Bike Rental and Weather Data across Dozens of Cities

Meelis Kull   MEELIS.KULL@BRISTOL.AC.UK

Intelligent Systems Laboratory, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

César Ferri   CFERRA@DSIC.UPV.ES

Departament de Sistemes Informàtics i Computació, Camí de Vera s/n, 46022, València, Spain

Adolfo Martínez-Usó   ADMARUS@DSIC.UPV.ES

Departament de Sistemes Informàtics i Computació, Camí de Vera s/n, 46022, València, Spain

Abstract

Effectiveness of an urban bike rental service depends on how well-planned are the layout of stations and the balancing activities which take bikes from full stations to empty stations. Layout and balancing decisions require the analysis and forecast of bike rental demand. To this date we are not aware of any public bike rental datasets which would cover many cities to facilitate large-scale analyses of demand. We provide a dataset of bike rental station status logs from 3584 stations in 27 cities of 11 countries, over the period of 7 months. The dataset is accompanied with weather information from the same cities over the same period. We have performed some example analyses to demonstrate that the dataset provides a rich source for many types of analyses, including analyses about the weekly profiles, station dependencies and relationships between weather and demand. We plan to provide updates to this dataset on a regular basis.

1. Introduction

Bicycle sharing/renting systems for sustainable urban mobility are gaining increasing popularity in many cities as an alternative to intensive car use (DeMaio, 2009). As with any public transportation system, success relies on a well-planned layout of stations and good connectivity. The virtue of bike sharing is that any station connects to any other, people can freely choose their starting points and destinations. The matters are complicated by the possibility that stations become empty or full. A good layout

can decrease the frequency of these events, but bike rental systems still need additional balancing activities. These activities are usually implemented by vans or trucks moving bikes from (almost) full stations to (almost) empty stations (Schlote et al., 2013).

For effective planning of stations layout and balancing activities it is important to perform demand analysis and forecast (Frade & Ribeiro, 2014). There are many factors influencing demand, such as time of day, time of week, season, weather, location and events (Fanaee-T & Gama, 2013). Planning the layout of stations requires the analysis of long-term patterns in the mobility of people, whereas balancing requires short-term predictions.

Bike sharing systems are implemented in more than a thousand cities in the world (DeMaio, P. and Meddin, R., 2013) and a significant amount of literature related to these systems has been published, see the review (Fishman, 2015) and Section 6 with related work. A public dataset about the bike rental usage in Washington DC (UCI Repository, 2015) was used to set a challenge to predict the number of rented bikes in a given hour (Kaggle, 2015). Another dataset has been published about Valencia, Spain, by a challenge which evaluates prediction of the number of bikes in stations with only 1 month of historical information, three hours in advance. The challenge promotes model reuse and provides models learned from other stations of Valencia with 2 years of history (REFRAME Project, 2015). However, there do not seem to be any public datasets involving many cities.

In this paper we propose a novel dataset with bike rental data from 3584 rental stations in 27 cities of 11 countries (Kull et al., 2015). The data span 7 months (from Sept 26, 2014 to Apr 26, 2015) and are accompanied with weather data from the same cities in the same period. The data collection is being continued and we plan to provide updates to the dataset on a regular basis.

We have performed some example analyses to demonstrate that the proposed dataset provides a rich source for many types of approaches. After presenting a detailed description of the data in Sections 2 and 3, we have discussed in Section 4 the bike rental demand estimation and profiles and analysed the differences in the predictive power of weekly (and daily) demand profiles across cities. We have also shown locational dependencies of demand with closer stations correlating on the average stronger than distant stations. In Section 5 we demonstrate that both temperature and the amount of rain correlate significantly with the bike rental demand. Currently the collected time-series are too short for long term trends and seasonality analyses, but we plan to provide updates on a regular basis. The related work is described in Section 6 and the paper closes with the conclusions in Section 7.

2. The Proposed Public Dataset

With this paper we present a novel public dataset of bike rental and weather information across dozens of cities. The dataset is publicly available at http://tinyurl.com/mo8e6co and it has been compiled using the following open data services:

- JCDecaux open data service for bike rental information (http://developer.jcdecaux.com). This service delivers both static data (station position, number of bike stands, payment terminal availability) and dynamic data (station state, number of available bikes, number of free bike stands and timestamp). Static data can be downloaded manually in a text file format or accessed through the API. Dynamic data are refreshed every minute and can be accessed only through the API by means of a personal and free API key provided by the service. Request to the dynamic data uses a *GET* call where one of the parameters is your "apiKey" and the user gets data in *JSON* format. Data are provided under an Open Data license[1].

