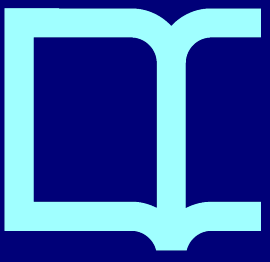

Urban cycles and mobility patterns
Exploring and predicting trends in a bicycle-based public transport system

Andreas Kaltenbrunner[a], Rodrigo Meza[a], Jens Grivolla[a], Joan Codina[a], Rafael Banchs[a]

[a] *Fundació Barcelona Media Universitat Pompeu Fabra, Diagonal 177, planta 9 08018 Barcelona, Spain*

Abstract

This paper provides an analysis of human mobility data in an urban area using the amount of available bikes in the stations of the community bicycle program *Bicing* in Barcelona. Based on data sampled from the operator's website, it is possible to detect temporal and geographic mobility patterns within the city. These patterns are applied to predict the number of available bikes for any station some minutes/hours ahead. The predictions could be used to improve the bicycle program and the information given to the users via the *Bicing* website.

Key words: Mobility pattern, community bicycle program, urban behavior

1. Introduction

Public bike sharing services are becoming more an more popular in the last few year. A still growing list of cities who provides such services systems can be found at the Bike-sharing world map at [1]. Since 2007 the city of Barcelona operates one of the largest bike sharing systems called *Bicing*, with about 6000 bikes distributed in about 400 station across the entire city. The system was very successful with more than 180.000 subscribers in 2009 according to a recent study performed by Barcelona's city council Lopez (2009). However, the same study also addresses the result of a consumer satisfaction study, which shows still some room for improvements. The two biggest problems detected, which cause user frustration, are (a) the impossibility to find a bike when a users wants to start his/her journey and (b) the impossibility to leave the bike in the user's destination due to empty or full stations. Without oversizing the system, there are basically two ways to solve these problems: Inform the user in advance about the best places to pick-up or leave the bikes and improve the redistribution of bikes from full to empty stations.

[1] http:// bike-sharing.blogspot.com

In this study we aim to contribute to the solution of these problems via the analysis of cyclic mobility patterns which lead to short term predictions of the number of available bikes in the stations. Such predictions would allow to improve the current web-service of *Bicing* and in turn increase users satisfaction with the system. Once this type of information is available, users may use mobile devices to access it. Knowledge of those patterns could lead to an optimization of the *Bicing* system itself, allowing the operator to predict shortage or overflow of bicycles in certain stations well in advance and adapt its redistribution schedule accordingly on the fly.

Furthermore we intend to show that this type of data allows also to infer the activity cycles of Barcelona's population as well as the spatio-temporal distribution of their displacements. Such knowledge may be interesting for city planers and may also represent a cheap way to compare the activity cycles between different cities.

To achieve these goal we use spatio temporal data, which has been obtained by a web mining process from the *Bicing* website and corresponds to the number of bicycles available for the users in a certain moment in time in every one of the approximately 400 different stations.

The rest of paper is organized as follows. We first review related work on the subject in 1.1 and give a more detailed description of the *Bicing* system in section 1.2. Afterwards we describe details of the data retrieval (section 1.3) and basic quantities of the collected data (section 1.4). In the results part of the article we first describe the patterns of activity in some stations in section 2 and then take a global picture analyzing the activity cycle of the entire city measured by the amount of bicycles in the stations (section 2.2) and their variation as spatial distribution (section 2.3). Then in section 3 we apply the findings to predict future activity. Finally, we present the conclusions in section 4.

1.1. Related work

Human mobility patterns have received a certain amount of attention in recent studies. However, it is not a straightforward task to obtain data which allows a large scale study, mostly due to privacy issues. Notable exceptions where the authors were able to overcome those difficulties include the use of geotagged photos (Girardin et al., 2008) and location data of mobile phones (Reades et al., 2007; Gonzalez et al., 2008; Song et al., 2010), or analyzing the circulation of individual banknotes (Brockmann et al., 2006) and civil aviation traffic (Hufnagel et al., 2004) to reconstruct geo-spatial data of human displacements in different distance-scales.

Most of these studies deal with the trajectories of individuals, but often (as in the case of our data) only aggregate spatio temporal data is available (e. g. the number of persons at time x in place y). An example for a study with such type of data can be found in (Reades et al., 2007). It uses aggregate mobile phone usage data to construct activity cycles for different locations, with clear differences between working day and weekend patterns as well as a characterization of certain areas within the city by a cluster analysis. Our study

shows how such results can be obtained as well via web-mining techniques from bike-sharing websites.

A similar yet less extensive study which does not include activity prediction has been performed by Froehlich et al. (2008).

Prediction of *Bicing* activity is a problem related to traffic congestion control, which has been analyzed traditionally for vehicular traffic. See for example (Hoogendoorn and Bovy, 2001) for a review on this subject. Related problems have also been investigated in the context of web-server traffic congestion where time series analysis techniques, especially the auto-regressive integrated moving average model or variants are widely used (Groschwitz and Polyzos, 1994; Papagiannaki et al., 2005), although other function approximation techniques, spanning from linear fits (Baryshnikov et al., 2005) to recurrent neural networks (Aussem and Murtagh, 2001), have been applied as well to obtain predictions. Here we use a technique based on activity cycles more related to (Kaltenbrunner et al., 2007) where different patterns reflecting a websites activity cycle were used to predict the number of comments a news-item would receive and implement as well time series analysis methods Box et al. (1990) in the form of an Auto Regressive Moving Average (ARMA) model.

