3 Modal share in the 10 cities

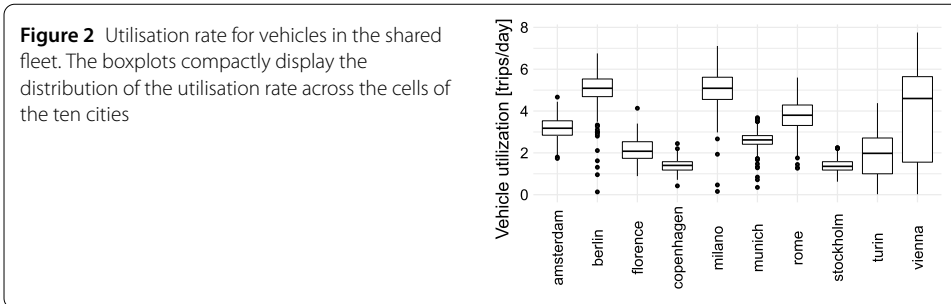
City	Bike (%)	Walking (%)	Public Transport (%)	Motorized (%)
Amsterdam	22.00	4.00	30.00	44.00
Berlin	9.90	7.20	41.00	38.50
Florence	4.04	7.68	19.10	66.82
Copenhagen	36.12	5.84	29.23	26.40
Milan	2.05	8.82	35.20	51.53
Munich	10.10	6.60	45.20	37.40
Rome	0.25	7.37	23.30	68.34
Stockholm	7.00	15.00	43.00	33.00
Turin	1.47	10.30	24.20	64.07
Vienna	7.00	27.00	39.00	26.97



of cities: one in which motorised modes dominate, one in which public transport (PT) and walking are more important, and one in which people move prevalently by bike.

In terms of pricing structure, the policy implemented by the car sharing operator at the time the dataset was collected was quite simple: the rental price is a linear function of the rental time (the specific price per minutes varies across the ten cities in the range [0.24, 0.46] €cent/min). No surge pricing nor proximity-based pricing were implemented in the ten cities. Also, there were no incentives for customers to change their destination and to bring back cars to areas where cars were more in demand. A per-kilometre fee is applied only when the car is used for more than about 200 km.

Finally, an interesting feature of this dataset is that it contains entries for two cities (Copenhagen and Stockholm in our analysis) for which the car sharing operator has now shut down service. An index that is often used as a measure of car sharing success is the *vehicle utilisation rate*, defined as the number of daily trips per vehicle. A higher value means that vehicles are used intensively in the city, hence the car sharing service is more profitable. Please note that long trips in which customers rent the shared vehicle for a long time are not the target of car sharing services but belong to the class of long-term rental. For this reason, the vehicle utilisation rate, with its ability to capture the short and frequent trips, is a direct measure of car sharing effectiveness. Figure 2 shows the utilisation rate in the ten cities. It is clear how vehicles in some cities are much more utilised than in others, even 2–3 times more. It is also interesting to note that the vehicle utilisation rate is the lowest in the two cities (Copenhagen and Stockholm) where the service has been shut down months after we had collected this dataset. Remarkably, in Turin and Vienna there is quite a lot of variability in the utilization rate. This is due to vehicles being injected or removed from the system during the data collection period.



4 Demand characterisation through sociodemographic indicators and urban diversity metrics

In this section we focus on the demand, i.e., on the number of pickup requests observed in the different areas of a city, and we investigate how they are related to sociodemographic and urban fabric indicators. We discuss these indicators (which are the explanatory variables for our model) below, together with a brief description of the spatial unit of analysis considered in this section.

Sociodemographic data: Sociodemographic indicators characterise the population in the different areas of a city. For this analysis, we need a granularity finer than city level.^b We were able to find open census data with the desired spatial granularity for the cities of Florence, Milan, Rome, and Turin. For their analysis, we focus on indicators related to the marital status, age group, educational attainment, employment status, and commuting habits. The census data are obtained from the Italian National Institute for Statistics (ISTAT) and correspond to the 2011 Italian Census.^c

Urban fabric data: The wealth of activities (cultural, commercial, recreational, etc.) taking place in a specific area is characterised using information about the Points of Interest (PoIs) collected from the location-based social network Foursquare.^d When a user enters a new PoI, they are prompted to enter one of the first-level categories defined by the platform, which are Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service. We do not consider the category Event because events are generally limited in time, hence they typically do not overlap with our period of observation of the car sharing dynamics. Using this information, the urban fabric is characterised computing the number of PoIs (per category and overall) in each area. We also include a measure of the diversity of the urban fabric in an area by exploiting the concept of venues entropy introduced in [47]. The venue entropy of an area a is obtained as:

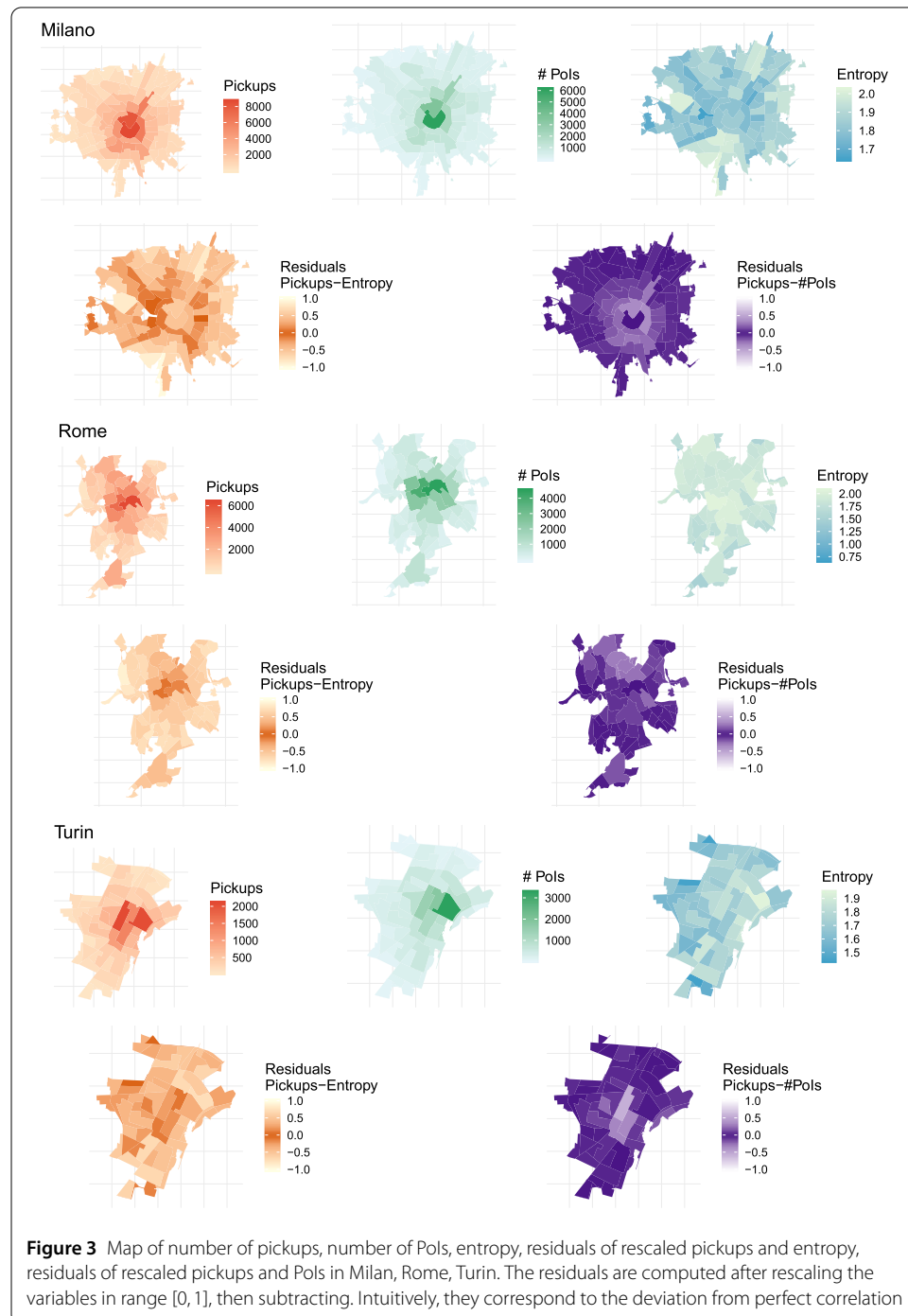
$$e(a) = - \sum_{c \in \mathcal{C}} \frac{n_c(a)}{n(a)} \times \log \left(\frac{n_c(a)}{n(a)} \right), \tag{2}$$