- OpenWeatherMap open data service api.openweathermap.org, see details at http://www.openweathermap.org/api

Data collection started on Sept 26, 2014 and has been running continuously since then. The proposed dataset has 7 months of data ending with April 26, 2015 but we plan to provide new versions of the dataset with new information on a regular basis. We are downloading bike rental data once every minute about all bike stations provided by the JCDecaux open data service. The current weather informa-

tion is downloaded once every 15 minutes about each city where the JCDecaux bike rental stations are. The information is queried by the geographical location which is obtained by taking the median of the longitudes and latitudes of the stations of that city. The detailed description of bike features is provided in Section 3 and of weather features in Section 5.

In order to minimise the amount of missing data we use 3 independent downloading servers to collect the data. Nevertheless, there are some missing values due to denials of service by the used open data services. In total, there are 0.7% of missing bike data (37 hours with no data from any stations) and 2% of missing weather data (different missing hours in different cities).

3. Description of Bike Data

The number of stations in the bike data is continuously growing. By the end of this dataset's Version 1, there are 3584 stations in 27 cities in 11 countries (at the beginning 3412 stations in 25 cities in 10 countries). Most of the stations (2295) are in the 12 cities of France, and among those most (1240) are in Paris, see Table 1. The station sizes are mostly between 10 and 50 with the median at 20 docks per station. The distributions of the station sizes for each city are shown in columns 5-7 of Table 1.

For each station we have the following data:

- **static data:** station number, name, address, geographical location (latitude, longitude), city;

- **almost static data (changing very seldom):** number of docks (stands), is the station open or not, does the station have banking facilities, is it a bonus station;

- **dynamic data:** time of last update, the number of available bikes at that time, the number of available empty docks at that time.

Although we query the status of all stations every minute, the status updates within the open data service are often less frequent (see details in Table 1). As a result, we obtain for each station a time series with varying interval. The average interval between two updates is for most stations in the range from 3 to 10 minutes with the median at 7 minutes. The distributions of the average interval for each city are shown in Table 1.

Most stations are almost always open and on the average filled to about 45%, the averages per city are given in Table 1. One of the challenges for the bike rental companies is to reduce the number of situations where a station is empty (because then the clients cannot rent a bike) or full (because then the clients cannot return a rented bike). These situations can be prevented or dealt with by taking bikes from

---

a full or nearly-full station to an empty or nearly-empty station. The average percentage of time when a station is empty or full for stations of each city are listed in Table 1.

4. Bike Demand Analyses

4.1. Demand Estimation

Our data source does not explicitly provide the number of rented bikes per time interval. However, we can get a lower estimate of this number from the updates to the number of bikes in the stations. If we see that the number of bikes in the station has decreased by k, then we know that at least k bikes have been rented in that time interval. Unfortunately, it is also possible that n bikes have been returned and $n + k$ have been rented out, for some positive integer n. Therefore, the observed decrease in the number of bikes between consecutive timepoints is a lower estimate of demand.

However, most of the demand should be captured by these decrements because out of all decrements, 71% of times the number of bikes decreases by exactly 1. The detailed histogram of changes per time interval is summarised by the following table:

<-3	-3	-2	-1	0	+1	+2	+3	>+3
<1%	1%	5%	18%	54%	15%	6%	1%	<1%

Another factor which affects our demand estimates is the activity of the bike company in balancing the number of bikes in the stations. If the company decides to remove some of the bikes from the station (and take these to another), then in our dataset it is indistinguishable from the same number of bikes being rented by the clients. The effect of the balancing activities on our demand estimates can be reduced by eliminating timepoints where the absolute change is larger than some threshold, because the company usually moves many bikes at a time. In the analyses of this paper we have not done this, because the effect is relatively small if the threshold is high, and for low thresholds too much client activity would be eliminated as well.

When the station is empty, then bikes cannot be rented and our demand index remains at zero. This is again an underestimate, because some clients might have wanted to rent a bike but we have no information about this. It would be possible to set the demand index to be not defined (a missing value) for this case, but we have not done this and instead use this lower estimate of demand.