When data in the from of individual trajectories is available a recent study Song et al. (2010) explored the possibility (and limits) of predicting a persons position using his/her previous mobility data.

1.2. Bicing

Bicing is an urban community bicycle program, managed and maintained in partnership by the city council of Barcelona and the *Clear Channel Communications* Corporation. *Bicing* is mainly oriented to cover small and medium daily routes of users within the Barcelona city area.

Users register into the system paying a fixed amount for a yearly subscription and receive an RFID Card that allows them an unlimited usage through the year, where the first half hour of usage is free and subsequent half hour intervals are charged at 0.30 euros up to a maximum of 2 hours. Exceeding this period is penalized with 3 euros per hour. There are approximately 400 stations distributed all through the city, where each station has a fixed number of slots, either empty (without a bicycle), occupied (holding a bicycle) or out of service, either because the slot itself or the bicycle it contains is marked as damaged. Whenever a subscriber needs to use a bicycle, he must select one from a station with occupied slots, travel to his destiny station, and leave it there on a free slot. The system registers every time a user takes or parks a bike in a slot. Bicycles can be withdrawn from the stations from Monday to Friday between 5:00 and 24:00. On Saturday and Sunday the service is open 24h. Outside of these time windows the bicycles can only be returned but not withdrawn.

There are two cases in which the system does not allow a user to fulfill his route:

1. The origin station does not have any available bicycles.
2. The destiny station does not have any empty slots to park in.

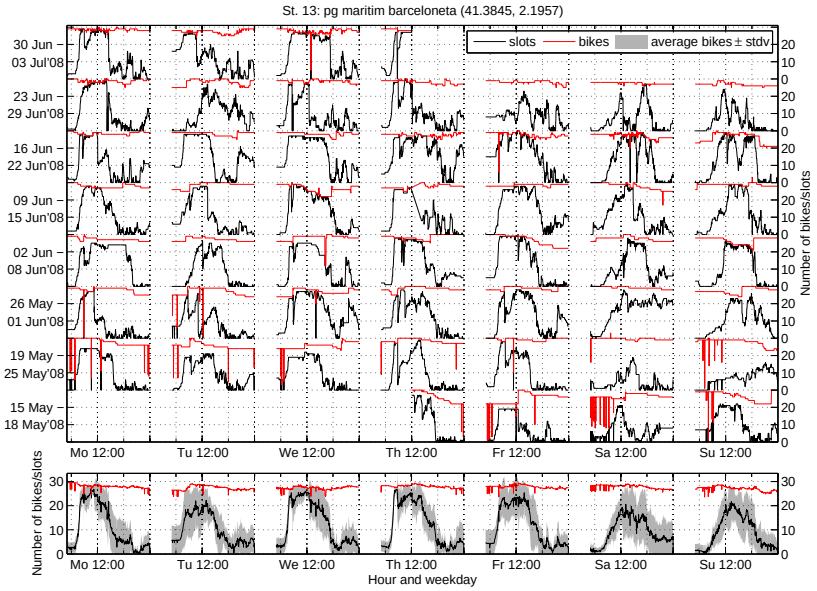

Figure 1: Time-series of the number of available bicycles (black line), and the total number of slots (red line) in an example station next to the beach. The two time-series are grouped by weeks and ordered bottom up according to their temporal sequence. The bottom row shows the average over these weekly patterns for this station. Gray areas correspond to mean ± one stdv.

When any of these situations occur, users needing a free slot or a bicycle have to choose between waiting at the station, going to another station or take other means of transportation. In order to reduce these type of situations, there are trucks which move bicycles from highly loaded stations to empty ones. However, in practice users do not wait for these trucks since they do not have a fixed schedule nor ensure a maximum response time to fix problems at a station.

To allow users to plan their routes in advance, the *Bicing* system provides on their website a map of stations[2], where users can check the status of the stations (amount of available bikes and empty slots) close to their departure and arrival points. However, this information is only available at the specific moment when the user queries the system. The service does not provide a history of previous loads to the stations[3] or an expected load of the destiny station at the time that the user gets to it.

[2] `www.bicing.com/localizaciones/localizaciones.php`

[3] A nice personal project (`http://statistings.com`) improves the service by providing the daily progression of the number of bicycles in the stations.

1.3. Data retrieval

The *Bicing* website provides an information service for users through the Google maps API. It shows a map of Barcelona overlayed with small markers indicating station positions and the amount of available bicycles and free slots for every station. Data is inserted into the map using JavaScript code with a string variable that contains a KML geospatial annotation document. This KML document defines the next information for each station:

1. station name
2. graphic icon to be inserted in the map
3. latitude and longitude
4. number of available bicycles
5. number of free slots

In order to analyze the dynamics of station loads, we have been collecting these KML documents since May 15th every two minutes, parsing it and storing in a MySQL database all the relevant information, such as the station name, localization, available bicycles and free slots. As the *Bicing* network changes from time to time, new stations are added automatically to the database when they first appear in the KML files collected from the *Bicing* website.

1.4. Basic quantities of the data collected

Due to a problem in the *Bicing* web-service, data after the 3rd of July was updated only once or twice a day and could not be used for our study. We base our results therefore on the data recollected during the 7 weeks between 12:00, May 15th and 12:00, July 3rd, 2008. We also initially did not collect data during *Bicing*'s closing hours on weekdays between 0:01 and 5:00, which restricts our analysis further to the time-window between 5:00 and 24:00.

