NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

DEPARTENT OF ECONOMICS

Faculty of Economic Sciences   
Bachelor’s programme ‘Economics’   
Bachelor Thesis

**Predictive modelling of bicycle availability**

**for bicycle-sharing systems**

Maria Golovina, BEC 145

Academic supervisor:

Stankevich Ivan Pavlovich

Moscow

2018

# Contents

[Contents 2](#_Toc514154067)

[Abstract 3](#_Toc514154068)

[Introduction 4](#_Toc514154069)

[Definition of key terms 6](#_Toc514154070)

[Development of public bike-sharing systems 7](#_Toc514154071)

[Related work 9](#_Toc514154072)

[Stochastic modelling of bicycle-sharing systems 11](#_Toc514154073)

[Estimation of arrival and departure rates 14](#_Toc514154074)

[Estimation of arrival and departure rates during over-demand periods 15](#_Toc514154075)

[Data description 16](#_Toc514154076)

[Introduction of heuristic for detecting rebalancing acts 17](#_Toc514154077)

[Realization of peak detection heuristic 22](#_Toc514154078)

[Estimation of arrival and departure rates 23](#_Toc514154079)

[Simulation 25](#_Toc514154080)

[Conclusion 30](#_Toc514154081)

[References 31](#_Toc514154082)

[Appendix 33](#_Toc514154083)

# Abstract

Station-based bicycle-sharing systems allow customers to pick up a bicycle from one of the stations distributed across the city and return it to another. Effective management of these systems accounts for the major part of their operational costs and gives rise to various optimization problems, including predicting station-level demand in order to understand where bicycle rebalancing should be done and how many bicycles should be relocated in order to meet the demand for bikes and empty docks at the stations. However, measuring actual demand at the stations might be a challenging task due to their finite capacity. When the station is empty or completely full the demand for either bicycles or docks is not observed.

This study focuses on predicting unobserved station-level demand for bicycles and docks using both data mining techniques and stochastic modelling. We simulate behavior of a station by modelling spatiotemporal arrival and departure rates as Poisson processes with piecewise-constant intensity rates. We justify using an adjusted formula that counts only points in time when both rates are observed to estimate arrival and departure rates. Obtained estimates of intensity rates are used to predict unobserved demand of a simulated station. Performance analysis of this model in comparison to results yielded by several popular machine learning algorithms show that proposed model outperforms them on a time horizon of 30 minutes, but does not beat Random Forest algorithm on time horizons of 1 hour and 2 hours in terms of RMSE.

# Introduction

In the past decade bicycle-sharing systems (BSS) have received increasing attention due to growing concerns about urban traffic congestion and climate changes. Integrating bicycle infrastructure with existing public transport system has a complex impact on public transport use. For instance, it has been shown that introduction of a BSS can decrease public transport use in the central part of a city but increase it on the periphery (Shaheen et al., 2011). Nevertheless, a significant number of studies found that BSSs ease traffic congestion, with the effect being especially prominent in large cities (Wang, Zhou, 2017). Experience of various US and Chinese cities shows that BSSs can directly reduce private car and taxi use (Martin, Shahee, 2014). Introduction of a bicycle-sharing system in Washington reduced traffic by 4% (Hamilton, Wichman, 2018). To sum up, bicycle is an environmentally friendly mode of transportation that promotes healthy lifestyle and can ease urban traffic network, which explains why local authorities support and often subsidize public bike-sharing systems.

While a new generation of dockless bike-sharing start-ups is gaining popularity in Asia and initially appears to be more attractive for investors, their pilot launches outside home markets proved to be unsuccessful in many European cities. The docked scheme is expected to remain prevalent in Europe in near future (this issue is discussed in more detail in Chal), which means that the problems confronting docked BSSs that are addressed in this work do not lose their relevance.

In docked BSS users borrow bicycles from one of the stations distributed across the city and return them to the same or to a different station. Docked bike-sharing service needs to be rebalanced over time to meet the demand for bikes as well as empty docks at the stations. The lack of one of these two resources may occur due to non-uniform distribution of rides between the stations under limitations of finite station capacity, an issue sometimes referred to as asymmetric demand-offer problem. In the rest of this paper the term over-demand will be used to describe situations when this problem leads to unsatisfied demand either for bikes, i.e. the station is empty, or for docks, i.e. the station is full, so cyclists can’t park and have to ride to a nearby station.

Bicycle rebalancing is crucial for operators to retain regular clients. These customers tend to buy annual or semi-annual subscription and use bicycles to connect to public transit network on their way home or to work. If the station is prone to over-demand, the probability of a situation when a customer doesn’t find a bike or an empty dock at a nearby station increases, so the system might ultimately become too unreliable, forcing the customer to buy his own bicycle or switch to a less environmentally-friendly mode of transport. As many of the city bike-sharing systems are operated by government agencies or in public-private partnerships, some operators even get penalized by the local government in proportion to the fraction of time the stations remain full or empty (e.g. V ́elib’ in Paris [Schuijbroek et al., 2017]). As a result, they have to implement rebalancing even if it accounts for the major part of their operational costs.

The need to tackle system imbalance gives rise to various optimization problems. This work focuses on predicting station-level demand for bicycles and empty docks in a station-based bike-sharing system combining stochastic modelling and data mining approach. One of the challenges of this problem that will be addressed is estimation of demand under the limitations of censoring, i.e. when demand is observed only if there is no over-demand at the station. Stochastic modelling will be used to simulate a station and estimate bike arrival and departure rates of the system as independent Poisson processes in a time-inhomogeneous queueing model.