We refer to the decrement in the number of bikes between consecutive timepoints as *pickup demand index* and to the increment as *dropoff demand index*. If the number of bikes increases (decreases) then we define the increment (decrement) to be 0. In this paper we only consider pickup demand which we hence call simply *demand index*. The de-
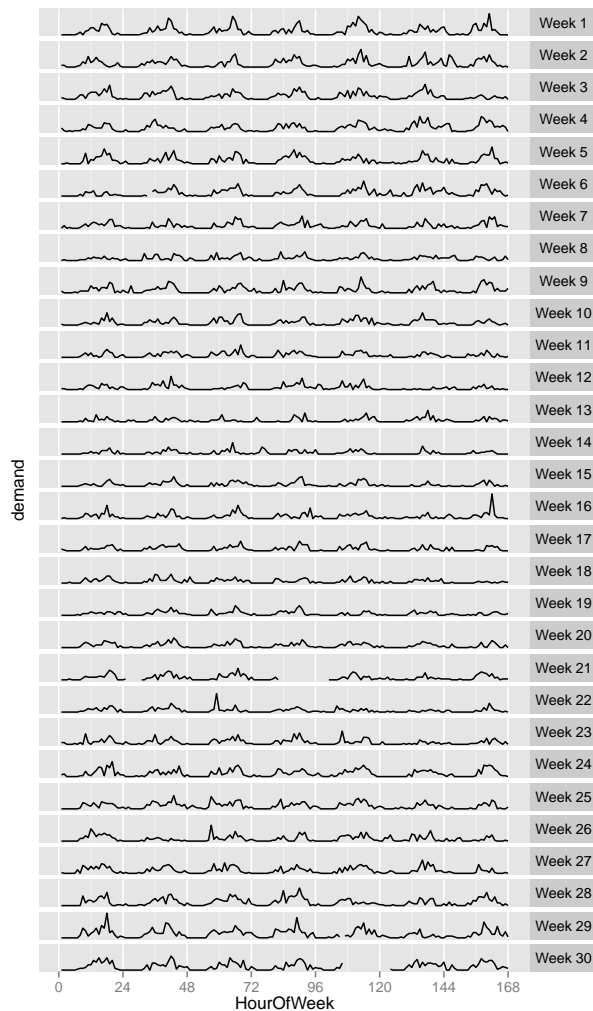

Figure 1. Demand index for the station `Paris.4002` across all 30 full weeks during 7 months. The vertical scale is the same in all rows, with maximum chosen to fit all data.

mand index for the station `Paris.4002` across all 30 full weeks during 7 months is visualised in Figure 1. The interruptions in weeks 21 and 30 are due to missing data. Lower curves in the middle rows show that demand is lower in winter.

4.2. Demand Profiles

It is well known that bike rental data have a strong weekly periodic component (Fishman, 2015). Therefore, we consider the weekly demand profiles. To obtain these, we first aggregate our data to obtain demand at every hour. Demand index at an hour is calculated as the sum of decrements across that hour.

#	city	country	# sta-tions	#docks (percentiles)			%open	%filled	update frequency in minutes (percentiles)			% empty	%full
				10th	50th	90th			10th	50th	90th		
1	Paris	FR	1240	18	30	52	100%	43 %	3.9	5.7	7.8	4 %	4 %
2	Bruxelles	BE	348	20	25	25	100%	45 %	6.4	8.8	9.7	3 %	2 %
3	Lyon	FR	347	13	19	29	100%	46 %	4.1	6.2	8.5	5 %	7 %
4	Toulouse	FR	281	16	20	24	100%	45 %	4.8	7.2	9.4	4 %	4 %
5	Valence	ES	276	15	20	25	100%	41 %	4.6	6.4	8.3	7 %	4 %
6	Seville	ES	260	15	20	25	100%	42 %	5.3	7.1	9.3	6 %	5 %
7	Marseille	FR	124	10	14	19	100%	40 %	7.3	8.4	9.4	8 %	4 %
8	Nantes	FR	103	14	15	25	100%	46 %	5.9	7.8	8.6	4 %	4 %
9	Dublin	IE	101	20	30	40	100%	42 %	3.4	5.5	6.9	8 %	4 %
10	Luxembourg	LU	72	15	20	20	100%	50 %	8.6	9.4	9.8	1 %	1 %
11	Goteborg	SE	62	20	20	30	78%	50 %	7.4	8.3	9.1	5 %	3 %
12	Cergy-Pontoise	FR	42	16	20	27	100%	36 %	8.2	9.3	9.9	0 %	0 %
13	Mulhouse	FR	40	11	14	18	100%	41 %	8.1	9.2	9.6	2 %	1 %
14	Vilnius	LT	38	12	18	24	99%	36 %	20.2	23.4	28.2	15 %	2 %
15	Ljubljana	SI	36	17	20	23	100%	29 %	5.1	6.4	8.1	12 %	1 %
16	Nancy	FR	30	15	18	28	100%	41 %	7.6	8.2	9.1	3 %	1 %
17	Besancon	FR	30	10	12	15	100%	49 %	7.7	8.8	9.4	3 %	4 %
18	Namur	BE	26	10	15	15	100%	53 %	8.7	9.4	9.7	0 %	1 %
19	Amiens	FR	26	18	20	20	100%	45 %	7.7	8.4	9.3	1 %	0 %
20	Rouen	FR	22	15	20	25	100%	45 %	6.5	7.7	8.4	1 %	1 %
21	Santander	ES	18	15	20	30	100%	46 %	9.0	9.5	10.3	0 %	0 %
22	Toyama	JP	18	16	16	25	100%	46 %	8.6	9.5	10.4	0 %	0 %
23	Lund	SE	17	20	23	30	100%	55 %	7.9	9.2	9.8	0 %	0 %
24	Creteil	FR	10	19	23	34	100%	38 %	8.6	9.6	9.9	0 %	0 %
25	Kazan	RU	7	20	25	27	100%	18 %	61.5	62.9	65.6	40 %	0 %
26	Lillestrom	NO	5	20	20	20	33%	38 %	9.1	9.9	10.8	20 %	6 %
27	Stockholm	SE	5	30	30	42	62%	56 %	9.4	9.6	9.9	0 %	0 %