where \mathcal{C} is the set of first-level Foursquare categories, $n_c(a)$ denotes the number of PoIs of category c in the area, and $n(a)$ is the total number of PoIs in a . Intuitively, the entropy measures the uncertainty in predicting the category of a venue taken at random from the area, so the harder the prediction, the greater the diversity.

Spatial unit of analysis: We are constrained to use the smallest census area for which data are provided. In case of census areas that only partially cover the car

sharing operation area, we consider the polygon resulting from their intersection and we rescale the sociodemographic indicators according to the percentage of overlapping. In order to have consistent estimates of the indicators inside each unit of analysis, we discard the census areas that overlap for less than 20% with the operation area.

The pickups events, the POIs and the entropy in the spatial units of analysis for the four cities are illustrated in Fig. 3.



4.1 Explanatory analysis

Methods: We investigate the relation between the total number of pickups (y) and the indicators discussed above (which we denote with x_k) using a multivariate linear regression model of the form:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \epsilon, \quad (3)$$

where $\beta_0 \dots \beta_j$ are the unknown parameters and ϵ is the error term. As expected for the kind of indicators that we are considering, multicollinearity is present in the data. In order to mitigate its negative effects, we use Lasso shrinkage [48] to estimate the coefficient of our linear regression.^e Another advantage of Lasso is that it also perform subset selection, whereby a reduced set of predictors that have the greatest effect on the response y is selected. In short, Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being smaller than a constant.

In standard linear regression, significance tests are used to test the statistical reliability of the rejection of the null hypothesis (i.e., that a coefficient β_k is zero). Recently, a significance test for the Lasso regression has been proposed [49], that factors in the selection bias related to the subset of selected predictors.^f We will use this significance test to provide the p-values for the coefficients of our regression.

As for the predictors that are skewed, we handle them applying a log transformation.

Results: The cities for which we were able to obtain census data with the required granularity are Milan, Florence, Rome, Turin. For Florence, the census areas with a significant overlap with the operation area were too few to get statistically meaningful results, so we also discarded this city. The Lasso regression results for the remaining Italian cities are shown in Table 4.

In all Italian cities (Table 4), a high educational attainment in a certain area is significantly associated with an increased demand for car sharing in that area. Vice versa, a low educational attainment is associated with lower demand in both Rome and Turin. This is largely in agreement with the findings from survey data discussed in Sect. 2.1: car sharing users tend to be better educated than the general population, and this signal is strong enough to be detected by the correlation between the demand and the demographic composition of neighborhoods.

Regularly commuting outside the reference municipality correlates negatively with car sharing demand (Milan, Rome). This is due to the fact that car sharing vehicles cannot be parked outside the operation area, hence they are not suitable for this type of commuting. They would be suitable, though, if paired with local public transport, using car sharing as a first/last-mile solution. This does not seem the case for Milan and Rome (the presence of transport facilities is not affecting the demand). It might be the case for Turin, as commuting outside the municipality is not considered a good predictor of the demand while a certain effect of the presence of transport facilities is detected. However, this effect is not statistically significant, hence this conclusion cannot be drawn from the data at hand. Our analysis seem to confirm the complex relationship between free floating car sharing and public transport discussed in Sect. 2.1: the synergy or friction between the two could be heavily dependent on local characteristics. An ad hoc analysis of this relationship would be an interesting follow-up work of the current investigation.

The marital status and age never correlate with the car sharing demand in the cities under study. The latter result is in contrast with the findings based on survey data, where

Table 4 Lasso regression for Italian cities. Each cell contains the coefficient estimated by Lasso regression. Statistical significance is reported as (***) = $p < 0.001$, (**) = $p < 0.01$, (*) = $p < 0.05$, (.) = $p < 0.1$

Predictors	Coefficients		
	Milan	Rome	Turin
Total population	-	-	-
# unmarried	-	-	-
# married	-	-	-
# separated	-	-	-
# widows	-	-	-
# divorced	-	-	-
Age < 5	-	-	-
Age 5–9	-	-	-
Age 10–14	-	-	-
Age 15–19	-	-	-
Age 20–24	-	-	-
Age 25–29	-	-	-
Age 30–34	-	-	-
Age 35–39	-	-	-
Age 40–44	-	-	-
Age 45–49	-	-	-
Age 50–54	-	-	-
Age 55–59	-	-	-
Age 60–64	-	-	-
Age 65–69	-	-	-
Age 70–74	-	-	-
Age > 74	-	-	-
# with university degree	0.22315 (***)	0.13065 (***)	0.22787 (***)
# with high school degree	-	-	-
# with middle school diploma	-	-	-
# with primary school diploma	-	-0.10589 (.)	-0.08575 (***)
# literate	-	-	-
# illiterate	-	-	-
# employed	-	-	-
# unemployed	-	-	-
# stay-at-home	-	-	-
# students	-	-	-
# other situations outside workforce	-	-	-
# commuting inside the municipality	-	-	-
# commuting outside the municipality	-0.1054700 (***)	-0.02632 (*)	-
# getting money	-	-	-
# Pals	-	-	-
Pals entropy	-	0.19928 (***)	-
# Arts & Entertainment	-	-	-
# College & University	-	-	0.04625
# Food	-	0.24642 (*)	-
# Nightlife Spot	0.2224 (**)	0.13826 (.)	0.28672 (***)
# Outdoors & Recreation	0.17047 (***)	0.08842	0.04746
# Professional & Other Places	0.17197 (*)	0.26091	0.26261
# Residence	0.12921 (**)	-	-
# Shop & Service	-	-	-
# Travel & Transport	-	-	0.02488

age always played a significant role in the profiling of car sharing users. One explanation could be that the age-related signal is weaker than the education-related one. Then, due to collinearity effects, the explanatory power of age is not considered sufficient by the Lasso. Another explanation is that age alone has never been explanatory, and its presence has always been due to its correlation with higher education attainments (in most OECD countries, young people are more educated than the elderly⁸).

In terms of urban fabric indicators, the presence of nightlife activities is associated with increased demand in all three Italian cities, while the presence of outdoor and recreational activities, as well as professional PoIs and residences, have a statistically significant effect in Milan only. Thus, leisure seems to be the main motivation behind car sharing trips in the three cities. Work-related trips are significant only in Milan. When comparing these results with the survey-based findings summarised in Sect. 2.1, no clear trend emerges. While leisure and work trips are a common finding, the signal associated with shopping activities goes completely undetected in these three cities.

