# Human mobility patterns and community detection in Bicing Public Bicycle Sharing system

Antonio Gonzalez Quintela & Cristina Palmero Cantariño

***Abstract*—The aim of this project is to obtain a first approach of the analysis of human mobility data in an urban area using Barcelona's sharing bicycling system called Bicing, in order to infer human mobility patterns and measure the pulse of the city. A spatio-temporal analysis of four weeks of usage is provided in order to uncover daily routines, cultural influences and the role of time and space in city dynamics. Activity information is then used to create a weighted undirected network which compares the activity patterns among stations. Finally, community detection techniques are applied so as to reflect these human patterns and deduce some conclusions from them.**

## I. Introduction

City-wide urban infrastructures are increasingly reliant on networked technology to improve and expand their services. Recent literature work has shown the value of sensing these digital footprints to uncover new insights into human behaviour, urban dynamics, and tourist movements [23]–[28]. In this paper, the underlying patterns of Barcelona are studied through the so-called public bicycle sharing service Bicing.

### A. Smart cities

The big urbanization process that the population around the world is experiencing provokes that most of the problems which humanity has to deal with are related to cities. It is estimated that cities will account for nearly 90% of global population growth, 80% of wealth creation, and 60% of total energy consumption in a few years.

Due to these difficulties, a brand new branch of technology which tries to tackle them is emerging, named Smart Cities. The main idea of the Smart cities is to use new technology and science to make the 'physical capital' (infrastructures) interact in a more fluent way with the 'social capital' in order to improve both of them in an interleaved fashion. Smart Cities are to improve some key issues such as economy, mobility, environment, people interactions, living and governance, although some of them overlap. In a more technical way of speaking, the challenges span from a new way to collect information, store data or process it to a development of pure science and new mathematical tools. Specifically, the former challenge tries to deal with vast quantities of raw information generated by citizens collected every day. And here it is when the term Big Data is introduced. This new level of information makes it possible for civic leaders to use data to better inform a city's daily operations and long-term planning. Cities can do a whole host of things with data from monitoring daily traffic conditions to assessing the future viability of a city's healthcare and education programs.

### B. Complex systems

When we talk about the more scientific developments in this way we are talking of complex systems. Smart cities try to deal with highly interacting systems which behave dissimilarly and owe some properties at different scales. The problems that these systems arise are nowadays a big challenge in most state-of-the-art issues and an important source of data to tackle real difficulties.

Complex systems try to deal with problems of dynamic systems with a lot of different parts which can interact in many different ways. This type of systems exhibit some properties such as self-organization, adaptability or power-law distributions which are commonly found in nature.

As the idiom *'we can not watch a film only watching one pixel'* claims, we need to understand all the possible interactions to be able to study complex systems, which evolve at the edge of criticality. In order to accomplish this task, Big Data is useful, which is not easily reducible in size without a loss of information. This could mean that the inherent behaviour which produces this data is a complex system. In this sense both walk together, and it is complex science which can develop tools and generate insights to handle and understand Big Data problems.

In words of Geoffrey West in Scientific American, *"'Big data' without a 'big theory' to go with it loses much of its potency and usefulness, potentially generating new unintended consequences"*.

### C. Mobility

Mobility acts as one of the most important concerns that can be faced applying these systems. In cities human mobility represents lots of great payoffs which could be: save energy, increase productivity or even ease the interaction between zones removing or breaking down social barriers. There exist many solutions proposed by the public policies around the world, which include the improvement of the public transport, more pedestrian streets, Public Bicycle Sharing services, promotion of carpooling and building the correct infrastructures to allow easy transportation between the different zones.

It is also a very cross-topic in the world of science. Mobility problems are closely linked with epidemic prevention, urban planning or emerging response.

### D. Public Bicycle Sharing

The history in public bicycle sharing (PBS from now on) services starts in the 60s in Amsterdam, a city with deep tradition in bicycle transportation, but it was not until


Fig. 1. Bicing station.

1998 when the technology made high impact in this kind of programs when Rennes starts its own PBS program.

This technology allows to implement the location of the stations and quantity of bicycles and spots in each station in a data-driven way. It require ways of transforming all the data recollected from the service into useful knowledge which can optimize the system and improve the user experience. The correct selection of all this parameters is a crucial question for the success of a PBS system.

One of the main issues of this system is the tradeoff between the number of bicycles which serves all the users and the costs of this extra offer in the services. For instance, it is possible to find a lack of bicycles in some stations during several periods of the day whereas others are lacking of free spots. In order to increase the efficiency from a economical point of view without losing any quality of the service a dynamic optimization has to be applied.

The correct use of the gathered information allows to predict the fluctuating demand in order to anticipate patterns and use it to balance the network. This task is done by motorized redistribution vehicles (trucks) and can be done in two different ways:

- In *static mode* the redistribution is carried out during the night.
- In *dynamic mode* the redistribution is carried out during the day. The bicycle repositioning is performed based on a prediction which has to be computed.

### *E. Bicing*

Bicing is the PBS system installed Barcelona city, mainly oriented to cover small to 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 which give them unlimited access throughout the year. The conditions of this system is that the first half hour of usage is free and afterwards subsequent half hour intervals are charged with an amount of 0.30 euros up to a maximum of 2 hours. After these 2 hours the penalization cost increase and it is recorded by the system. Too many penalizations of this type could suppose the loss of the service rights.

Bicycles are checked out by swiping an RFID membership card at a Bicing station kiosk (Figure 1), which then unlocks a bicycle and displays its rack location on an LCD screen. Check-out information is uploaded to a web server that provides real-time information about the number of available bicycles and vacant slots at each station. The stations are distributed along the whole city and there are around 400 stations so far (figure 2). The number of slots per station varies between 15 and 39. These slots can be either empty (without a bicycle), occupied (holding a bike) or out of service, either because the slot itself or the bicycle it contains is marked as damaged.

Bicing service is available 365 days per year with the following timetable:

- Monday-Thursday: all day except from 02h to 05h. During this period, only bicycle returning is allowed.
- Friday: all day except from 03h to 05h. During this period, only bicycle returning is allowed.
- Saturday, Sunday and public holidays: 24 hours.

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

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

In either of these situations, users have to decide whether to wait at the station or going to another station. When the destiny station lacks of empty slots, the user has 10 additional minutes to find another station where to park the bike. In order to reduce these kind of situations, there are trucks which transport bicycles from overloaded stations to empty ones. However, these trucks do neither have a fixed scheduled nor ensure a maximum response time to fix problems at stations.