Table 1. Summary of 3584 bike rental stations in 27 cities of 11 countries.

A weekly demand profile is a vector of length 168 where the N-th element quantifies the average demand in the N-th hour of the week, where Monday 0am-1am is the first and Sunday 11pm-12pm is the 168th hour of the week. For a given station, the Nth element of the weekly profile is calculated as the average demand index across all Nth hours of the week in the time-series. Figure 2 shows the weekly profiles for a random selection of 30 stations. For a city, we define its weekly profile as the average of weekly profiles of all stations. Figure 3 shows the weekly profiles of all cities in our dataset.

4.3. Demand Deviations from Profiles

The amount of regularity in the rental stations can vary and some stations follow their weekly profile to a stronger extent than others. In the following analysis we quantify regularity across cities.

In order to measure regularity of a station we calculate Mean Absolute Error (MAE) between the actual demand and the demand suggested by the weekly profile. To make the errors between stations more comparable we divide this error by the mean absolute error of the constant model, to obtain Relative Absolute Error (RAE). For comparison we

calculate MAE and RAE for daily profiles as well. The daily profile is a vector of length 24 where the Nth element quantifies the average demand in the N-th hour of the day and is calculated as the average demand index across all Nth hours of the day in the time-series.

The results are presented in Table 2. After the city names in the first column, the next three columns contain the averaged MAE using all the stations in each city. The last two columns show the same but for the RAE. wk stands for the weekly profile, da for the daily profile and av for the constant model, which is worked out as the mean of the total demand. According to the table, cities with few stations such as Stockholm, Santander, Lillestrom or Creteil are the least regular in following the weekly and daily profiles, while Bruxelles is an exception to this rule. On the other hand, cities like Paris or Dublin shows the highest regularity when compared to the other cities.