5 Demand forecasting

In this section, we focus on the elements that influence the short-term behaviour of a car sharing system and we exploit them to forecast the demand. As we are interested in a finer spatial granularity (e.g. block level), we depart from the census areas used in the previous section. We thus need to identify a meaningful spatial unit to define car availability in a given area. In fact, differently from station-based car sharing, in free floating car sharing there is no natural “aggregation” point for vehicles, which can be freely picked up and dropped off anywhere within the operation area. We can still perform a spatial analysis of car sharing usage by dividing the operation area into smaller cells and studying what is the behaviour, over time, in each of these cells. In this work we consider cells with side length 500 m, which is the maximal walking distance typically accepted by car sharing users [28, 50].

Demand predictability is one of the crucial aspects for every transportation system. In car sharing, in particular, it is of utmost importance for vehicle redistribution, whose goal is in fact to proactively move vehicles in order to address the future demand. In [29], comparing the time series of empty cells over time against that of available vehicles, we have shown that there typically a lot of empty cells but at the same time there are also a lot of available vehicles. This situation hints at a strong concentration of vehicles in certain areas, vehicles that could be proactively moved to where the customers most need them. Vehicle redistribution is typically performed periodically (e.g., every hour) and can be represented as a continuous cycling between three phases: (i) the *forecast* phase, when the expected pickups and drop-offs during the next relocation window are predicted; (ii) the *selection* phase, when the areas with vehicle surplus and vehicle deficit are identified and matched; and (iii) the *dispatching* phase, when the relocation workforce is assigned the previously defined relocation tasks [28].

Our goal in this section is not to develop a new custom-built method for demand prediction in car sharing systems, but rather to compare state-of-the-art solutions that belong to different forecasting approaches (see Sect. 2.2 for the discussion on existing methods) in order to understand their performance in the ten cities under study. Indeed, while prior work on demand prediction has focused on individual cities, it is important to assess the robustness of the most representative methods to cope with the heterogeneity of travel behaviours and urban fabric. The target audience of this analysis are researchers working on designing optimised transport models for car sharing who might benefit from knowing what is the best, off-the-shelf, approach to prediction, so that they can focus their efforts on optimising the selection and dispatching phase discussed above. Similarly, third-parties developers will benefit from this type of analysis. For example, one could think to set up a service (similar in vein to OpenStreetCab [51]), whose goal is to provide the best option

price-wise between Uber and NYC taxis for a given trip) whereby the most *reliable* car sharing service is recommended (e.g., one that guarantees that a car will be available in the evening when one drives back home). Third-parties apps will most likely have access only to the public data made available by the car sharing operators (similar to the data we are dealing with).

Problem definition: The goal of demand prediction is to establish the vehicle deficit/surplus at the cells. It can be described using the general formula we presented in [41], which we discuss hereafter in a simplified version. If we denote with T the interval at which relocation is performed, every T minutes the car sharing operator will compute, for each cell i , the expected balance \hat{b}_i of vehicles at cell i for the next T minutes, which can be described as follows:

$$\hat{b}_i = v_i + \hat{drop}_i - \hat{pick}_i, \quad (4)$$

where v_i is the number of cars currently parked at station i , while \hat{drop}_i and \hat{pick}_i are, respectively, the forecast number of drop-offs and pickups in the next time interval. Please note that v_i is a known quantity as it photographs the current situation at cell i . Instead, \hat{drop}_i and \hat{pick}_i have to be estimated from what has happened in the past.^h In the following, we show how statistical learning can help fill this gap and thus close the relocation cycle.

Let us focus on a tagged cell i belonging to the set of all cells \mathcal{C} . We denote the set of days in our observation period with \mathcal{D} . Then, we divide each day $d \in \mathcal{D}$ in bins of length T (i.e. we discretize time). The prediction problem at hand is a typical one: we have historical data (a set of N observations) about pickup and drop-offs at cell i in each bin t for each day in \mathcal{D} . We have to predict what will happen in each bin of the next days. In the following, we use the general term *event* to denote either pickup or drop-off events.

Features: For each cell i , we extract the following features for prediction:

- number of events $e_{(i,d,t)}$ observed in cell i at time t of day d
- the time of the day (corresponding to bin t)
- the day of the week (Sunday, Monday, etc.)
- whether the day is a weekday or not
- average number of events $\hat{e}_{(i,d,t)}$ observed at bin t of day d in the neighbouring cells (we consider 2-hop neighbours only).

Methods: We use the first 80% of the days in the dataset for training, and we predict the remaining 20%.ⁱ We set the time window T to 1 hour, implying that we want to forecast pickups and drop-offs happening in a one-hour time frame. We only consider cells that have more than 30 events during the observation period. Then, we run the prediction algorithms and we measure the prediction error in terms of Root Mean Squared Error (RMSE).

We now define a set of relevant prediction techniques to be evaluated on the datasets at hand. It is important to point out that car sharing operators do not disclose any detail on their approach to demand prediction. Thus, comparing against state-of-the-art industrial benchmarks is not an option. The first two solutions that we consider are simple baselines based on historical averages/medians. With regards to our discussion in Sect. 2.2, the third one is representative of the class of time series prediction. Then, we pick two approaches for the machine learning category: neural networks (which have been already used in the literature for car sharing [24–26]), and Random Forest (which has been shown

to be extremely effective when applied to bike sharing booking predictions [31, 32]). Finally, we test a technique in the custom forecasting category, specifically the one proposed in [28]. In the following we provide a description of each technique.

Prediction based on Historical Average (HA): this prediction function returns the average number of events observed in the same time window across different days. In other words, the predicted number of events \hat{y}_t at a certain time t in the future is obtained as $\hat{y}_t = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} e_{(i,d,t)}$. As car sharing typically exhibits marked differences between weekdays and weekends [20], we also test a version of the algorithm (denoted as HA+) that distinguishes between working days and weekends. A similar function has also been used as benchmark in the related literature on bikesharing forecasting [32, 33].

Prediction based on Historical Median (HM): the prediction function returns the median number of events observed in the same time window across different days, i.e., $\hat{y}_t = \text{median}_{d \in \mathcal{T}}(e_{(i,d,t)})$. This function is expected to perform well in cases where the distribution of pickups/drop-offs is highly skewed. As for the previous algorithm, we also test a version (denoted as HM+) that distinguishes between working days and weekends.

ARIMA: the Autoregressive Integrated Moving Average technique is a popular time series forecasting method. It is a generalisation of the ARMA model used in [22, 33]. Typically, ARIMA models are denoted with $\text{ARIMA}(p, d, q)$, where p is the order (number of time lags) of the AR component, d is the degree of differencing, and q is the order of the MA component. Here we use the seasonal version of the above ARIMA model, estimating the parameters for both the non-seasonal and the seasonal component (this allows us to detect cyclic behaviour, if it exists). We remind that in a seasonal ARIMA model, seasonal AR and MA terms predict the target variable using data values and errors at times with lags that are multiples of S (the span of the seasonality). For each cell the best configuration of the ARIMA parameters is selected according to their Corrected Akaike Information Criterion (AICc) value, using the `auto.arima` function of R's `forecast` package. The search range for the parameters is the default one in the `auto.arima` function. Being this a time series method, only the temporal information of each observation and the actual observed values are fed to the model.