## II. This study

Human mobility patterns have received certain amount of attention in recent studies. These patterns reflect the culture and the spatial layout of the city. It is shown in [1] that individuals follow simple and reproducible patterns of mobility in their everyday displacements.

The aim of this study is fourfold:

- Obtain a description of the general patterns and activity cycles.
- Obtain a description of particular patterns regarding their layout and other characteristics such as social differences among zones of Barcelona city.
- Obtain a network identifying the most probable routes between stations.
- Propose certain improvements to the system, both in the balance system and in the user service.

The human behavior patterns and the modularity of a city where some different spaces perform the same social task (residential neighbourhoods or office areas) in different locations of the city suggest that some Bicing stations may show similar cycles depending on the activities occurring around them.

To find such similarity patterns, a study of the mobility in Barcelona is to be investigated, using Bicing system as the data baseline supplier. The drawback of this system is that it do not have a public listing of bicycle rides that

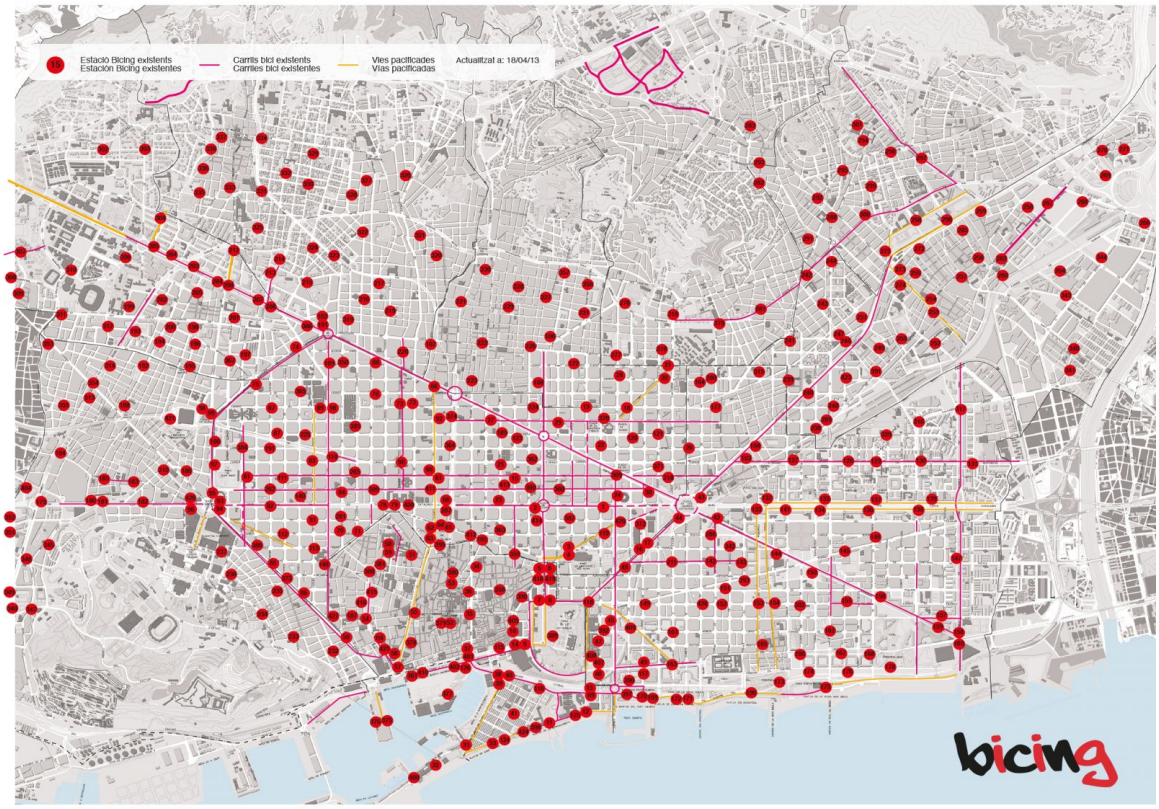


Fig. 2. Bicing service map as it was in April 2013.

users make from a departure to an arrival station, which would have eased the creation of the network. Instead, it only supplies discrete information about the number of bicycles and free slots available in each station.Hence, in this case, only aggregate spatio-temporal data is available to infer human mobility patterns. Specifically, the data going to be used to infer hyphoteses about this topic is the following for each station:

- Bike availability during time.
- Activity time-series. It is the measure of the activity for each station and its definition is the derivative of the number of bicycle time-series of a given station.
- Position of the station.
- Elevation of the station.
- Population related to the location of a given station.
- Matrix of distances in time between the stations. It is a matrix of which number of columns and rows corresponds to the number of stations, and each cell determines the time any bicycle takes to go from one station to the other (at time of delivering this report, Bicing company has not made available to us this information yet).

## III. Data gathering

In order to construct the network and analyze the dynamics of station loads, a first data retrieval stage was done. The data has been collected from CityBik API [1], which provides real-time information of the different PBS around the world.

[1]http://www.citybik.es/

TABLE I
Gathered Data from CityBik and Google APIs

| Variable | Description |
|---|---|
| id | CityBikes station ID |
| internal_id | Real station ID |
| cleaname | Station name and address |
| name | Station name |
| bikes | Number of bikes in the station |
| free | Number of free slots in the station |
| timestamp | Last time the station was updated |
| nearby_stations | ID of the closest stations |
| lat | Latitude in E6 format |
| lng | Longitude in E6 format |
| elevation | Elevation in meters |

In this case, Bicing data has been gathered in JSON format from April 29th to May 27th every 5 minutes. As the Bicing network changes from time to time, new stations are added automatically to the database when they first appear in the JSON files collected from the API website. Since the elevation data was out of the scope of that API, another gathering procedure was carried out thanks to Google's elevation API in order to fetch elevation information of each station. A Java application was developed in order to fetch the data, parse it and store in a MySQL database all the relevant information, detailed in table I.