In total, we collected data from 377 stations with a total of approximate 8700 free slots (three stations, which never contained any bicycles, were omitted from the analysis). The number of slots per station varies between 15 and 39 and the maximum amount of bicycles in the stations observed in our data was 3657.

2. Activity cycles

After having explained the data we are going to use in this study we will analyze it in this and the following sections. We will start with an analysis of the activity cycles we can obtain from the amount of bicycles available at the different stations. First, we focus on the local cycles, one for every station. We will later aggregate these cycles to infer activity cycles of Barcelona's population in 2.2. When taking into account the geographic distribution these cycles allow to visualize the mobility patterns of the city as we will show in 2.3.

We will later examine the usefulness of these cycles to predict the future amount of bicycles in the stations in section 3.

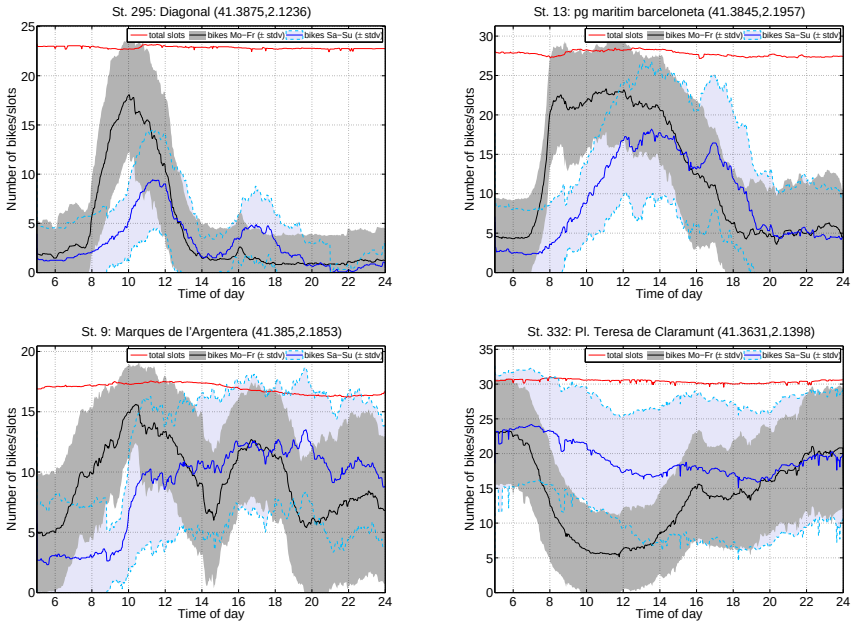

Figure 2: Average number of available bicycles during working days (black), and weekends (blue lines) for four example stations with different types of activity cycles. Red curve gives the average total number of slots in the station. Gray and blue areas correspond to mean ± one stdv.

2.1. Local activity cycles

Before we begin calculating activity cycles we take a closer look at the data recovered from *Bicing*'s web-service.

The top plot in Figure 1 shows an example of the recovered time-series data from a station close to the beach, a hospital and some office and university buildings. The recollection started on Thursday, 15-05-2008 (bottom of the subfigure) and subsequent weeks are drawn with an offset towards the top of the figure. The black lines indicate the amount of available bicycles. For control reasons we also draw the sum of bicycles and empty slots (red line), which in case the station were 100% operationally should correspond to the total number of its slots. However, since often some slots or bicycles are marked as defect and cannot be used, the red lines show some fluctuations. Sometimes they experience a sudden drop during short time intervals (e.g. on Saturday, 17-05-2008 morning), probably caused by a sporadic malfunction in *Bicing*'s data collection system.

Although the data is quite noisy with some sudden drops in the number of bikes, maybe caused by replacement trucks which move bicycles from occupied stations to empty ones, the mean weekly activity pattern shown in the bottom subplot of Figure 1, allows to average out those fluctuations quite well. We therefore have chosen to ignore those unpredictable truck events in the rest of

this study. The relative small standard deviations (black areas) show that the observed patterns are quite stable during the 7 weeks of data we analyzed. Note especially the near zero deviation at the sharp rise in the morning which can be observed from Monday to Friday. The greater standard deviation of the Tuesday pattern is caused by the local holiday on June 24th, whose bicycle pattern is more similar to those of a typical Sunday. We clearly observe two different patterns for weekend and working days.

This is confirmed by a more detailed analysis of these two patterns in Figure 2, where weekend (blue lines) and weekday patterns (black lines) from four different stations are compared. To calculate those patterns we first delete all the elements of the time-series where the total number of slots in the station is below a certain threshold (10). This allows to eliminate most of the moments where we believe the data to be erroneous (e.g. the drops in the red line in Figure 1). We then average those filtered time-series over the days of the corresponding categories and apply a median filter with window length 3 to filter the noise further.