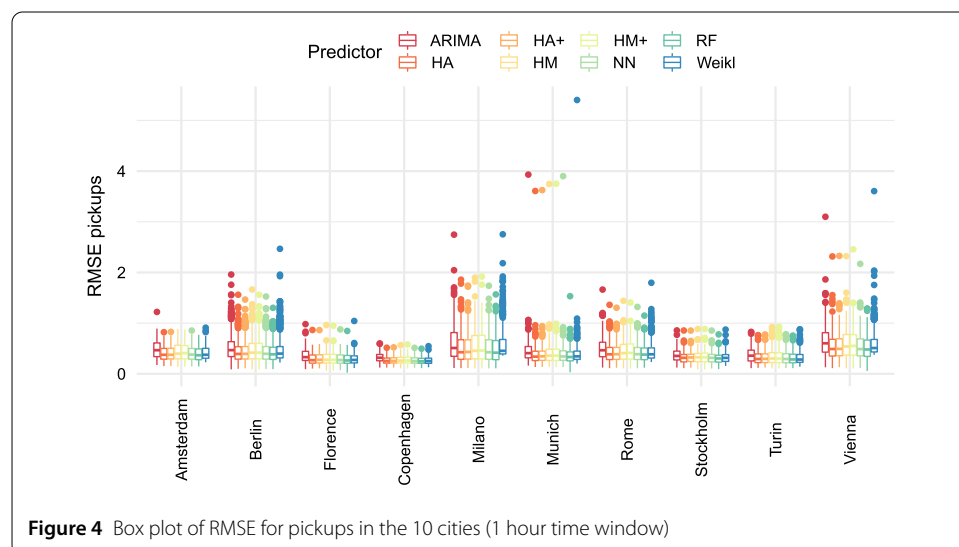
Random Forest (RF): tree-based learning method that aggregates the prediction results of several decision trees obtained by randomly selecting, each time, only a subset m of the original p features (those described in the features section above). In order to select the most appropriate m , we used 5-fold cross validation and we vary m in $\{2, 4, 5\}$. We use the implementation in the R package `randomForest`, together with the `caret` package for training and prediction.

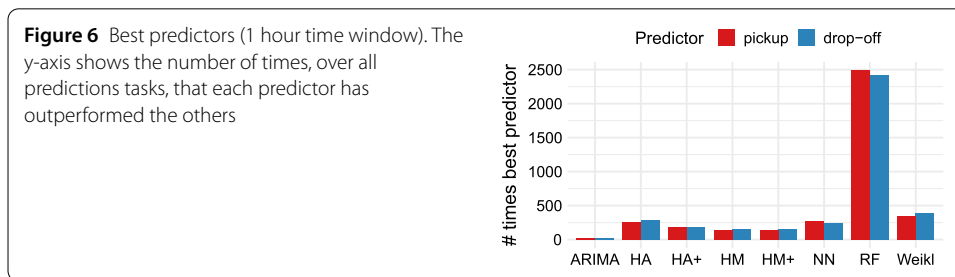
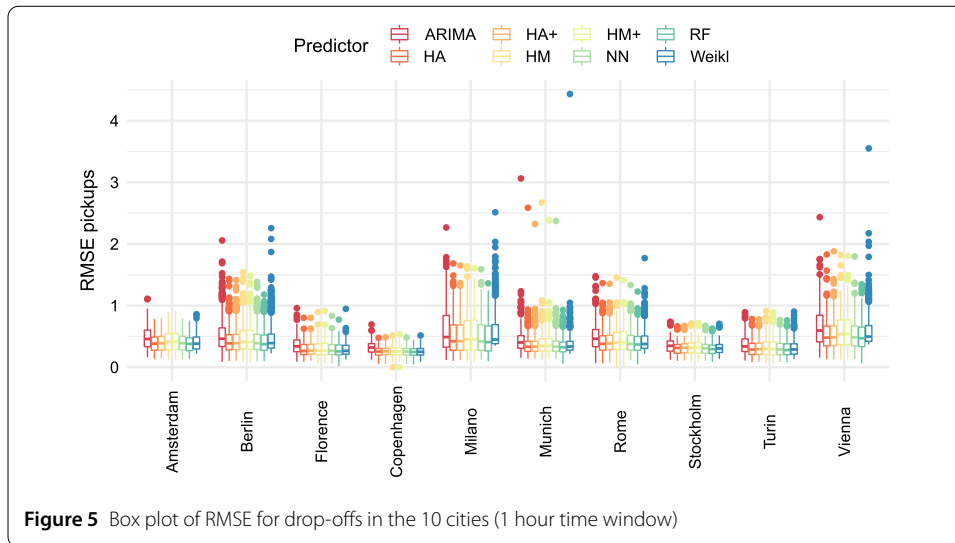
Neural Network (NN): relying on the same settings as in [23], we use a single layer perceptron with as many neurons in the input layer as the features described above, one hidden layer (searching for the best number of neurons between 1 and 30), single output neuron, backpropagation, hyperbolic tangent activation function, linear output function. Categorical features have been represented using dummy variables. Then, input and output data were scaled to the range $[-1, 1]$, which is the sensitive range of the hyperbolic tangent activation function. We rely on the implementation in R package `RSNNS`, together with the `caret` package for training and prediction. Parameters selection is again performed using 5-fold cross validation.

Algorithm in Weikl and Bogenberger [28] (WEIKL): one of the very few custom proposals in the literature on car sharing, the rationale of this algorithm is to represent each timeslot

of each day through a vector, whose components are the number of events at each cell during the timeslot. Let us focus on a tagged timeslot t . These vectors describing the spatial demand for timeslot t across each day make up a matrix of size $|\mathcal{C}| \times |\mathcal{D}_{\mathcal{T}}|$ (where \mathcal{C} denotes the set of cells and $\mathcal{D}_{\mathcal{T}}$ denotes the set of days in the training set). The $|\mathcal{C}|$ -dimensional representation of each day is then simplified using Principal Component Analysis, and only the first two principal components are retained. This two-dimensional description of the days is then clustered using k -means, in order to group together days featuring the same demand behaviour. In the original paper, how the optimal number of groups is obtained is not specified, so we decided to rely on the gap statistic [52], a state-of-the-art solution that is able to handle also the single-group case (i.e., to detect when the optimal choice is to not split in groups). Once this has been done for all timeslots, a so-called from-to matrix is built, computing the probability that days in a certain group g_i in timeslot t would be in group g_j in timeslot $t + 1$. Using this from-to matrix, it is possible to compute the demand variation from a timeslot to another for each group. This concludes the training phase of the algorithm. In the prediction phase, the demand in timeslot $t - 1$ is mapped into one of the groups computed in the training phase (by closest centroid matching). Then, the number of forecasted events for timeslot t is obtained from the computed expected demand variation for the group. Please note that in [28], each day was divided in timeslots of non-uniform size. For fairness with the other prediction algorithms, we use timeslots of fixed size T . We have implemented this method in R.