4.4. Locational Dependencies

Due to shared weather and time factors, the demand time-series of different stations in the city are correlated. In addition, the correlation can be even higher due to the locations of stations. We illustrate this in Figure 4 by plotting for
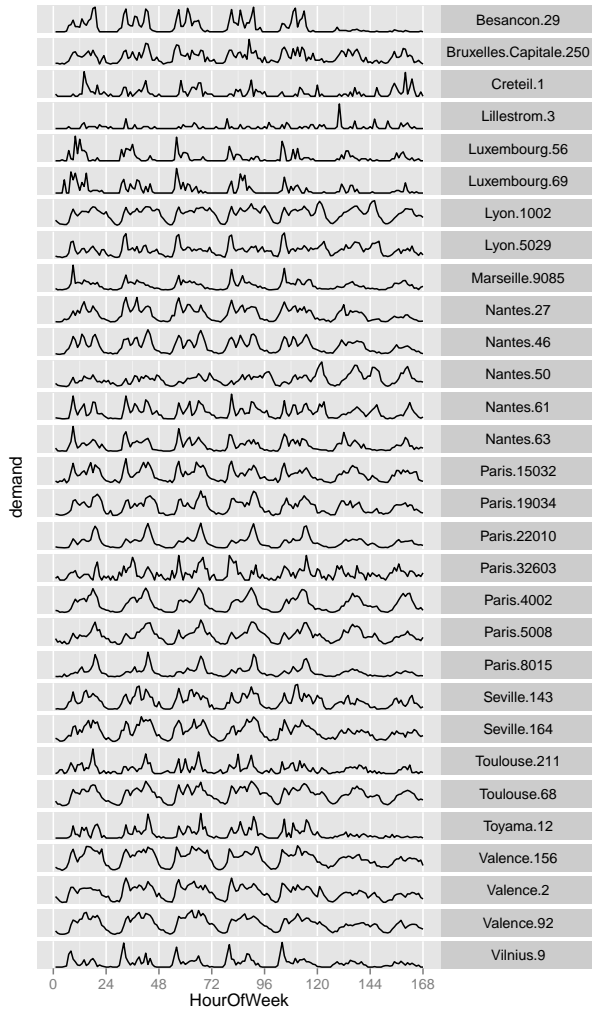
Figure 2. Weekly profiles of randomly selected 30 stations. The maximum of the vertical scale is different for each station, chosen to fit the data.
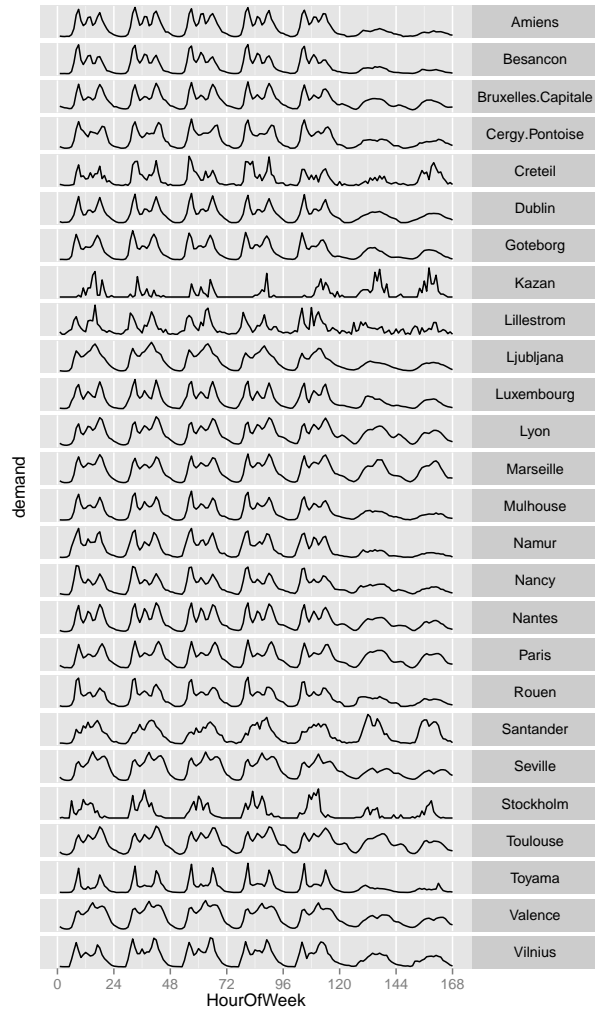

Figure 3. Weekly profiles of all the 27 cities. The maximum of the vertical scale is different for each station, chosen to fit the data.

each two stations in Dublin their distance against the correlation (across the whole time-series) of their bike rental demand indices. The plot shows that nearby stations have on the average higher correlation than distant stations.

Table 3 shows the results of the analysis across all cities with at least 10 rental stations. Average distance between two different stations varies from 1km to 6km and average correlation from 0.05 to 0.50. In all except two cities the correlation between the above-the-diagonal elements of the distance matrix and correlation matrix is negative. This indicates that on average, nearby stations have higher correlations in demand than distant stations.

5. Weather and Demand Dependency

In this section, we study dependency between demand of bikes and weather conditions. For that reason, we have collected a set of meteorological parameters for the analysed period. These features represent a complete information of the climatological conditions (temperature, rain, wind, humidity, pressure) for the set of cities where we have bike demand data. The climatological data have been obtained from the API of *openweathermap.org*[2] and the generated dataset can be found in (Kull et al., 2015). In the dataset we collect the hour averages of a wide set of climatological parameters. In this part, we concentrate on the meteorological factors that, a priori, seem to be the most promising in

---

[2]http://openweathermap.org/

city	wk_mae	da_mae	av_mae	wk_rae	da_rae
Amiens	0.59	0.71	0.95	0.69	0.80
Besancon	0.52	0.61	0.80	0.66	0.77
Bruxelles	0.49	0.54	0.62	0.82	0.89
Cergy Pont	0.24	0.26	0.32	0.79	0.86
Creteil	0.33	0.36	0.58	0.85	0.91
Dublin	1.57	1.99	3.15	0.52	0.65
Goteborg	0.85	0.93	1.09	0.79	0.86
Kazan	0.14	0.16	0.17	0.80	0.95
Lillestrom	0.59	0.64	0.86	0.90	0.93
Ljubljana	1.35	1.55	1.97	0.70	0.80
Luxembourg	0.33	0.37	0.45	0.76	0.84
Lyon	1.31	1.48	1.96	0.70	0.78
Marseille	0.64	0.69	0.86	0.77	0.82
Mulhouse	0.45	0.50	0.62	0.73	0.82
Namur	0.36	0.41	0.49	0.75	0.85
Nancy	0.73	0.82	1.00	0.73	0.82
Nantes	0.95	1.10	1.42	0.70	0.79
Paris	1.56	1.78	2.44	0.66	0.75
Rouen	0.65	0.74	1.01	0.68	0.76
Santander	0.29	0.30	0.37	0.84	0.88
Seville	0.93	1.09	1.40	0.70	0.80
Stockholm	0.43	0.47	0.71	0.84	0.89
Toulouse	1.02	1.15	1.45	0.75	0.83
Toyama	0.50	0.55	0.68	0.75	0.83
Valence	1.25	1.44	1.92	0.68	0.77
Vilnius	0.79	0.88	1.09	0.73	0.82