We first focus only on the weekday patterns. The top right subplot corresponds to the station analyzed in Figure 1 in more detail. We observe very different patterns in the different stations. Station #295 (top left) is close to a university and shows a quite narrow peak in the number of bicycle in the station between 8:00 and 13:00, typical for a university with morning classes only. The following two stations are also close to universities (top right and middle left subplots). However, their observed patterns are somehow different. All three stations show the initial rise in activity in the morning. Sharp in station #13 (top right) and less pronounced in #9 (bottom left). Station #13 is also close to some important office buildings and a hospital which might explain the sharp raise in activity around 8:00, more prone to a fixed working schedule in companies or hospitals than varying starting hours of university classes. The location close to the beach probably causes the lower decay in the number of bikes in the afternoon hours where beach traffic collides with the leaving students and office and sanitary workers. Station #9 shows more variability. Although more spread than station #295 the morning peak is quite similar. However, this station experiences a second peak starting at 15:00 and reaching its maximum at 16:00 in the afternoon, This might either be caused by people leaving the university to take their lunch elsewhere or a change of shift between morning and afternoon lesson students. Finally, this station also experiences an increase in activity after 20:00 caused with high probability by the popular close-by area of bars and restaurants called "Born".

Finally, station #332 (bottom right subplot) shows an opposite cycle compared to the previous ones, typical for residential areas, where people leave the region during the morning to return later in the afternoon or late evening.

The onsets of activity in the weekend patterns (blue lines) occur later than during working days, or is nearly absent as can be observed for example in station #332, where only some minor activity is observed. Station #295 shows an interesting bimodal distribution on weekends, which might be caused by a nearby shopping center which attracts afternoon visitors on weekends.

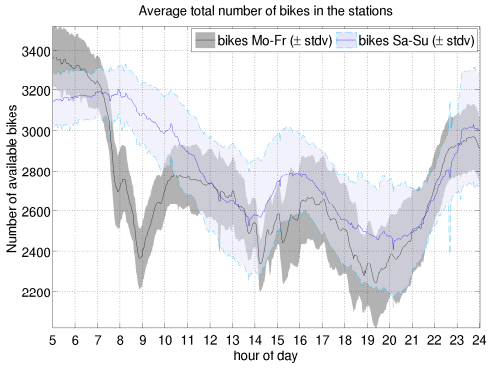

Figure 3: Average of the total amount of bicycles available in the stations.

2.2. Global activity cycles

If we look instead of the local cycles in the particular stations at the sum of bicycles available at all stations during a certain hour of the day, we get an idea of the global mobility cycle of Barcelona.

In Figure 3 we plot these average cycles of available bicycles for the working days from Monday to Friday (black curve) and the weekend (blue curve). To filter the worst noise out of the data, caused by malfunctions in the system, we use only measurements where the total sum of slots (free and occupied) in all the stations is greater than 8000 and furthermore we apply again a median filter with a window length of 3 to achieve smother curves.

The less bicycles are available for rent in the station the more displacements using them are being performed. First, we analyze the traffic during working days (black line). We observe a first local minimum (i.e. a local maximum in displacements) a little earlier than 8:00, and a second lower one at 9:00. These two minima correspond to the typical starting hours in offices, which in Barcelona varies normally between 8:00 and 10:00. This is further confirmed by the fact that the curve reaches a local maximum at this hour, the time when late starters finally reach their working or study locations. A third lower minimum is observed around 14:00, which might be caused by students who leave their classes. The number of available bicycles increases during people's lunch breaks (typically between 14:00 and 16:00), but when the local maximum at the end of this time span is reached it decays again. Finally, the global minimum number of available bicycles (the maximum in displacements) is reached slightly after 19:00 in the afternoon. Typical finishing time of many working schedules.

The weekend pattern is different in the sense that it does not show the early morning minima. Instead we observe the maximum of available bicycles around 8:00, the equinox between late home-comers from the last parties and early birds starting their day with a bicycle ride. The use of the bikes steadily augments until their number in the stations reaches a local minimum at 14:00 just before lunch time, during which it increases again. Afterwards the number

of available bicycles decays again and follows a similar pattern as during working days, although the local maximum at 16:00 occurs slightly earlier and the global minimum slightly later (at 20:00) and is less pronounced than during working days. It is therefore difficult to separate working day from weekend activity only based on afternoon activity, as can be observed as well for most of the stations presented in Figure 2.

Note that initially we only collected data between 5:00 and 24:00, which corresponds to the opening hours of *Bicing* from Monday to Friday. However, although the users are not allowed to withdraw from a station outside of this time schedule, they can return a bicycle also between 24:00 and 5:00. This explains the difference in the number of bicycles available at the beginning and end of the above described cycle.

The small standard deviations (gray and blue areas in Figure 3) show that the observed cycles are quite stable throughout the period the data was collected. The weekend deviation is slightly greater than its working day counterpart which is caused by the greater number of working days in our data set (35 vs 14) and the more flexible personal time-schedules on weekends.

2.3. Mobility patterns

To get a spatial picture of the mobility pattern in the city, we use these local activity cycles together with the stations geo-coordinates (longitude and latitude) and place the difference in the number of bicycles in the stations compared to their amount at 5:00 on the map of Barcelona for different times of the day. Afterwards we interpolate a 3D surface using this difference as color-encoded height[4]. Red stands for a positive difference, i.e. more bikes can be found in this stations than at the beginning of the day, while blue regions show areas whose number of bicycles has been reduced. Green areas indicate a more or less constant relation between incoming and outgoing bicycles. Figure 4 shows such geo-patterns for 6 different hours using the stations working day cycle[5]. At 6:30 (top left subfigure), no big difference form the initial distribution of bicycles in the stations can be observed. At 9:30 however (top right subfigure), just after the morning minimum in Figure 3, we observe quite a different picture. Several areas change color either into deep red or dark blue. Blues regions correspond to mainly residential areas, from which people move out, while the red hot-spots are found mainly close to university and business quarters[6]. Interestingly, although the number of bicycles in the station increases by roughly 400 until 12:00 in Figure 3, the snapshot of the geo-pattern (not shown) at this moment in time does not change very much. The only noticeable difference is that in already red

[4] Alternatively one can repeat the same procedure with other starting times (e.g. 16:00 to emphasize afternoon patterns).