The second data retrieval stage gathered census data divided

TABLE II
CENSUS GATHERED DATA

| Variable | Description |
|---|---|
| cod_SC | zip code |
| densidad | population density |
| poblacio2009 | population in 2009 |
| e15_24 | Age: 15-24 years |
| e25_44 | Age: 25-44 years |
| e45_64 | Age: 45-64 years |
| e65_o_mas | Age: from 65 years |
| directius | # CEOs |
| Tecnics_Cientifics_Intellectuals | # techs & researchers |
| Administratius | # administratives |
| Operaris | # operators |
| Treb_No_qualificats | # non qualified workers |
| Empresaris | # entrepreneurs |
| assariats_fixos | # salaried |
| assalariats_eventuals | # seasonal wage earners |
| tasa_segundas_residencia | second residences rate |
| tasa_oficinas | offices rate |
| tasa_comercios | business rate |
| tasa_industrias | factories rate |

in neighbourhoods, in order to relate it to the different stations and possible patterns that they show. This is detailed in table II.

## IV. DATA CLEANSING

The data scraped from Bicing is noisy as a result of temporary station closures, technical issues at the stations caused by maintenance work, internet connectivity failures, server-side SQL time-outs, and broken bicycles and parking slots. Data cleansing is therefore the earliest stage and a critical issue to ensure that the data used to create and study the network is valid and able to lead to an accurate representation of the reality. Several stations where therefore removed from the data due to extremely anomalous behaviour. For instance, stations whose number of bicycles surpass their total number of slots have been omitted. The remaining data was then filtered with a median filter of window size of value 3 in order to remove isolated service fluctuations. The theoretical intention was to fetch data every 5 minutes, but due to some of the problems mentioned before, time intervals came not to be of 5 minutes strictly. Furthermore, some information was missing in some time periods. These issues provoke that the need of apply an interpolation arose in order to have aligned time series of each station. Once applied, we obtained a preprocessed time series for each station with the number of free slots and available bicycles every 5 minutes from April 29th to May 27th, thus having 288 samples per day and station.

Regarding census data, it is assigned as features to each station directly by taking into account the proximity of the neighbourhood to the station using latitude and longitude values.

## V. ACTIVITY CYCLES

Once the data about the activity in each station is available, it can be used to infer very identifiable results, such as the aggregate behaviour of the city per station by computing the cumulative aggregate activity for all working and weekend days, resulting in two different patterns, one for working days and one for weekends, due to the fact that people tend to move through the city in a different way depending on the day of the week (during working days, mobility patterns show that people move from home to work and the other way round, whereas during weekends, social life and hobbies take place). During the data gathering period, two public holidays in working days took place, May 1st and May 20th. As they are punctual holidays (not long weekends), they could imply a different third category pattern, but in order to simplify the study these two days have been omitted. Finally, a vector of 288 samples is obtained for each station and pattern.

From this aggregate of station bike availability during time, station activity can be inferred by using a measure called Activity Score (AS), which measures how active a station is at a given time.

$$AS(t) = B_t - B_{t-1} \tag{1}$$

where $B_t$ is the number of bicycles at time $t$.

Therefore, two different time series are considered, each with two different patterns for working days and weekends:

- Activity score
- Mean number of bicycles

### A. Motorized redistribution vehicles identification

As explained in section I.D, bicycles are redistributed among the different stations by trucks that control crowded and empty stations so as to achieve an equilibrium in the system and to maintain its effectiveness. This redistribution provokes punctual peaks in the number of bikes flow of each station, so they can be detected by using a sliding window among the raw (before applying the median filter which smoothes the time-series) daily time series applying a threshold. Those peaks that surpass the threshold value would be considered as truck action. Figure 3 shows an example of a truck activity. In this case, truck removes bicycles from the station, passing from around 18 to 2 bicycles in a very short period of time (21-05-2013 at around 08h). This peaks are removed after the filtering process of the pre-processing stage, so they have to be analyzed before applying it, that is, over the raw data.

## VI. SPATIO-TEMPORAL PATTERNS

It is easy to imagine that a city as huge as Barcelona can contain different zones of different behaviour, that is, it is separated in residential zones (usually at suburbs), commercial and business districts (city center) and leisure areas (coast, Tibidabo, Montjuic), connected via narrow streets, one-way avenues and a multitude of public transportation options. It is therefore clear that these zones will have different behaviour patterns because the spatial layout of a city has an obvious influence on the movement patterns and social behaviours

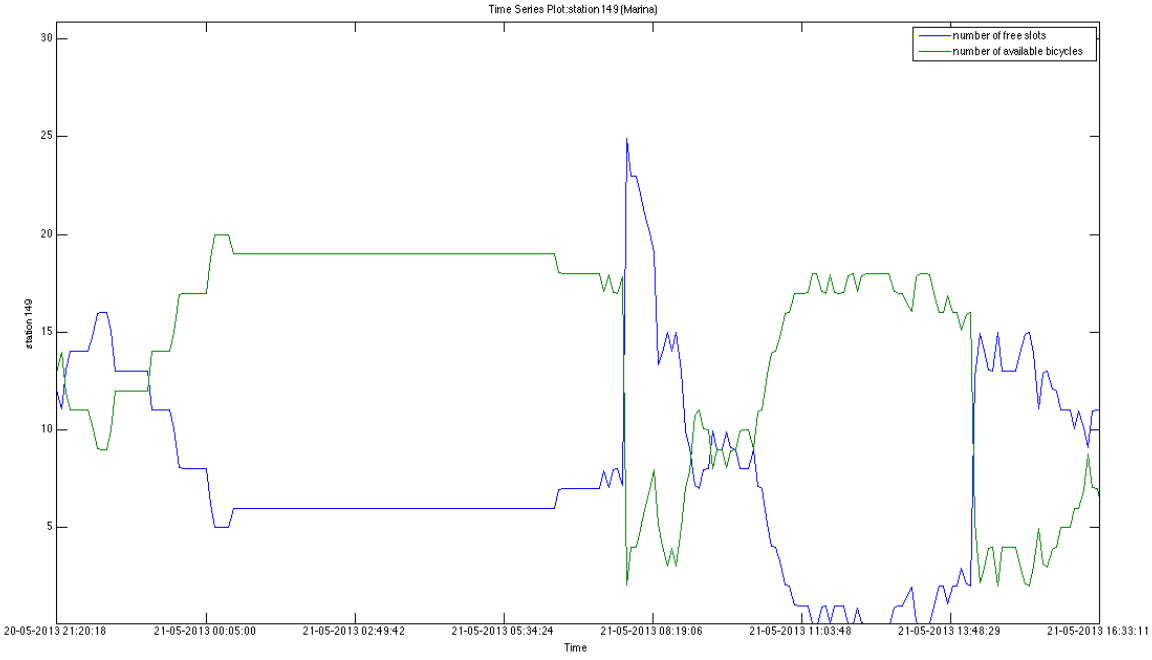


Fig. 3. Example of truck activity among raw data (before filtering) in Station 149, c/Pujades 55, Marina. In this case, a truck removes bicycles from the station, passing from around 18 to 2 bicycles in a very short period of time (21-05-2013 at around 08h), corresponding to the largest activity peak.