Table 2. Results for the set of analysed cities. From left to right, MAE of the weekly profile, MAE of the daily profile, MAE of the constant model, RAE of the weekly profile and RAE of the daily profile.

Figure 4. Pairwise correlations vs pairwise distances between stations of Dublin.

finding dependencies between demand and climatological conditions: temperature, rain/snow and wind speed. Temperature is expressed in Kelvin degrees, wind is expressed in meters per second and finally rain represents precipitation in millimetres of rain or snow.

Table 4 contains the results of the study about weather and demand dependency. The first three columns contain the average weather conditions for all the cities. According to average temperature Seville is the hottest city (288) while Kazan is the coldest (269). Dublin and Goteborg have the highest values of average wind speed while Ljubljana and Toyama present the lowest values. This last city of Japan is with Santander the most rainy cities according to precipitation average. On the other extreme Kazan presents the lowest rain mean. The following columns analyse the correlation between demand and weather factors. cor_T column has the correlation between temperatures and demand. This correlation is computed station by station, and then, we calculate the average of all stations for each city. As expected, in this case we find positive correlations since warm temperatures invite to use bicycles. However, these correlations can also represent that bikes are most used in the day light hours, when most of journeys are realised and temperatures are warmer. In order to avoid this problem
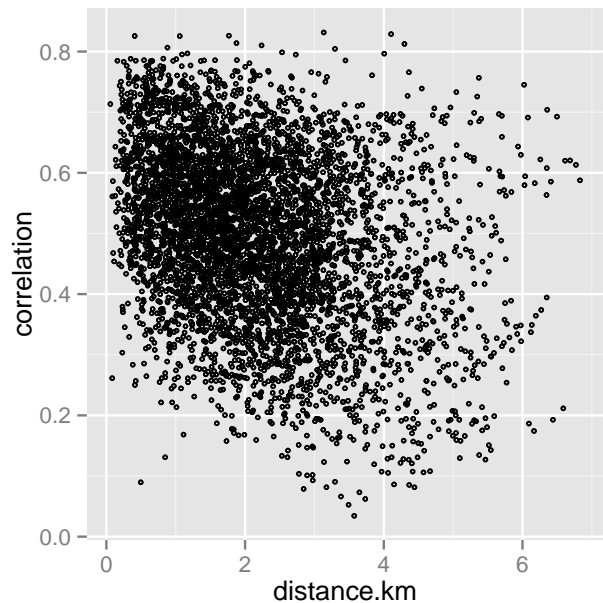
and to observe better the effect of temperature in demand, we introduce the following column cor_TH. In this case, we compute correlations between demand and temperature by the hour of the day. In this way, we have 24 correlations for each station. We then average all the results for station and city. Now this column reveals a foreseeable behaviour, the cities where we find more (positive) correlation between demand and temperatures are those with coldest averages. In fact correlation between columns cor_TH and av_T is -0.298. This reveals that in cities with warm temperatures, this factor does not affect as much as in the coldest cities. The following column cor_W represents the correlation between demand and wind speed. Here, in most cases we find a negative correlation showing that strong wind is reducing the demand of bikes. Again, we can study the relation between windy cities (av_W) and dependency on wind (cor_W). In this case we find a strong negative correlation -0.581 indicating that in windy cities, demand is more negatively affected. Finally, column cor_R includes correlation between demand and precipitation for the analysed cities. Logically, all correlations are negative since biking under rain conditions is not appealing. Again we find that in most rainy cities the negative correlations are higher (correlation between column cor_R and av_R is -0.180).