[5] A similar but simpler spatio temporal visualization by Fabien Girardin using just the evolution of bicycles in the stations during one day can be found at `http://www.girardin.org/fabien/tracing/bicing/`

[6] For a comparison with land-use date see pages 38 (for university areas) and 42 (areas with high commercial activity) of (Rueda, 2002).

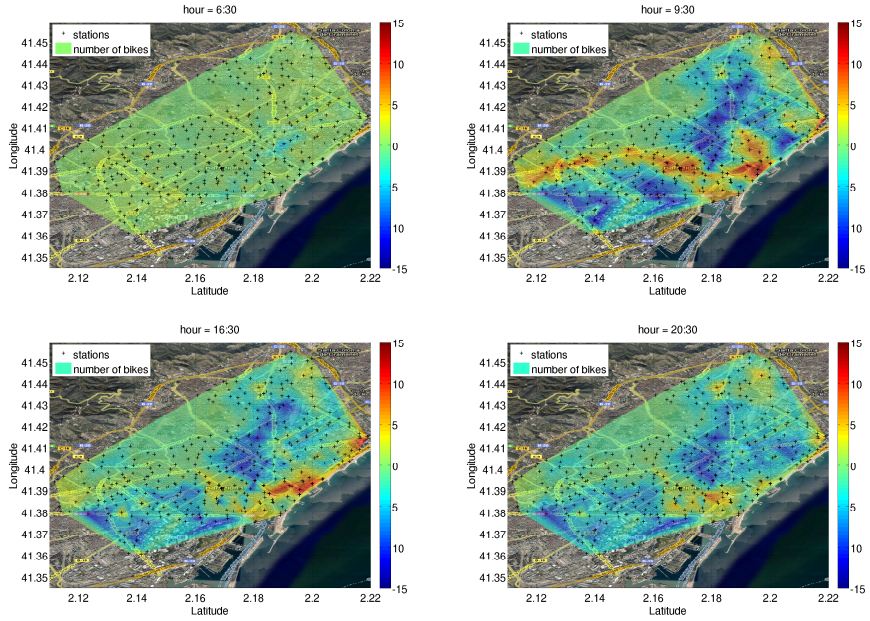

Figure 4: Geographic mobility patterns: Black crosses indicate the location of *Bicing* station and the color-overlay the average variation during working days in the number of available bikes from the level at 5:00. Blue tones indicate regions which loose bikes while red tones stations which increase their number of bicycles.

regions the amount of bicycles increases slightly even more. We can conclude that the morning peak in activity leads to quite a narrow band of stations with high bicycle concentration. The band crosses the city starting at its westmost entrance, where one the mayor university area of the city lies, and follows the Diagonal through a business district towards Passeig the Gracia, where it turns right and heads down passing by one of the mayor business and shopping areas and the University of Barcelona to meet the city center and later the sea. There it turns left again to follow the beach towards Port Olympic, leaving out one station in the also mainly residential area of Barceloneta and passing by several campuses of Universitat Pompeu Fabra. Close to Port Olympic we also find important office buildings as well as in a narrow band which grows from there northwards towards Glories. Another area which receives a big surplus in activity is Diagonal Mar, the east-most point of Barcelona, also a region with important business activity and a large shopping center.

In the afternoon the picture changes, at 16:30 (bottom left subfigure) a lot of bicycles have moved away from the previously described hot-spots, and the residential areas get some of their lost bikes back. Only the regions close to Port Olympic remains deeply red, probably now caused mainly by beach traffic. Also Diagonal Mar maintains its bicycles. At 20:30 (bottom right), finally, also those bikes head home again, only some regions in the city center still have a

surplus of bicycles, probably caused by people enjoying Barcelona's nightlife. Those regions maintain their bicycles still at 23:30 (not shown) when most of the remaining stations have recovered all their bikes and their original green tones. Those stations will recover their bikes during the night.

3. Prediction of activity

In this section, we present initial results on the prediction of bicycles or free slots at a given station at a given time. We compare several simple prediction models, and establish evaluation measurements as well as a baseline with which other (more complex) models can be compared. We then present a more advanced time-series analysis technique that can use information not only from the given station but also its surroundings.

3.1. Basic predictors

Our initial set of prediction models is based on the current state of the station as well as aggregate statistics of the station's usage patterns. As the simplest baseline we chose to predict the current state of the station (number of bikes or free slots) for any time in the future. If there are currently 5 available bicycles, the system will predict that in 10 minutes there will still be 5 bicycles available. This corresponds to the best prediction algorithm one can apply using only the present situation as displayed on the actual *Bicing* website.

The next set of models is based on extrapolating from the current state using the tendencies registered on other dates. To the current number of bikes we add the expected change based on the average gradient in the aggregate model. The aggregate model in this case can be based on all days other than the one for which predictions are made[7], or can be limited to the same day of the week, or split between weekdays and weekends/holidays.