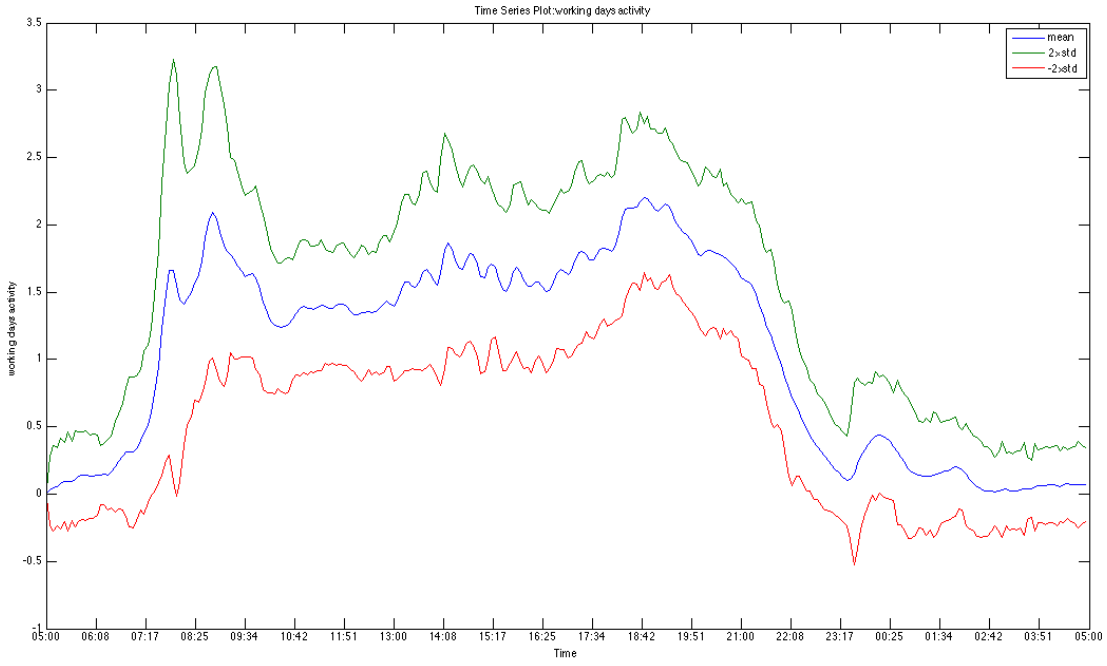


Fig. 4. Accumulated global activity for all stations during working days. Blue denotes activity average, whereas green and red denote the boundaries of two times the standard deviation.

found therein. For instance, the number of bikes during working days will decrease dramatically in the morning in residential areas, thus increasing proportionally in business districts. However, this morning commute is absent in public holidays or weekends, changing the mobility flow to leisure areas. Hence, it is logic that this dissimilarities are to be

seen among the different activity patterns of the stations. Depending on their location, positive or negative peaks of activity in punctual and identifiable times take place. Probably, all these places with patterns in common are strongly related by some features. In figure 5 we can observe an example of the differences among residential and commercial patterns.

Human mobility patterns are not only location-related, but also temporal. These temporal patterns of a city are a reflection of the daily routines of its citizens. For instance, it could be observed in 5 that there exists a "lunch spike" occurring at 14h, reflecting that Spaniards ten to eat a late lunch compared to Nordic people.

Thus, both time and space have an important role in Bicing usage, which is influenced by daily routines, culture and location of the different stations. This usage involves a multitude of underlying motivating factors such as commuting, shopping and having lunch. Figure 4 shows the accumulated global activity for all stations during working days. We can observe the most important peak during the morning when citizens moves from their residences to their work or studies location. The activity decreases afterwards maintaining a fluctuating pattern from 10h to 17h, only disturbed by the so-called lunch spike explained before. The activity increases again from 17h to around 20h when people go back home, and then it slowly decreases until 23h, when people probably take the bicycle as a night walk. After this last period of activity, the activity decreases until the end of the service.

Local and global patterns could be related to emergence [7] and self-organization [6] topics. Emergent systems have the following characteristics:

- The whole is more than the sum of the parts.
- The organization is always a bottom-up rather a top-down nature.
- Biological emergent systems often involve chemical signaling.

Self-organization can be understood in terms of the second and third stages of thermodynamics. The second stage describes a system in which the flow is linearly related to the force [4]. Such a system tends towards a steady state in which entropy production is minimized, but it depends on the capacity of the system for self-organization. In a third stage system, flow is non-linearly related to force, and the system can move far from equilibrium. This system maximizes entropy production but in so doing facilitates self-organization [5]. In this case, bicycles move along the network from station to station but in the end the number of bicycles keeps constant.

In order to identify this patterns and the similarities between stations the activity patterns of all stations will be compared by using activity time-series of each station. The result is a complete directed weighted graph with the distances given by a comparison method. So, we perform a comparison of the activity time-series of each station with each other obtaining a complete graph. The drawback of using time-series is the lack of alignment. Then, a basic time-series matching problem is considered, which compares activity patterns of all stations to the rest and place them into similar groups or clusters. This can be achieved by hierarchical or partitional clustering such as k-means or nearest neighbours. Similarity depends on the features to consider (i.e. how we will describe or compress the sequences). The benefit of non-metric distance functions is that they match flexibility, are robust to outliers, stretch in time and space and offer support for different sizes and lengths, although the speeding-up search can be difficult. They thus allow to align time series. An alignment is the association of each point in one time series to another point in the other time series. In order to compare two warped time series, the non-metric Dynamic Time Warping (DTW) will be used applying Euclidean distance as cost indicator. DTW compares two time series while accounting for the warp using Dynamic Programming. The alignment cost can then be used as pattern similarity indicator. In general the warping is resolved by accounting for added samples (the second example is performed slower) or deleting samples (the second example is performed faster). Thus, this framework allows us to classify the stations according to their temporal patterns. Local distances are computed using the Euclidean distance metric:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{2}$$

Then, the overall computational difference between the entire sequences is computed, named as global distance:

$$\begin{aligned} D(i,j) &= min[D(i-1,j-1), D(i-1,j), D(i,j-1)] \\ &+ d(i,j) \end{aligned} \tag{3}$$

DTW was used in preference to a Euclidean distance measure because we were interested in comparing overall temporal patterns and wanted to allow for up to one hour of temporal shifts in the data (bicycle rides between stations usually do not take more than 30-45 minutes in average). Thus, DTW based metric with a one-hour Sakoe-Chiba band was applied (12 samples) [11], obtainin a matrix of $424 \times 424$ distances.