city	average distance	average correlation	correlation of distance and correlation
Amiens	0.96	0.28	0.10
Besancon	0.90	0.33	-0.10
Bruxelles	5.44	0.10	-0.32
Cergy-Pont	3.79	0.07	-0.25
Creteil	1.88	0.05	-0.10
Dublin	2.17	0.50	-0.31
Goteborg	1.63	0.33	0.04
Ljubljana	2.39	0.45	-0.62
Luxembourg	3.09	0.13	-0.23
Lyon	3.66	0.34	-0.43
Marseille	2.57	0.20	-0.28
Mulhouse	1.54	0.19	-0.37
Namur	1.44	0.14	-0.15
Nancy	1.58	0.24	-0.39
Nantes	1.81	0.31	-0.21
Paris	5.92	0.38	-0.16
Rouen	1.66	0.25	-0.40
Santander	2.97	0.08	-0.59
Seville	3.35	0.32	-0.54
Toulouse	3.33	0.24	-0.53
Toyama	1.01	0.24	-0.48
Valence	3.12	0.36	-0.41
Vilnius	1.83	0.38	-0.31

Table 3. Average distance between stations, average demand correlation between stations, and the correlation between the distance and demand correlation.

cities	av_T	av_W	av_R	cor_T	cor_TH	cor_W	cor_R
Amiens	280.35	4.49	0.31	0.14	0.09	0.003	-0.01
Besancon	279.30	2.98	0.21	0.12	0.07	-0.012	-0.05
Bruxelles	280.54	3.61	0.18	0.11	0.08	-0.008	-0.03
Cergy-P.	280.93	3.67	0.15	0.07	0.05	0.002	-0.01
Creteil	281.01	3.71	0.11	0.06	0.06	-0.007	-0.01
Dublin	279.93	6.23	0.11	0.10	0.04	-0.009	-0.02
Goteborg	278.75	6.31	0.15	0.19	0.17	-0.177	-0.07
Kazan	269.25	4.72	0.03	0.18	0.18	-0.040	-0.03
Lillestrom	275.00	3.01	0.10	0.15	0.17	0.002	-0.02
Ljubljana	277.55	1.47	0.17	0.37	0.27	0.041	-0.01
Luxembourg	278.73	3.05	0.19	0.14	0.11	-0.009	-0.03
Lyon	281.32	2.77	0.22	0.24	0.17	0.039	-0.07
Marseille	285.24	5.46	0.20	0.20	0.12	-0.050	-0.05
Mulhouse	278.87	2.39	0.14	0.14	0.09	0.001	-0.03
Namur	279.69	4.28	0.16	0.09	0.06	0.002	-0.02
Nancy	279.57	3.45	0.17	0.15	0.12	-0.029	-0.02
Nantes	282.87	3.93	0.15	0.17	0.10	-0.009	-0.05
Paris	281.06	3.71	0.12	0.22	0.16	-0.004	-0.04
Rouen	280.82	4.21	0.16	0.16	0.11	-0.002	-0.03
Santander	282.77	2.53	0.51	0.14	0.10	0.008	-0.05
Seville	288.73	2.89	0.10	0.23	0.11	0.043	-0.05
Stockholm	276.41	3.64	0.06	0.09	0.09	-0.001	-0.01
Toulouse	283.25	3.53	0.21	0.22	0.14	0.015	-0.07
Toyama	277.66	1.67	0.46	0.18	0.11	0.048	-0.06
Valence	286.35	3.79	0.07	0.28	0.14	0.035	-0.05
Vilnius	275.29	4.60	0.06	0.26	0.19	0.084	-0.08

Table 4. Results for the set of analysed cities: average temperature, rain and wind speed, correlation between demand and temperature, correlation between demand and temperature corrected by hour, correlation between demand and wind and correlation between demand and rain/snow.

6. Related work

Bike-sharing systems are gaining popularity in many cities (DeMaio, P. and Meddin, R., 2013) as an alternative to intensive car use. The first systems go back to the decade of 1960s and have experimented a rapid expansion with bike-sharing programs worldwide (DeMaio, 2009). In a bicycle sharing system, bicycles are made available for short-term (30-45 minutes) use to individuals, allowing people to borrow a bike from point A and return it at point B. A comprehensive review of recent bike-share literature can be found in (Fishman, 2015).

Most of the literature concentrates on solving the tendency of these systems to become quickly unbalanced, with some stations being mostly used to pick up bicycles and other more to return bicycles (Schlote et al., 2013). Thus, one of the problems in bike-sharing systems is the estimation of the potential demand to the service. This estimation is the main objective in (Frade & Ribeiro, 2014), where a methodology is determined according to the characteristics and trips of the Portuguese town of Coimbra. Demand and availability prediction on various time scales by means of Generalised Additive Models is performed in (Chen et al., 2013). In addition, this work also estimates the waiting time for the next available bike if the current availability is zero.