We evaluate the different models by measuring the mean error (difference between predicted and actual availability of bicycles) over all stations and all available dates. This is done for different time offsets, i.e. predicting 10 minutes, 20 minutes, or several hours into the future.

Figure 5 (top) shows an example for the fit obtained using the baseline model (i.e. predicting the current state 2 hours into the future) and (middle) a gradient based prediction (using only data of to the same day of the week) for one particular station and day. The blue curve corresponds to the actual number of bicycles (filtered with a median filter) in the station, while the red one indicates the prediction. In this example we achieve a much lower prediction error (indicated by the light blue areas) using the gradient of the average activity cycle (green curve in Figure 5 middle).

This is confirmed further by Figure 5 (bottom) where we compare the overall performance of our prediction algorithms as explained above. For very short

[7] In a real application setting this would obviously be limited to days prior to the current date.

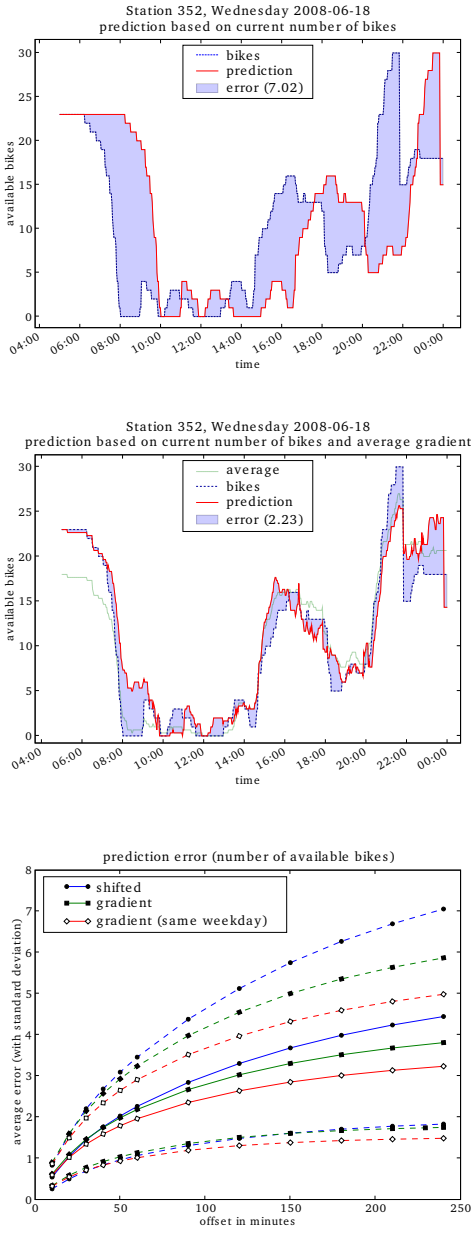

Figure 5: Prediction of bicycle availability, (top) using only the current value and (middle) adjusted by the average gradient over the other weeks. The bottom subfigure shows the average error depending on the time offset of the prediction (2 hours for the example day of station #352 shown in the middle and left subfigure).

periods (10 minutes) there is no notable difference between the baseline and other models, which may partly be due to a large number of low activity stations where predicting no change is the safest bet for very short time scales. However, we notice a significantly better performance of prediction algorithms using the activity cycles for larger offsets.

3.2. Using time series analysis for prediction

Differently from the approach using the daily average variations of available bicycles at each station as explicative variable for prediction, we explore now the use of time series analysis methods Box et al. (1990) for predicting bicycle availability at the stations.

More specifically, an Auto Regressive Moving Average (ARMA) model will be considered for implementing the predictor. This specific approach allows for taking into account the recent history of both, the current station and its closest surrounding stations, to predict bicycle availability. As its name implies, an ARMA model incorporates two fundamental models: an Auto Regressive (AR) component which is able to exploit relevant information related to the autocorrelated nature of the time series, and a Moving Average (MA) model which is able to incorporate information from additional sources of information generally denominated "inputs".

A general form for an ARMA estimator is as follows:

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{j=1}^{m} \sum_{i=1}^{q} b_{(i,j)} I_{t-i}^{j}$$

where X is the time series to be predicted, p and q are the orders of the auto regressive and moving average models, respectively, m is the total number of "input" time series I_j, t is the time index for each time series, and a_i and $b_{(i,j)}$ are the model coefficients that have to be computed during the training phase.

ARMA models are trained by means of an optimization procedure aiming at minimizing the fitting error within selected training dataset. In our case, we selected a continuous section of about 800 hours of data for training the models. Similarly, a non-overlapping data section of 30 hours was selected for evaluating prediction quality, which was measured in terms of the average absolute error.

In all experiments presented here, we used a history of 20 minutes (10 samples) of both, the same station the predictions are generated for (AR component) and the surrounding stations (MA component) to generate the predictions. Hence, the corresponding orders of our ARMA model are $q = p = 10$. Before applying time series analysis to station data, all time series were smoothed with a FIR low-pass filter based on a Hamming window.