Distances obtained with DTW are then transformed to weighted edges between stations. In complex networks context, community detection can be used as a network clustering method in order to obtain certain representative groups of Bicing stations that share a similar behaviour pattern. The similarity metrics defined so as to build the graph are the following:

- Considering a complete graph and pruning the self-loops (zeros in the distance matrix).

$$w_{ij} = \frac{C}{D_{ij}} \qquad \text{and} \qquad w_{ii} = 0 \tag{4}$$

  where $D_{ij}$ is the result of apply DTW comparing the time-series of the station $i$ with the one of the station $j$ and $C$ a normalization constant. It has been selected because of its simplicity and allows to keep positive weights due to its monotonic decreasing function.
- Considering the case described before but now pruning out all the edges which have a weight lower than a threshold that has to be decided. This results in a more sparse graph adequate to apply an easier community detection algorithm.


Fig. 5. Time in X axis and number of bicycles in Y axis. Top: Station 286, c/Bolivia 76, near Glories commercial centre, UPF university and many business (commercial district pattern). Bottom: Station 309, c/Sant Ramon Nonat 26, in Collblanc (residential zone pattern). We can clearly observe the differences among patterns. Marked in bold are (1) commute to work in the morning, (2) lunch time and (3) going back home in the evening.

Figure 6 left shows the weight distribution obtained by the inverse of the DTW distances. We can observe that it is approximated to a Poisson distribution. Figure 6 right studies stable communities. After applying an incremental pruning by scanning different thresholds we can observe that after a value of 150, the number of communities detected starts to increase exponentially until 400 when it saturates.

Applying community detection to the weighted graph three different clusters are obtained, shown in figure 7. We can clearly observe three different clusters depending on the station activity. Taking into account their location in the map, red nodes are related to residential areas, whereas blue nodes are mostly related to commercial and business areas. Not surprisingly, stations considered as example in 5 are correctly

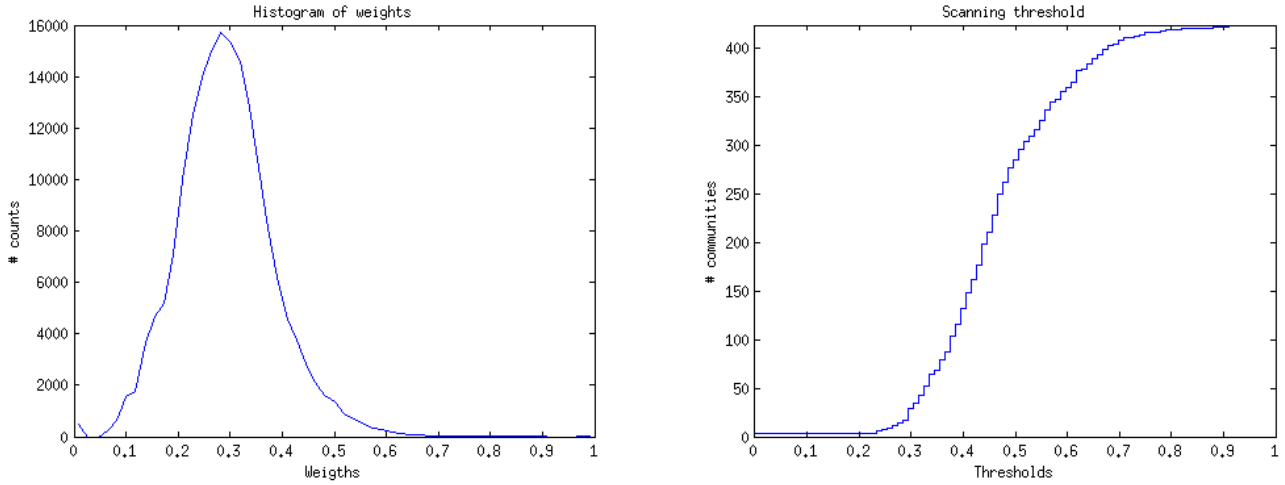


Fig. 6. Left: Histogram of weights obtained from DTW distances of working day patterns. Right: result of applying different thresholds when detecting communities.

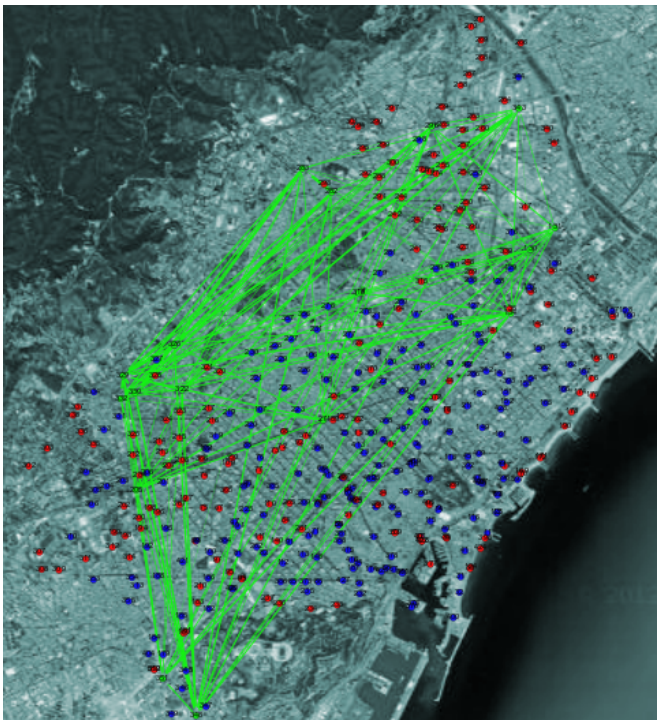

Fig. 8. Relations among stations that form the green cluster.