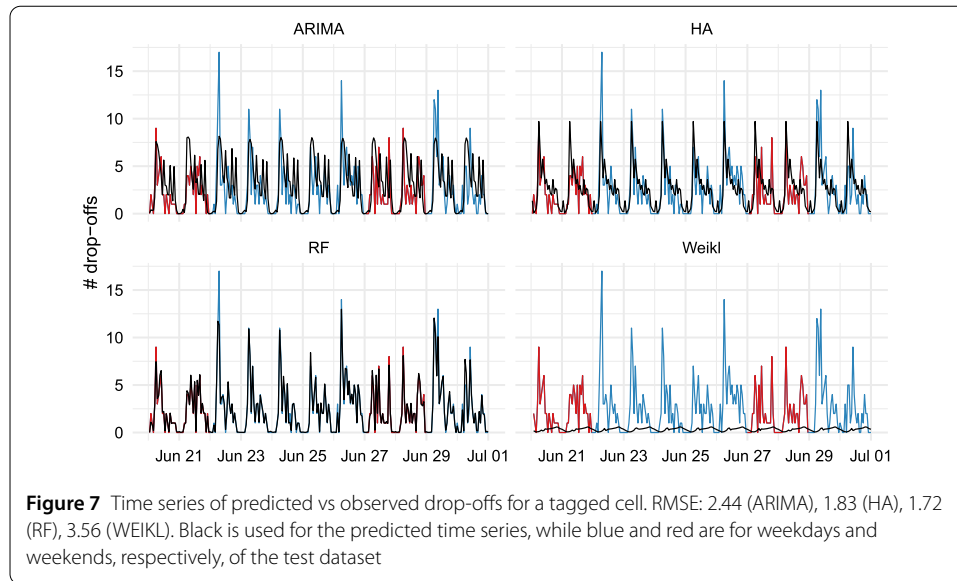
Results: The results are shown in Figs. 4 and 5, for pickups and drop-offs respectively. For most cities and for all algorithms, the error is small, with forecasts off, on average, by less than one drop-off/pickup for the vast majority of cells. However, there are a few cells for which the prediction error is high. After an in-depth analysis of the nature of these cells, we discovered that they are typically in very busy areas (e.g. near the airport), where both the high volume of traffic and the bustier nature of arrivals and departures may explain this variability. Also, the RMSE for pickups tends to be slightly higher than for drop-offs. In terms of which prediction algorithm works best, Fig. 6 shows that Random Forest provides the most accurate predictions for the vast majority of cells. The WEIKL algorithm is the second best, but its performance is very close to that of the NN approach and, surprisingly,





to the simple Historical Average. HA+ and HM+, the versions of HA and HM algorithms that take into account the difference between weekdays and weekends, do not outperform in general their simpler counterparts. ARIMA, used also in [22] for forecasting car sharing demand, provides consistently the worst predictions.

In Fig. 7, in order to showcase the main strengths and weaknesses of the prediction techniques used, we focus on a tagged cell (specifically, on one for which the error is generally large) and we plot the time series of the predicted drop-offs (black curve) against the observed drop-offs (blue and red, in order to distinguish between weekdays and weekends). For the sake of readability, we consider one strategy per class of prediction approach: HA for the simple baselines, ARIMA for the time series forecasting class, RF for the machine learning approaches, and WEIKL for the custom solutions. The ARIMA model tends to replicate the same daily patterns across all days in the test set, since the ARIMA model is not able to capture multiple seasonalities, which are instead present in the data. By using predictive models that explicitly handle these multiple seasons (such as [53]), the quality of prediction could be significantly improved. A similar problem seems to hold for HA: it tends to replicate a “model day”, which is always the same. Instead, the predictions provided by the Random Forest algorithm are the most flexible ones, as they seem to adapt individually to each day. However, despite this flexibility, there seems to be an inherent variability in certain cells in the datasets (Figs. 4–5) that makes prediction difficult. The tagged cell considered here is also useful to illustrate the weakness of the WEIKL solution. Since it groups together many cells to extract a typical behaviour of the system in a given timeslot, the cells with a small number of events (which are many) tend to dominate

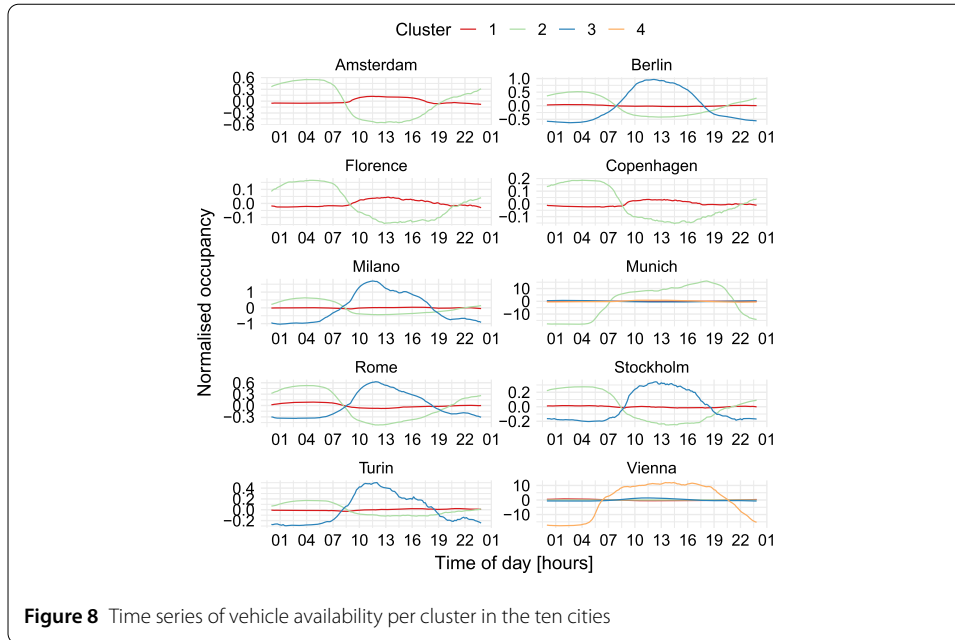


over the more active ones (like the tagged cell considered here). Thus, in these cases, the predictions are significantly off with respect to the actual behaviour of the cell.

6 Spatiotemporal usage patterns

It is expected that cells in a car sharing system are used differently by the users, but how many different usages can be identified? In order to answer this question, in the following we carry out a classification of cells based on their usage pattern. To this aim, we focus on the time series of vehicle availability in each cell and we measure how close this time series is with what we observe in other cells. We measure the time series distance using the Dynamic Time Warping (DTW) technique [54] (with Sakoe-Chiba band), then we cluster cells based on their DTW-distance using Partition Around Medoids (PAM) clustering. For each city, the optimal number of clusters is obtained using the silhouette method. In order to be able to compare our time series, we discretise time into bins with a duration of 10 minutes. For each cell, we extract one availability value per bin by averaging the availability in the bin in different days. In addition, in order to detect variation above and below the average behaviour, we normalise the measured availability using the average availability at the cell.

The results are shown in Fig. 8. The optimal number of clusters is 2 in Amsterdam, Florence, and Copenhagen, 3 in Berlin, Milan, Rome, Stockholm, Turin, and 4 in Munich and Vienna. However, the fourth cluster, when present, is a very special cluster, composed of just a single cell. This single cell is a very special one in the city ecosystem, and in both cities where the fourth cluster is present, this cluster comprises the airport zone. If we plot the availability time series within each cluster (Fig. 8, obtained by computing the average availability in the cells belonging to the cluster), it is striking to see that the clusters highlight very characteristic cell usage. Some cells have above average availability at night and below average availability during the day. Other cells have exactly the opposite behaviour. Finally, there is a group of cells with an intermediate behaviour, where apparently no significant difference in usage is detected over the whole day. It is easy to map this behaviour into the “nature” of the area covered by the cell: people leave residential areas in the morning



and come back in the evening, while the opposite is true for commercial/business areas. Similar classes were identified in [20] for station-based car sharing, and in [31] for bike sharing. Figure 8 also highlights the outlier behaviour of the airport zone (which constitutes the fourth cluster, when available). Airports in Munich and Vienna see a huge variation in availability; however, the behaviour of their time series is simply a scaled version of the commercial/business pattern discussed before. Due to the magnitude of the airport clusters' time series, the behaviour of the other clusters of Munich and Vienna is barely visible in the plot. If we zoomed in, we would see the typical patterns that can be seen more clearly in cities with no airport within the operation area.