The first experimental result, which is depicted in Figure 6, was intended to determine the optimal number of surrounding stations to be used in order to achieve minimum prediction error. The bars presented in the figure show the average absolute error over the 30 hours of evaluation data for a set of 10 different stations when considering 1, 5, 10, 15, 20, 25 and 30 surrounding stations for training the prediction models. In all cases, the predictions were

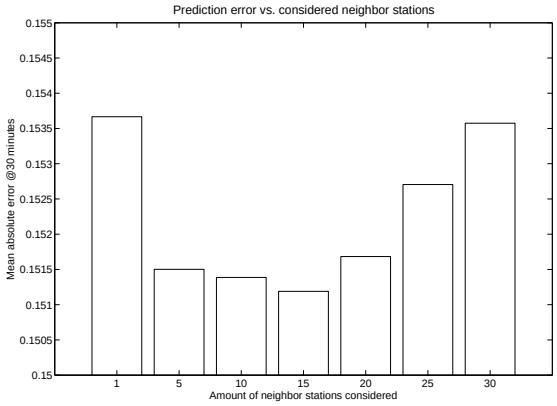

Figure 6: Dependence of the average prediction error on the number of surrounding stations used.

generated for a time interval of 30 minutes ahead. As seen from the figure, optimal prediction is achieved when considering the information related to the 15 surrounding stations, while including the information related to only the closest station or too many stations (25 and over) significantly deteriorates the prediction error. This result reveals that the dynamics of neighboring stations definitively have an important incidence on the ability of predicting bicycle availability at a given station. Further experimentation has shown that, in general, considering a number of surrounding stations between 5 and 20 will provide a good predictive power.

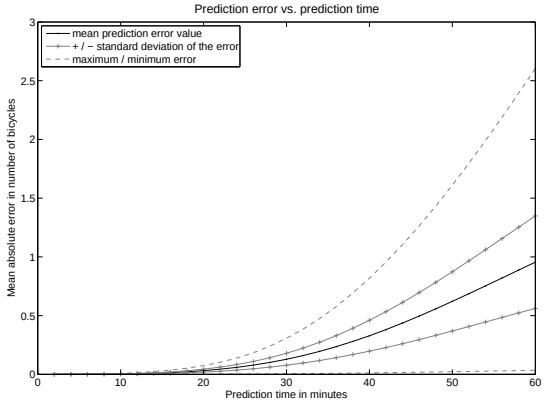

Figure 7: Average prediction error depending on prediction interval, with maximum/minimum error and standard deviation.

In the second experimental result, we evaluate how the prediction error increases as the time interval for predictions is increased. Figure 7 shows the increment in the mean absolute error when the prediction time interval varies between 2 minutes (just the next sample) up to 60 minutes. In all cases, the predictions were generated by considering the 5 closest surrounding stations. Curves in the figure illustrate the average, standard deviation, maximum and minimum error values for prediction errors computed over 357 stations out of the 377 available stations (20 stations had to be discarded because of the amount of noise and errors in their corresponding data series). As seen from the figure, at a 30 minute prediction interval, the average prediction error is below 1 bicycle, reaching a maximum value of 3 bicycles after one hour interval. It is important to mention at this point that, although predictors are indeed providing good estimates, such a small error values for time intervals below 20 minutes are also a consequence of the low-pass filtering applied to the data.

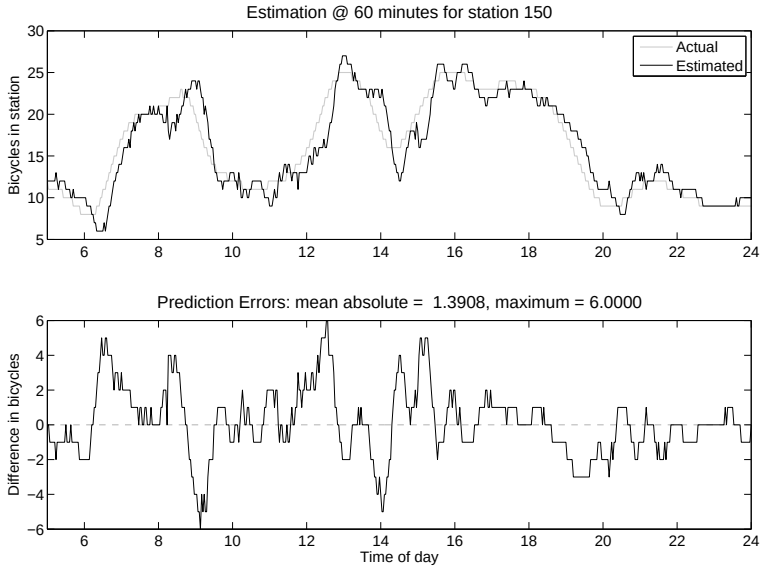

Figure 8: Prediction vs. actual bike availability for station #150, 60 minutes in advance, using data from 5 surrounding stations. Bottom subplot shows the difference between actual and predicted number of bicycles

Finally, a specific prediction curve is presented in Figure 8. In this case, predicted and the actual number of available bicycles at station #150 over the 30 hours interval of evaluation data is presented. The prediction presented here was generated for a time interval of 60 minutes ahead and was computed by considering the closest 5 surrounding stations. In the lower panel of the figure, the prediction error curve is provided. Notice that, although mean absolute

prediction error is relatively low (1.39 bicycles), the maximum error value along the prediction interval is 6 bicycles.

The results are not directly comparable with those obtained using the basic gradient based predictors, due to the evaluation being done on one 30-hour period whereas the previous evaluation was averaged over all days between 2008-05-15 and 2008-06-29, as well as slight differences in the smoothing and cleaning of the original data. However, it seems that ARMA can provide an important improvement over simpler methods (which already improve significantly over the baseline of only knowing the current state). It is particularly interesting to see the significant positive impact of using the neighboring stations to improve the prediction.