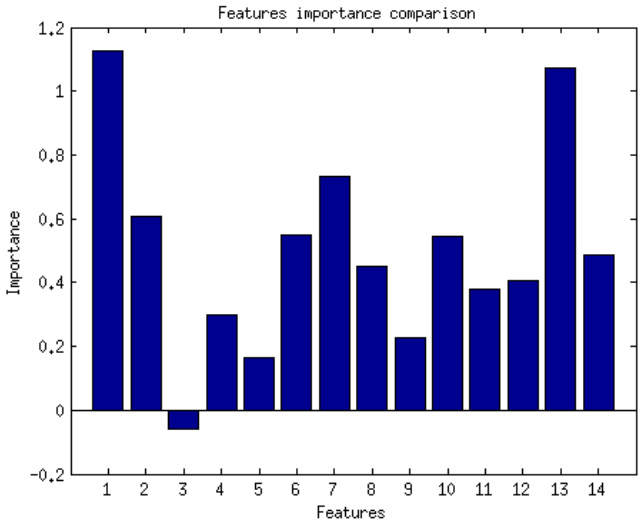


Fig. 9. Comparison between the importance of the features used in the study obtained by applying Random Forest, which are (1) elevation, (2) population density, (3) ages from 15 to 24 years, (4) ages from 65 years, (5) number of technicians and researches, (6) number of administrative people, (7) number of operating people, (8) number of non qualified workers, (9) number of entrepreneurs, (10) number of fixed salaried, (11) second residences rate, (12) offices rate, (13) commerce rate and (14) factories rate.

detected. The third cluster, marked in green, is mainly located in elevated and extreme zones, such as Montjuic, Pedralbes, Carmel and Horta, which indicates that the activity in these kind of areas is pretty low, probably because people are not used to go uphills. However, there are a few stations that are more centered. The relations inferred is that the latter stations had hardly any activity due to malfunctions of the service, therefore they have been related to those that have a few. Figure 8 better shows these relations.

Random Forest algorithm [29], [30] has been applied to the features of each station to assess their importance implied the clustering process. As we can observe in figure 9, the two most important features are elevation (1) and business rate (13), which confirms the hypothesis that community detection has claimed before by separating communities in residential and business areas and also taking into account elevated stations.

Figure 10 considers the relation between the elevation of each station and the average of the number of available bikes. It is shown that the number of stations and bicycles available decreases inversely proportional to the elevation. Anyway, the redistribution carried on by motorized vehicles may alter the original results.

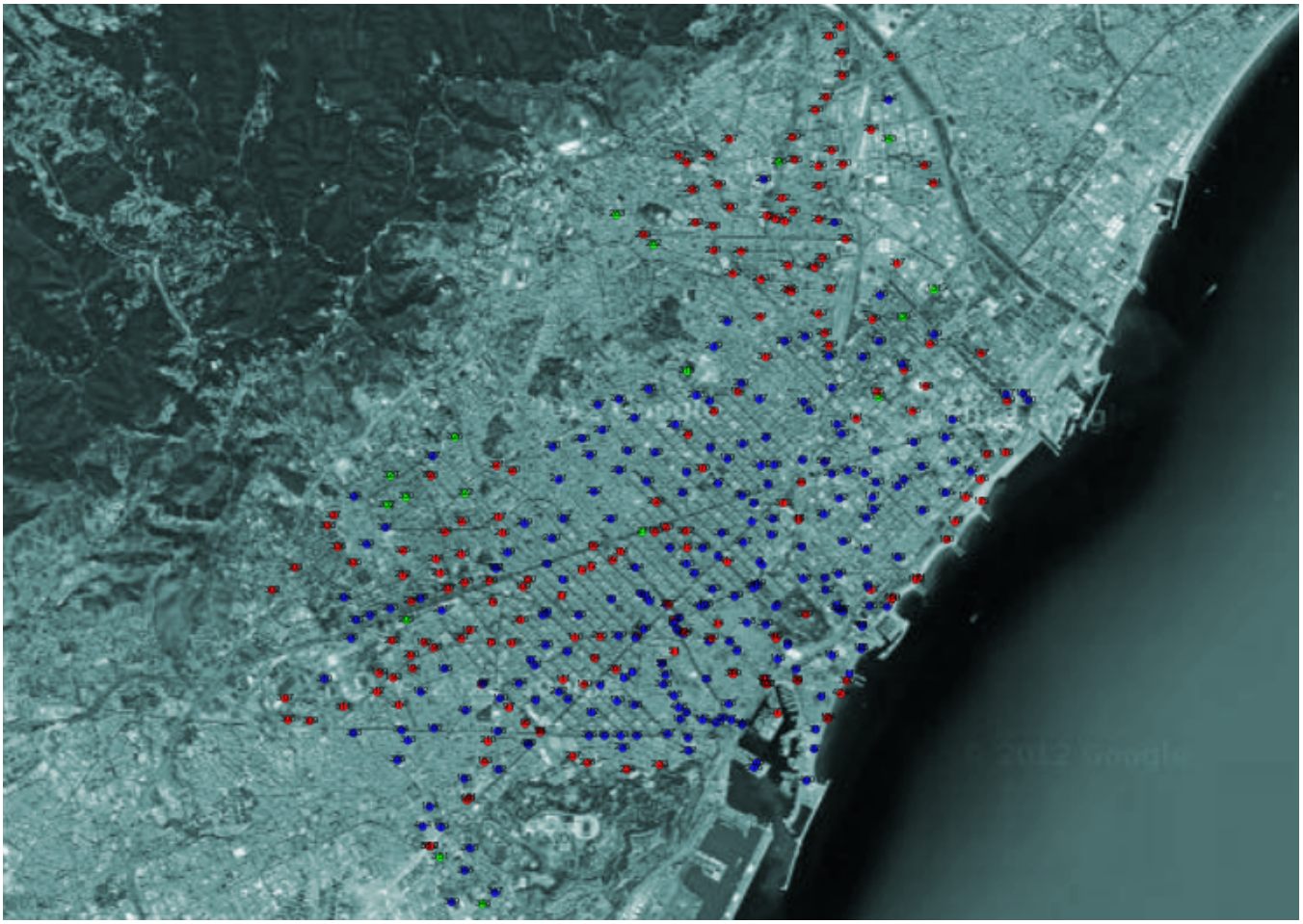

Fig. 7. Community detection applied to the weighted graph formed by the different Bicing stations, which are over-imposed in the Barcelona map by latitude and longitude coordinates. Stations are denoted as nodes. Blue nodes represent commercial and business areas whereas red ones represent residential zones. Green nodes denote stations of hardly, if any, activity, corresponding to elevated and extreme zones, but also to malfunctioning stations.