Based on the above discussion, we can associate each cluster with the trend in its corresponding availability time series. Thus, we identify four main behaviours: cells with mid-of-day availability peak, cells with night peak, cells with no significant peak, and cells whose availability variations are much higher than in other cells. We use the labels *day*, *night*, *neutral*, and *high-intensity* to refer to these four classes. In the following, we investigate to which extent the behaviour of cells is spatially autocorrelated. To this aim, since cell labels are categorical, we use the Join Count statistics [55]. With this approach, for each cell n , we count how many of its neighbouring cells belong to n 's class and we compare this result with what would be obtained if classes were distributed uniformly at random across cells. Since the high-intensity class comprises at most one cell per city, we discard it from the analysis. The results for all cities are shown in Table 5. Cells exhibiting an availability peak at night are spatially autocorrelated in all ten cities. Cells with a mid-of-day peak are spatially correlated in all cities except for Florence and Copenhagen. Out of the seven cities featuring neutral cells, the spatial autocorrelation is significant for only three of them. We can conclude that, in general, the availability of vehicles in cells tends to be spatially autocorrelated, hence neighboring cells tend to have shortage/abundance of vehicles at the same time. This further motivates the use of vehicle availability information in neighbouring cells for demand forecasting (RF and NN in Sect. 5 indeed rely on

Table 5 Join Count Statistics. Column *Count* contains the total number of matches; column *Exp. (rand)* contains the number of matches expected under random; *Test Stat.* contains the test statistics; ^(†) denotes p-values greater than 0.05 (not statistically significant). Neighbour cells are obtained using the Queen criterion (i.e., assuming moves like the Queen in a chess game); spatial weights are binary

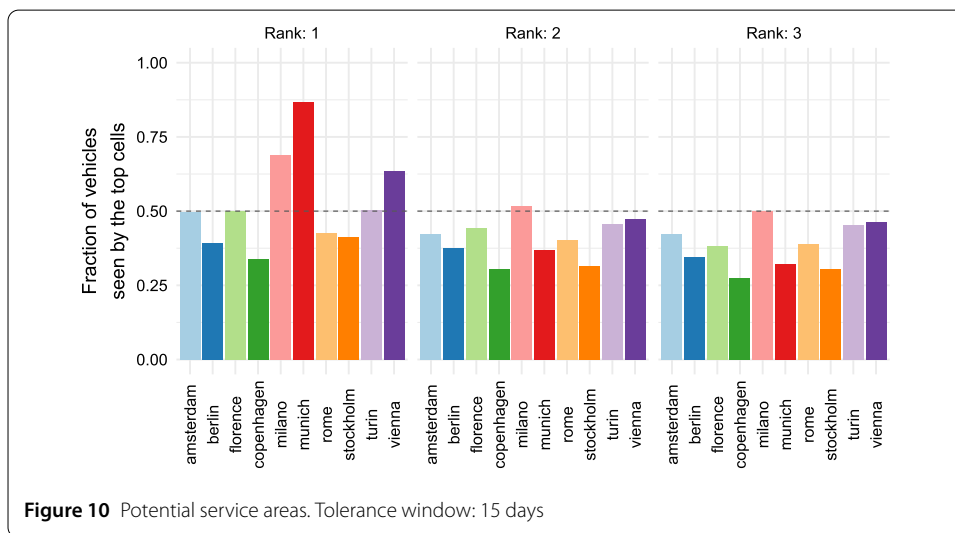
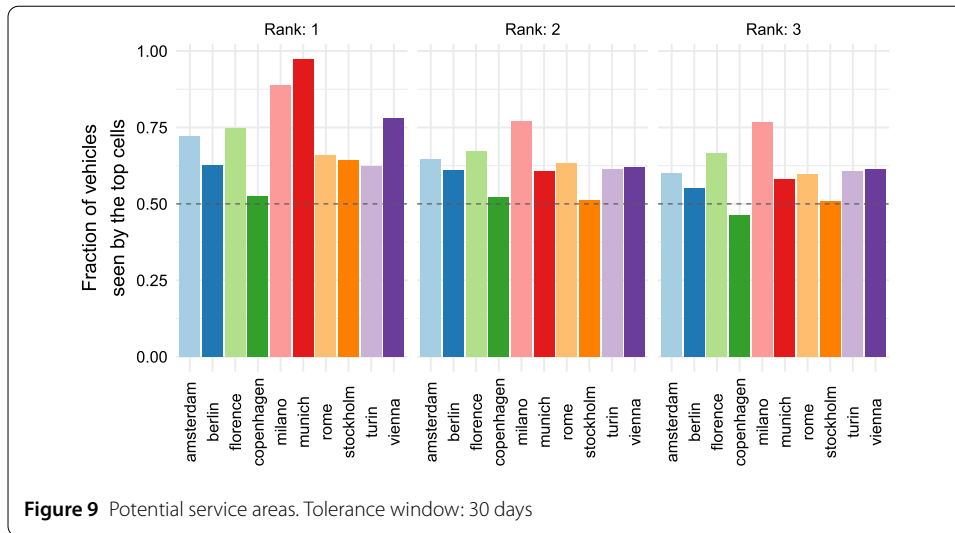
City	Cluster Type	Count	Exp. (rand)	Test Stat.	p-value
Amsterdam	day	753	726.56	2.22	1.33e-02
	night	132	53.13	12.40	1.25e-35
Berlin	neutral	1623	1571.39	2.44	7.42e-03
	night	501	295.93	14.81	6.38e-50
	day	254	65.88	25.53	4.31e-144
Florence	day	400	403.76	-0.33	6.29e-01 ^(†)
	night	109	86.40	2.95	1.60e-03
Copenhagen	day	385	385.50	-0.05	5.22e-01 ^(†)
	night	86	59.11	4.37	6.18e-06
Milan	neutral	837	793.31	2.82	2.40e-03
	night	349	174.15	16.54	9.25e-62
	day	89	10.18	26.01	1.92e-149
Munich	neutral	613	588.51	1.64	5.10e-02 ^(†)
	night	221	98.37	14.37	3.95e-47
	day	82	26.10	11.79	2.13e-32
Rome	neutral	623	609.62	0.90	1.84e-01 ^(†)
	night	163	88.58	9.24	1.26e-20
	day	265	75.74	25.09	3.35e-139
Stockholm	neutral	192	198.33	-0.69	7.56e-01 ^(†)
	night	66	50.95	2.53	5.72e-03
	day	22	9.13	4.60	2.10e-06
Turin	neutral	422	408.32	1.26	1.04e-01 ^(†)
	night	152	100.44	6.61	1.96e-11
	day	41	4.01	19.43	2.20e-84
Vienna	night	372	231.90	11.83	1.29e-32
	neutral	657	599.45	3.75	8.80e-05
	day	85	18.59	16.45	3.90e-61

this information and their performance is quite good, with RF being the most performing prediction algorithm overall).