4. Conclusions

We have shown that mining usage data from community bicycle services allows to infer the activity cycles of a large city's population as well as the spatio-temporal distribution of their displacements. There are clear patterns of user behavior by station and type of day. Visualization of the average daily variation in activity allows to observe that stations with similar behavior also often correspond to adjacent areas in the map revealing residential, university and leisure areas. The cycles allow a prediction of the amount of available bicycles in the stations, which is significantly better for time windows greater than 20 minutes than the current approach on the *Bicing* website where only the actual number of bicycles/free slots is shown. Use of more sophisticated time-series analysis techniques (ARMA) and in particular the incorporation of information from surrounding stations allows to improve these prediction further.

Many enhancements and other approaches remain to be tested, including the incorporation of knowledge about interventions of *Bicing* trucks and other events that deviate from the "normal" trend into the more successful prediction methods. Weather conditions and many other factors (events, geographic characteristics, etc.) may also be taken into account.

We believe that the findings on predicting the amount of bicycles in the stations could easily lead to an improvement of the *Bicing* web site's bike availability information, by including a short time outlook into the future. It may also help to improve the *Bicing* service itself, avoiding a future empty or full station through an improved manual redistribution of the bikes via trucks. Both aspects would help to improve user satisfaction with the service, and make people more likely to use *Bicing* with important sustainability impacts.

The knowledge gained from analyzing the mobility patterns in Barcelona could be very helpful in planning the future deployment of the *Bicing* system throughout the city as well as identifying hotspots in the current infrastructure. The predictive models could be applied to resource optimization, in particular in relation with route planning of maintenance trucks and balancing of bicycle distribution in the city.

It would be interesting to contrast our results with more specific usage statistics. The *Bicing* system must internally produce more information that is not public, such as the origin/destination of individual users. Access to this data would allow to produce more precise models and make better predictions. Other information is not even available to the *Bicing* operator: e.g. the users that could not take/leave a bicycle because the station was empty/full. A survey aimed at obtaining a more detailed picture of the *Bicing* users and their motivations, currently being carried out by Jon Froehlich et al.[8], could help uncover this information.

A growing number of community bicycle services are appearing world wide[9], some of them with a similar web-service as the one we used to obtain our data, which is sure to generate increasing interest in this research topic. We are currently collecting data from many cities around Europe in order to do a comparative study of activity patterns between these cities.

Acknowledgments

We thank Fabien Girardin for pointing out some references and four anonymous referees for helping to improve this manuscript.

References

Aussem, A., Murtagh, F., 2001. Web traffic demand forecasting using wavelet-based multiscale decomposition. International Journal Of Intelligent Systems 16 (2), 215–236.

Baryshnikov, Y., Coffman, E. G., Pierre, G., Rubenstein, D., Squillante, M., Yimwadsana, T., 2005. Predictability of web-server traffic congestion. In: Proceedings of the WCW'05. IEEE Computer Society, Washington, DC, USA, pp. 97–103.

Box, G., Pelham, E., Jenkins, G., 1990. Time Series Analysis, Forecasting and Control, 3rd Edition. Prentice-Hall.

Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439 (7075), 462–465.

Froehlich, J., Neumann, J., Oliver, N., 2008. Measuring the pulse of the city through shared bicycle programs. In: Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems (UrbanSense08).

[8] `https://catalysttools.washington.edu/webq/survey/jfroehli/56481`

[9] A huge list of such services can be found at the Bike-sharing world map at `http://bike-sharing.blogspot.com`.

Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., Blat, J., 2008. Digital footprinting: Uncovering tourists with user-generated content. IEEE Pervasive Computing 7 (4), 36–43.

Gonzalez, M. C., Hidalgo, C. A., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.

Groschwitz, N., Polyzos, G., 1994. A time series model of long-term NSFNET backbone traffic. In: Proceedings of the IEEE International Conference on Communications (ICC'94). pp. 1400–4.

Hoogendoorn, S. P., Bovy, P. H. L., 2001. State-of-the-art of vehicular traffic flow modelling. Proceedings of the I MECH E Part I Journal of Systems & Control in Engineer 215, 283–303.

Hufnagel, L., Brockmann, D., Geisel, T., 2004. Forecast and control of epidemics in a globalized world. PNAS 101, 15124–15129.

Kaltenbrunner, A., Gómez, V., López, V., 2007. Description and prediction of slashdot activity. In: Proceedings of the 5th Latin American Web Congress (LA-WEB 2007). IEEE Computer Society, Santiago de Chile.

Lopez, A., 2009. El transporte publico individual de barcelona. In: II Jornadas de la Bicicleta Publica. Sevilla, Spain.
URL `http://www.bicicletapublica.org/PDF/angel_lopez.pdf`

Papagiannaki, K., Taft, N., Zhang, Z. L., Diot, C., 2005. Long-term forecasting of Internet backbone traffic. IEEE Transactions On Neural Networks 16, 1110–1124.

Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C., 2007. Cellular census: Explorations in urban data collection. IEEE Pervasive Computing 6 (3), 30–38.

Rueda, S., 2002. Barcelona, ciutat mediterrnia, compacta i complexa. Una visi de futur ms sostenible. Ed. Ajuntament de Barcelona.
URL `http://www.bcn.cat/agenda21/publicacions/ColleccioAgenda21.htm`

Song, C., Qu, Z., Blumm, N., Barabasi, A. L., 2010. Limits of predictability in human mobility. Science 327 (5968), 1018.