## VII. Probable route identification

According to the fact that in some periods of time there is a peak of negative activity in one station and a positive peak of activity in other, the spatial distance between both stations and the distance in time between both peaks could be highly related, thus existing a great probability that the bicycle that provoked the negative peak of the first station is the same that gave rise to the positive peak of the second station, being the former the departure station and the latter the arrival one. It is possible to deal with the problem of estimating those routes that are most likely to be transited by the users from the aggregated data that is available to us, thus recovering all the individual bicycle rides between all the stations and day periods. That is to say, not to consider only aggregated data to study the mobility patterns and therefore infer the network, but create a dynamic network taking into account all the possible variations during time. In order to create the network it is not only needed the station temporal patterns but also the relative location between them which are defined by the temporal distance of a bike ride, beginning in the departure station and finishing in the arrival one. Note that in the actual context of the service provider this problem is not interesting at all, since individual bicycle movements among stations can be tracked by the system administrator. However, when this information is not available, estimating the most popular routes from the aggregated data is a challenge. Basically, this problem is not solvable at all from the observation data allow. But we claim that it is possible to approximate suboptimal solutions for it by means of conditioning the observed data with some additional information, such as the distances among the stations, the average bicycle velocity, and some other common sense implications. For instance, users are more likely to use the service to cover intermediate distance, taking into account that using the service for less than 30 minutes is free. Moreover, citizens usually take the bicycle only for rides from high to lower elevation.

Taking the temporal series of number of bikes per station, each of them has an inherent reason to be explained. There are individual bicycles moving from departure to arrival stations. Observing directly the data we can only enquire knowledge of

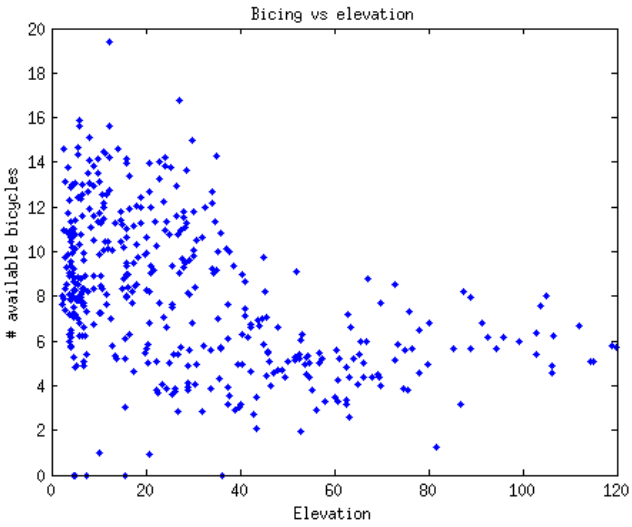


Fig. 10. Number of bicycles in each station depending on the elevation of them.

the number of bicycles each station gains or loses during time every 5 minutes as explained in section V. From this activity we can ensure that there are a minimum of this number of bicycles interacting (arriving or going away) with this station.

The objective is to link departures and arrivals of different stations in order to get a 1-factor of a bipartite graph with maximum associated value. All the edges have a weight given by the probability to be linked according to the distance in time of the two events and the probability distribution of a time duration of a bicycle ride taking off from the initial station and arriving to the final station. The formalization of the problem is:

Problem such as $G(U, V, E)$ where

- $U$: departures from the stations
- $V$: arrivals to the stations
- $E$: rides between stations

An algorithm that can be used in this type of problems has to be found, taking into account that in the worst case we will have $100000$ rides to infer which is the number of bicycle rides per day. This means a bipartite graph of $100000 \times 100000$ nodes, each half in one part of the graph [13]–[20] . Traditional algorithms blow up with this huge challenge. For instance, the traditional Munkres algorithm [2] has a complexity of $O(n^4)$ and its modification by Kuhn [3] has a complexity of $O(n^3)$. It is therefore needed to use some randomized algorithm in order to make this problem feasible.

Once the time distances matrix is available, a cut-off distribution has to be applied in order to assign probabilities to each edge. The fact of having a cut-off value causes the bipartite graph to be more sparse, thus provoking a reduction in the possible matchings search space. To optimize the $-log[p(x)]$ of probabilities, a genetic algorithm could be used.

## VIII. Proposals and future work

Due to computational cost and time limitations several open problems have been left for possible future work. The most interesting and challenging ones could be the following:

- Consider the motorized redistribution vehicles in the bicycle pattern flow, using some proposed methods in the literature as in [21], which uses Petri Nets. Since trucks tend to visit stations once in a defined period of time, a sliding window could be also used to detect their appearance, focusing in peaks which do not repeat during the defined interval.
- Assign census information using probabilities instead of computing the distance directly.
- Instead of using an aggregate of the activity for each station, split each weekday into seven time bins:
  - Early morning (05-09h).
  - Mid-morning (09-13h).
  - Lunch (13-15h).
  - Afternoon (15-17h).
  - Early evening (17-20h).
  - Late evening (20-00h).
  - Night (00-05h).

  For each time bin, calculate the average number of available bicycles, the difference between the number of bicycles at the beginning and end of the bin's edge, and the measure of the station's activity. This approach would increase the accuracy of the results and the final weighted network, thus having impact in the number of communities detected.
- Instead of splitting each weekday into time bins as proposed before, use a sliding window along the data.
- Use a community detection method which only takes into account stable nodes, that is, nodes that does not change between communities when applying pruning.
- Consider the use of a probabilistic community detection as in [22].
- Detect communities among time, as done in [31], [32] which work with *Multiplex Networks*.
- Apply hierarchical clustering to the communities detected in order to assess the different sub-communities within them.
- Ability to predict the behaviour of the system in the future, using the knowledge of the activity and the available time-series, using methods such as ARMA [33], [34], probabilistic graphical models or time delay neural networks [35] , considering exogeneous factors.

## IX. Conclusions

A first approach of an analysis of human mobility behaviour and an identification of activity patterns has been achieved. The first and probably the most matter of concern is the pre-processing of the data, which has been proved to be a key issue in order to assess accurated results. After observing them, it can be concluded that several malfunctioning stations were not omitted because their punctual cases were not taken into account in this first stage. That is, these stations do not even get to a minimum of bikes, thus showing that they are out of

work or that Bicing system is not acquiring correctly the data from them. This lack of knowledge has caused, in part, the emergence of the green cluster and the outliers shown in figure 10. It is also remarkable the importance of the data status once gathered; Bicing system is not as perfect as it could be, thus favouring the obtention of non precise data.