7 Locating cleaning and maintenance areas

A critical operational aspect for car sharing is how to perform cleaning and maintenance. When not done properly, it may even be a critical factor of the service shutdown, as in the case of Parisian car sharing Autolib.^k In order to perform cleaning and maintenance, the car sharing workforce is typically remotely dispatched to collect vehicles that are in need of either. However, moving workers around is expensive, and more efficient solutions could be found based on the vehicle usage in the city. As a case study, in the following we discuss how to identify potential service areas within the operation area. A potential service area is a location vehicle pass by with very high probability. A workshop could be deployed in this area, and this would make cleaning and maintenance operations much more efficient.

We can use our dataset to understand if these potential service areas exist or not in the cities covered by the car sharing service under study. To this aim, we define a reference window W , corresponding to the accepted tolerance for taking out a vehicle for maintenance. Based on data from active car sharing operators, we assume that reasonable values for W are between 15 and 30 days. Then, for each cell, we count the number of distinct vehicles seen by the cells during W . Figures 9 and 10 show the results for the top three



cells in each cities, i.e., the three cells that see the highest number of distinct vehicles during two different time windows ($W = 30$ and $W = 15$ days, respectively). Assuming that a (somewhat generous) threshold of 50% vehicles would be acceptable for the car sharing operator to justify the opening of a workshop in the area, all cities with the exception of Florence would accommodate three workshops satisfying this requirement when $W = 30$. The scenario $W = 15$ is by far more challenging: six cities would be able to open at least one workshop, but only one city could open two and three. The top ranking cell for cities whose operation area covers the airport is always the cell that includes the airport, which thus becomes a strategic asset in car sharing operations, in addition to being a huge generator of car sharing traffic.

8 Conclusions

In this work, we have collected web-based data about free floating car sharing in 10 European cities, cities that are heterogeneous both in terms of car sharing success and mode split. We have studied how the car sharing demand relates to sociodemographic and urban

indicators, showing that the car sharing demand is positively correlated with high educational attainment and nightlife activities, while being negative correlated with the percentage of people commuting outside the municipality. These findings both confirm and extend the results in the related literature obtained from survey data. Then, focusing on the predictability of future car sharing requests, we have shown that they can be forecasted quite accurately using state-of-the-art prediction algorithms, and we have highlighted the very good performance of Random Forest as predictor. Finally, we have proposed a strategy for selecting the area in which maintenance facilities should be deployed, and we have shown how the airport zone can become a strategic asset for car sharing operators, due to the fact that the high volume of traffic generated by the area makes it extremely convenient to deploy cleaning and maintenance facilities there.

Funding

This work was funded by the ESPRIT, REPLICATE and SoBigData projects. The ESPRIT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653395. The REPLICATE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 691735. The SoBigData project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654024.

Abbreviations

PT, Public Transport; RMSE, Root Mean Squared Error; PAM, Partition Around Medoids; AICc, Corrected Akaike Information Criterion.

Availability of data and materials

The data that support the findings of this study cannot be publicly shared. Information is available from the corresponding author upon reasonable request. The ISTAT data are publicly available and can be found at <https://www.istat.it/it/archivio/104317>.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Designed the study: CB RB. Analyzed the data: CB HL. Wrote the paper: CB RB. All authors read and approved the final manuscript.

Authors' information

CB and RB are permanent researchers at IIT-CNR. HL was a postdoctoral researcher at IIT-CNR at the time the study was carried out.

Endnotes

- ^a Only for these cities we were able to find fine-grained geospatial census data significantly overlapping with the car sharing operation area.
- ^b Please note that EU countries are legally bound to provide census data to the Eurostat database at most at the level of NUTS 2 (regions). The actual database (<https://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/database>) contains data up to NUTS 3 level (provinces) but this is not enough for our purposes. For this reason, we resorted to individually checking the countries' official institutes for statistics.
- ^c <https://www.istat.it/it/archivio/104317>
- ^d Through the Foursquare Places API it is possible to browse the venues in a certain geographic area. Since the standard API returns at most 50 venues per input area, each city is split into several browsing areas, whose size is properly dimensioned to ensure that all the available venues are acquired.
- ^e We use Lasso regression as implemented in the R package `glmnet` [56], using 10-fold cross validation for parameter estimation.
- ^f As an example, running an Ordinary Least Square linear regression on the selected subset of predictor and calculating the p-values associated to the coefficients would yield a very optimistic estimate of the significance, due to the fact that the subset of predictors is not selected independently of the data.
- ^g <https://data.oecd.org/eduatt/population-with-tertiary-education.htm#indicator-chart>.
- ^h Note also that, for the sake of clarity, in Equation 4 we are intentionally neglecting the contribution of relocated vehicles that have yet to arrive at the cell from the previous relocation interval. This does not affect the forecast results discussed in this section because this number would be known in advance anyway and, thus, would not be part of the prediction process.
- ⁱ Please note that standard k-fold cross validation cannot be performed with time series because time series data are not independent across time. The approach used in this paper is the same used in [32].

- ^j Note that the set of initial features ($p = 3$) is expanded after applying OneHot encoding. For example, the categorical day of the week is split into 6 binary features.
- ^k <https://www.thelocal.fr/20180619/wheels-set-to-come-off-paris-autolib-electric-cars>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 April 2018 Accepted: 19 February 2019 Published online: 01 March 2019