These bike-sharing systems are not only analysed according to the use of the stations but also from the point of view of studying how the location of stations can be optimised, which is one of the keys to success. In (García-Palomares

et al., 2012), authors determine the optimal location of the stations by means of using a Geographic Information System (GIS) and compare their results to the most commonly used modelling approaches. With the same aim, authors in (Hu & Liu, 2014) propose, on the one hand, to minimize the total cost of the city-bike system, including fixed cost, operating cost, passenger's travel cost and dispatch routing cost; and, on the other hand, to find an optimal location for constructing rental station and truck dispatching depot in Taiwan.

Not too many approaches have been published where exogenous variables that affect the bike-sharing system performance have been taken into account. Also in (Chen et al., 2013), authors presented prediction algorithms where the weather and time of the day were taken into account, leading to significantly improved performance. Time of the day is also studied in (Froehlich et al., 2009) as a part of their analysis of usage from Barcelona's shared bike system. In (Kaltenbrunner et al., 2010), after a comprehensive study of the mobility patterns within the bicycle program in Barcelona, the weather is only referred in the conclusions as another possible factor of influence together with events, geographic characteristics, etc. To the best of our knowledge, the only repositories for bike sharing that incorporate weather information can be found in (UCI Repository, 2015) and (REFRAME Project, 2015), where the bike-sharing program from Washington, D.C. (US) and Valencia (Spain), respectively, are complemented with weather fea-

tures. As weather really matters, we have provided weather information at city level in our dataset.

So far, all the papers introduced work in a single city schema. The first paper that takes a global view of several bike-sharing systems was published in (OBrian et al., 2014). The work offers a very interesting view of different bike-sharing policies and discusses about metrics for classifying among bicycle sharing systems. Authors analyse 38 bike-sharing programs that operate in Europe, Middle East, Asia, Australasia and America, using an extensive database that, to the best of our knowledge, is not publicly available.

7. Conclusions

In this paper we have proposed a public bike rental and weather dataset covering 3584 rental stations in 27 cities of 11 countries. To our knowledge, this is the first public dataset with bikes and weather information from multiple cities. Using some simple analyses we have demonstrated that the dataset provides a rich source for many types of analyses, facilitating development of improved methods for planning locations of stations and balancing activities. We plan to provide updates to this dataset on a regular basis.

References

Chen, Bei, Pinelli, F., Sinn, M., Botea, A., and Calabrese, F. Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. In *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on*, pp. 53–58, Oct 2013.

DeMaio, Paul. Bike-sharing: History, impacts, models of provision, and future. *The Journal of Public Transportation*, 12(4):41–56, 2009.

DeMaio, P. and Meddin, R. The bike-sharing world map, 2013. URL http://bit.ly/K9pKmO.

Fanaee-T, Hadi and Gama, Joao. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pp. 1–15, 2013.

Fishman, Elliot. Bikeshare: A review of recent literature. *Transport Reviews*, pp. 1–22, 2015.

Frade, Inês and Ribeiro, Anabela. Bicycle sharing systems demand. *Procedia - Social and Behavioral Sciences*, 111:518–527, 2014.

Froehlich, Jon, Neumann, Joachim, and Oliver, Nuria. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI, pp. 1420–1426, 2009.

García-Palomares, J.C., Gutiérrez, J., and Latorre, M. Optimizing the location of stations in bike-sharing programs: A GIS approach. *Journal of Applied Geography*, (32): 235–246, 2012.

Hu, Shou-Ren and Liu, Chao-Tang. An optimal location model for a bicycle sharing program with truck dispatching consideration. In *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1775–1780, Oct 2014.

Kaggle. Bike sharing demand competition, 2015. URL http://www.kaggle.com/c/bike-sharing-demand.

Kaltenbrunner, Andreas, Meza, Rodrigo, Grivolla, Jens, Codina, Joan, and Banchs, Rafael. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

Kull, Meelis, Ferri, César, and Martínez-Usó, Adolfo. The bike-sharing demand, 2015. URL http://users.dsic.upv.es/~flip/BikeSharingDemand/.

OBrian, Cheshire, and Batty. Mining bicycle sharing data for generating insights into sustainable transport systems. *The Journal of Transport Geography*, (34):262–273, 2014.

REFRAME Project. MoReBikeS: 2015 ECML-PKDD Challenge on Model Reuse with Bike rental Station data, 2015. URL http://reframe-d2k.org/index.php/Challenge.

Schlote, A., Chen, B., Sinn, M., and Shorten, R. The effect of feedback in the assignment problem in shared bicycle systems. In *International Conference on Connected Vehicles and Expo (ICCVE)*, pp. 960–961, Dec 2013.

UCI Repository. Bike sharing dataset data set, 2015. URL http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.