The obtained behaviour patterns help to conclude the clear pulse of the city in working days, to therefore infer a weighted undirected network that compares the different activities and characteristics of each station. The application of community detection techniques to this network has demonstrated the existence of underlying clusters along the system, although there could be more sub-clusters within them. Nevertheless, taking into account our priors of the system we expected a greater modularity, maybe lowered by the motorized redistribution vehicles which alter the real data flow. Note that this information cannot be exact due also to exogeneous factors, albeit the results are promising and encouraging to continue studying the proposals of section VIII.

To conclude, Big Data is more than simply a matter of size, it is an opportunity to find insights in new and emerging types of data, to better serve citizens, and to answer tough questions. Using a data-driven approach, a city can transform how it fundamentally organizes and operates, making it a better place to live, work, and play. Anyway, quantity alone never leads to quality and hence smartness. If the data we use is of low quality, then instead of leading to smarter cities we could get misinformed cities. Big data is great as long as it conforms to the quality requirements.

## REFERENCES


[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. *Understanding individual human mobility patterns.* Nature, 453(7196):779-782, June 2008.

[2] J. Munkres, *Algorithms for the Assignment and Transportation Problems*, Journal of the Society for Industrial and Applied Mathematics, 5(1):32-38, 1957 March.

[3] Harold W. Kuhn, *Variants of the Hungarian method for assignment problems*, Naval Research Logistics Quarterly, 3: 253-258, 1956.

[4] Katchalsky A, Curran PF. *Non-equilibrium Thermodynamics in Biophysics*. Cambridge, MA: Harvard University Press; 1967.

[5] Prigogine I, Stengers I. *Order Out of Chaos.* Toronto: Bantam Books; 1984

[6] Addiscott TM. *Entropy, non-linearity and hierarchy in ecosystems.* Geoderma. 2010;160:5763.

[7] Addiscott TM. *Soil mineralization: An emergent process?* Geoderma. 2010;160:3135

[8] Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., Blat, J. (2008). *Digital footprinting: uncovering the presence and movements of tourists from user-generated content.* IEEE Pervasive Computing.

[9] Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). *Mobile landscapes: Using location data from cell-phones for urban analysis.* Environment and Planning B: Planning and Design, 33(5). 727 748.

[10] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.

[11] H. Sakoe, S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* IEEE Transactions on Accoustics, Speech and Signal Processing, vol. ASSP-26, NO. 1, February 1978

[12] Berndt, Donald, and James Clifford. *Using dynamic time warping to find patterns in time series.* KDD workshop. Vol. 10. No. 16. 1994.

[13] Fukuda, Komei, and Tomomi Matsui. *Finding all minimum?cost perfect matchings in Bipartite graphs.* Networks 22.5 (1992): 461-468.

[14] Varadarajan, K. R. (1998, November). *A divide-and-conquer algorithm for min-cost perfect matching in the plane.* In Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on (pp. 320-329). IEEE.

[15] Gabow, H. N., Tarjan, R. E. (1989). *Faster scaling algorithms for network problems.* SIAM Journal on Computing, 18(5), 1013-1036.

[16] Lin, G., Tegos, T., Chen, Z. Z. (2005). *Heuristic Search in Constrained Bipartite Matching with Applications to Protein NMR Backbone Resonance Assignment*. Journal of Bioinformatics and Computational Biology, 3(06), 1331-1350.

[17] Cook, W., Rohe, A. (1999). *Computing minimum-weight perfect matchings.* INFORMS Journal on Computing, 11(2), 138-148.

[18] U. Zwick, *Lecture Notes on Maximum matching in bipartite and non bipartite graphs.* School of Computer Science, Tel Aviv University, December 2009

[19] Michel X. Goemans, *Lecture notes on bipartite matching*, Massachusetts Institute of Technology, February 2009

[20] Zhen, J., Guangming, D. (2010). *Weight Identification of a Weighted Bipartite Graph Complex Dynamical Network with Coupling Delay.* Journal of Inequalities and Applications, 2010.

[21] Labadi, K., Benarbia, T., Hamaci, S., Darcherif, A. M. (2012). *Petri Nets Models for Analysis and Control of Public Bicycle-Sharing Systems.*

[22] Ferry, J. P., Bumgarner, J. O., Ahearn, S. T. (2011, July). *Probabilistic community detection in networks.* In Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on (pp. 1-8). IEEE.

[23] Froehlich, J., Neumann, J., Oliver, N. (2008). *Measuring the pulse of the city through shared bicycle programs*. Proc. of UrbanSense08, 16-20.

[24] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R. (2010). *Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system.* Pervasive and Mobile Computing, 6(4), 455-466.

[25] Froehlich, J., Neumann, J., Oliver, N. (2009, July). *Sensing and predicting the pulse of the city through shared bicycling.* In Proceedings of the 21st international jont conference on Artifical intelligence (pp. 1420-1426). Morgan Kaufmann Publishers Inc..

[26] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R. (2008). *Bicycle cycles and mobility patterns-Exploring and characterizing data from a community bicycle program.* arXiv preprint arXiv:0810.4187.

[27] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R. (2010). *Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system.* Pervasive and Mobile Computing, 6(4), 455-466.

[28] Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R. (2010). *Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system.* Pervasive and Mobile Computing, 6(4), 455-466.

[29] Breiman, L. (2001). *Random forests.* Machine learning, 45(1), 5-32.

[30] Zhang, J., Zulkernine, M. (2006, April). *A hybrid network intrusion detection technique using random forests. In Availability, Reliability and Security*, 2006. ARES 2006. The First International Conference on (pp. 8-pp). IEEE.

[31] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., Onnela, J. P. (2010). *Community structure in time-dependent, multiscale, and multiplex networks*. Science, 328(5980), 876-878.

[32] Bansal, S., Bhowmick, S., Paymal, P. (2011). *Fast community detection for dynamic complex networks.* In Complex Networks (pp. 196-207). Springer Berlin Heidelberg.

[33] McLeod, A. I., Li, W. K. (1983). *Diagnostic checking ARMA time series models using squared?residual autocorrelations.* Journal of Time Series Analysis, 4(4), 269-273.

[34] Ng, S., Perron, P. (1995). *Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag.* Journal of the American Statistical Association, 90(429), 268-281.

[35] Huang, J. Q., Lewis, F. L. (2003). *Neural-network predictive control for nonlinear dynamic systems with time-delay.* Neural Networks, IEEE Transactions on, 14(2), 377-389.