References

1. Mitchell WJ, Borroni-Bird CE, Burns LD (2010) Reinventing the automobile. Personal urban mobility for the 21st century, vol g. MIT Press, Cambridge.
2. "Cars are parked 95% of the time." Let's check! <http://www.reinventingparking.org/2013/02/cars-are-parked-95-of-time-lets-check.html>. Accessed 5 April 2018
3. Fact #613: March 8, 2010 Vehicle Occupancy Rates (2010). <https://energy.gov/eere/vehicles/fact-613-march-8-2010-vehicle-occupancy-rates>. Accessed 5 April 2018
4. European Environment Agency: occupancy rates of passenger vehicles. <https://www.eea.europa.eu/downloads/90455cbdbff89c2c6149387ee11e4ea/1441389594/occupancy-rates-of-passenger-vehicles-1.pdf>. Accessed 5 April 2018
5. Shaheen S, Cohen A (2015) Mobility and the sharing economy: impacts synopsis—Spring 2015. Technical report, Transportation Sustainability Research Center, University of California, Berkeley
6. Kortum K (2014) Driving smart: carsharing mode splits and trip frequencies. In: Transportation research board 93rd annual meeting
7. Schwieger B, Victorero-Solares P, Brook D (2015) Global carsharing operators. Report 2015. Technical report, Team Red
8. Stopher PR, Greaves SP (2007) Household travel surveys: where are we going? *Transp Res, Part A, Policy Pract* 41(5):367–381
9. Meier RL (1962) A communications theory of urban growth. Published for the Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University by MIT Press, Cambridge
10. Kopp J, Gericke R, Axhausen KW (2015) Do sharing people behave differently? An empirical evaluation of the distinctive mobility patterns of free-floating car-sharing members. *Transportation* 42(3):449–469
11. Wittwer R, Hubrich S (2018) Free-floating carsharing experiences in German metropolitan areas. *Transp Res Proc* 33:323–330
12. Giesel F, Nobis C (2016) The impact of carsharing on car ownership in German cities. *Transp Res Proc* 19:215–224
13. Becker H, Ciari F, Axhausen KW (2017) Comparing car-sharing schemes in Switzerland: user groups and usage patterns. *Transp Res, Part A, Policy Pract* 97:17–29
14. Urban Audit Database. <http://ec.europa.eu/eurostat/web/cities/data/database>. Accessed 5 April 2018
15. Adam M-B (2005) Car-sharing: where and how it succeeds. Transportation Research Board of the National Academies
16. Project M (2010) The state of European car-sharing. Technical report
17. Schmöller S, Weikl S, Müller J, Bogenberger K (2015) Empirical analysis of free-floating carsharing usage: the Munich and Berlin case. *Transp Res, Part C, Emerg Technol* 56:34–51
18. Willing C, Klemmer K, Brandt T, Neumann D (2017) Moving in time and space—location intelligence for carsharing decision support. *Decis Support Syst* 99:75–85
19. Kortum K, Schönduwe R, Stolte B, Bock B (2016) Free-floating carsharing: city-specific growth rates and success factors. *Transp Res Proc* 19:328–340
20. Boldrini C, Bruno R, Conti M (2016) Characterising demand and usage patterns in a large station-based car sharing system. In: The 2nd IEEE INFOCOM workshop on smart cities and urban computing. IEEE, pp 1–6
21. Wang H, Cheu R, Lee D-H (2010) Dynamic relocating vehicle resources using a microscopic traffic simulation model for carsharing services. In: 2010 third international joint conference on computational science and optimization. pp 108–111. IEEE, <http://ieeexplore.ieee.org/document/5532914/>
22. Müller J, Bogenberger K (2015) Time series analysis of booking data of a free-floating carsharing system in Berlin. *Transp Res Proc* 10:345–354
23. Cheu RL, Xu J, Kek AGH, Lim WP, Chen WL (2006) Forecasting shared-use vehicle trips with neural networks and support vector machines. *Transp Res Rec* 1968(1):40–46
24. Xu J-X, Lim JS (2007) A new evolutionary neural network for forecasting net flow of a car sharing system. In: 2007 IEEE congress on evolutionary computation. pp 1670–1676. IEEE, <http://ieeexplore.ieee.org/document/4424674/>
25. Schulte F, Voß S (2015) Decision support for environmental-friendly vehicle relocations in free-floating car sharing systems: the case of Car2go. *Procedia CIRP* 30:275–280
26. Alfian G, Rhee J, Ijaz M, Syafrudin M, Fitriyani N (2017) Performance analysis of a forecasting relocation model for one-way carsharing. *Appl Sci* 7(6):598
27. Boyacı B, Zografos KG, Geroliminis N (2015) An optimization framework for the development of efficient one-way car-sharing systems. *Eur J Oper Res* 240(3):718–733
28. Weikl S, Bogenberger K (2013) Relocation strategies and algorithms for free-floating car sharing systems. *IEEE Intell Transp Syst Mag* 5(4):100–111
29. Boldrini C, Bruno R, Laarabi HM (2017) Car sharing through the data analysis lens—KNOWMe: 1st International Workshop on Knowledge Discovery from Mobility and Transportation Systems. Technical report. [arXiv:1708.00497](https://arxiv.org/abs/1708.00497)
30. O'Brien O, Cheshire J, Batty M (2014) Mining bicycle sharing data for generating insights into sustainable transport systems. *J Transp Geogr* 34:262–273
31. Sarkar A, Lathia N, Mascolo C (2015) Comparing cities' cycling patterns using online shared bicycle maps. *Transportation* 42:541–559
32. Yang Z, Hu J, Shu Y, Cheng P, Chen J, Moscibroda T (2016) Mobility modeling and prediction in bike-sharing systems. In: Proceedings of the 14th annual international conference on mobile systems, applications, and services. ACM, New York, pp 165–178

33. Gast N, Massonnet G, Reijsbergen D, Tribastone M (2015) Probabilistic forecasts of bike-sharing systems for journey planning. In: Proceedings of the 24th ACM international on conference on information and knowledge management. ACM, New York, pp 703–712
34. Trasarti R, Pinelli F, Nanni M, Giannotti F (2011) Mining mobility user profiles for car pooling. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 1190–1198
35. Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, Ratti C (2014) Quantifying the benefits of vehicle pooling with shareability networks. *Proc Natl Acad Sci USA* 111(37):13290–13294
36. de Almeida Correia GH, Antunes AP (2012) Optimization approach to depot location and trip selection in one-way carsharing systems. *Transp Res, Part E, Logist Transp Rev* 48(1):233–247
37. Biondi E, Boldrini C, Bruno R (2016) Optimal deployment of stations for a car sharing system with stochastic demands: a queueing theoretical perspective. In: The 19th IEEE intelligent transportation systems conference, 2016. IEEE, pp 1–7
38. Pavone M, Smith S, Frazzoli E, Rus D (2012) Load balancing for mobility-on-demand systems. *Int J Robot Res* 31(7):839–854
39. Kek AGH, Cheu RL, Meng Q, Fung CH (2009) A decision support system for vehicle relocation operations in carsharing systems. *Transp Res, Part E, Logist Transp Rev* 45(1):149–158
40. Febbraro AD, Sacco N, Saeednia M (2012) One-way carsharing: solving the relocation problem. *Transportation research record. Transp Res Rec* 2319:113–120
41. Boldrini C, Bruno R (2017) Stackable vs autonomous cars for shared mobility systems: a preliminary performance evaluation. In: IEEE MoD@ITSC'17: modelling, analysis and control of intelligent mobility-on-demand systems workshop
42. Rottondi C, Verticale G, Neglia G (2014) On the complexity of optimal electric vehicles recharge scheduling. In: Green communications (OnlineGreencomm), 2014 IEEE online conference on. IEEE, pp 1–7
43. Biondi E, Boldrini C, Bruno R (2016) Optimal charging of electric vehicle fleets for a car sharing system with power sharing. In: IEEE energycon. IEEE, pp 1–6
44. Nourinejad M, Zhu S, Bahrami S, Roorda MJ (2015) Vehicle relocation and staff rebalancing in one-way carsharing systems. *Transp Res, Part E, Logist Transp Rev* 81:98–113
45. Uesugi K, Mukai N, Watanabe T (2007) Optimization of vehicle assignment for car sharing system. In: Knowledge-based intelligent information and engineering systems. Springer, Berlin, pp 1105–1111
46. Jorge D, Correia GHA, Barnhart C (2014) Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems. *IEEE Trans Intell Transp Syst* 15(4):1667–1675
47. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C (2013) Geo-spotting. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '13
48. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58:267–288
49. Tibshirani R, Johnstone I (2014) A significance test for the lasso. *Ann Stat* 42(2):413–468
50. Herrmann S, Schulte F, Voß S (2014) Increasing acceptance of free-floating car sharing systems using smart relocation strategies: a survey based study of car2go Hamburg. In: International conference on computational logistics. Springer, Berlin, pp 151–162
51. Noulas A, Salnikov V, Lambiotte R, Mascolo C (2015) Mining open datasets for transparency in taxi transport in metropolitan environments. *EPJ Data Sci* 4:23
52. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc, Ser B, Stat Methodol* 63:411–423
53. De Livera AM, Hyndman RJ, Snyder RD (2011) Forecasting time series with complex seasonal patterns using exponential smoothing. *J Am Stat Assoc* 106(496):1513–1527
54. Esling P, Agon C (2012) Time-series data mining. *ACM Comput Surv* 45(1):12
55. Cliff AD, Ord JK (1981) Spatial processes: models & applications. Taylor & Francis, London
56. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