The rest of the paper is organized as following. First, we briefly overview the history of bike-sharing systems and explain why docked BSSs will not ubiquitously make way for dockless systems in the near future. Then we review related academic literature. In the next chapter basic properties of stochastic processes are introduced and modelling stations through Poisson processes is justified. describe the data used in this study. We bring this section forward, because the specifics of station availability data justify the attention that is paid to unobserved demand and is mainly used for visual research. Then, we introduce a stochastic model that is aimed at estimating arrival and departure rates and is based on the queue model proposed by Gast et al. (2015) and use it to simulate behavior oh a station and model unobserved demand. The scope of this study does not include other stages of bicycle rebalancing, such as determining optimal hour and an optimal or near-optimal route for relocation. Nevertheless, predicting demand for bikes is an important step that anticipates choosing an efficient repositioning strategy. The estimated numbers of demanded bicycles and empty docks are passed to the model solving pickup and delivery vehicle routing problem. The model then suggests an efficient (usually cost-efficient, but it depends on the design of the model) route for one or several vehicles.

It is worth noting that to the best knowledge of the author of this study, there are no other works that use statistic modelling to enrich historical data, making it possible to train a model that predicts actual demand without censoring, which is insightful for operators of the bike-sharing system.

# Definition of key terms

Over-demand state: a situation when the station is either empty or completely full and does not meet demand for bicycles or empty parking slots, respectively.

Over-demand: unsatisfied demand in an over-demand state, i.e. demand either for bikes when the station is empty or for docks when the station is full and a customer can’t park and has to ride to a nearby station.

Arrival and departure rates: probability of arrival per time unit and departure per time unit, respectively. This is a definition of Poisson process intensity adapted for the processes of arrival and departure of bicycles. Arrival and departure rates are also referred to as pick-up and drop-off rates in related literature.

Rebalancing act: an act of moving bicycles by operator in attempt to satisfy over-demand and avoid over-demand state or reduce its duration.

Occupancy: the number of bicycles at a station at given time.

# Development of public bike-sharing systems

The literature on development of bike-sharing systems agrees on four main generations of BSSs (Shaheen et al., 2010). The pilot BSS was launched by Provo activists in Amsterdam in the 1960s. It involved giving out white bicycles and leaving them for free communal use haphazardly throughout Amsterdam. As the bikes were not provided with a lock, both this scheme and its analogues (e.g. Cambridge in 1993) eventually were shut down due to theft and vandalism.

At the beginning of 1990s the second generation of coin-deposit bike-sharing systems was launched in Denmark. In this generation docking stations were introduced. There was still no charge for use, but a coin deposit was required to unlock the bicycle. As the system could be used anonymously, it was still theft-prone.

The need to solve this problem gave rise to the third generation of BSSs which remains the most common one nowadays. These BSSs are designed with a program-specific theft-deterrent system: users are no longer anonymous and have to provide their ID, mobile phone number or bank card in order to get a bicycle. These BSSs also employ docking stations. Third-generation BSSs gained more popularity due to incorporating information technology and tracking information about users, trips and stations to improve the service. This generation expanded quickly from 13 BSSs in 2004 to more than 850 in 2014. By the end of 2017 there were more than 1500 functioning BSSs around the world (Meddin, DeMaio, 2017). Modern third-generation BSSs have automatic docking stations, allow users to unlock a bike using magnetic stripe card or smartphone and provide a user-friendly app that shows location and availability of each station.

The fourth generation of dockless bike-sharing systems appeared in China in 2014 and is currently rapidly gaining popularity in Asia. Dockless bike-sharing services have no over-demand issue, are exempt from expenses of maintaining and rebalancing the stations and therefore are cheaper to launch and do not require public subsidy. However, dockless BSSs still have rebalancing expenses and could be restrained by authorities. It seems that it could be difficult for dockless BSS operators to maintain a large bike fleet without causing city regulations. For instance, on 18 January 2018 Dallas City Manager T.C. Broadnax published a letter to bike-sharing operators in response to multiple complaints on bicycles cluttering sidewalks.[[1]](#footnote-1) Since then bike-sharing regulations have been drafted and proposed to the city council, and the fees might be introduced in the near future.[[2]](#footnote-2)

What is more, docked bike-sharing systems are more prone to theft and vandalism, which makes them economically unsustainable. For instance, in February 2018 a Hong Kong dockless bike-sharing start-up Gobee.bike has terminated its service in France, reporting that 60% of their bikes were stolen or vandalized during the first 4 months after they entered the market. Earlier this year Gobee.bike abandoned Milan, Rome, Brussels and other European cities.[[3]](#footnote-3)

The damage to bike fleet in docked BSSs is inevitable, but its level is much lower and it varies dramatically between cities and bike-sharing schemes. In 2012 more than 37% of Vélib bikes were damaged or stolen in Paris with the incidents clustering around low-income districts without video surveillance, while Belfast previously reported that about 15% of their fleet was stolen or vandalized each year on average since their launch in 2015. While CCTV can be employed near bike stations to prevent theft, it is impossible to do so with dockless BSS, which makes them easier to steal.[[4]](#footnote-4) As a result, docked bicycle sharing is expected to keep the lead in Europe in the near future, which means that the problem of rebalancing bike stations remains relevant.

# Related work

Incorporation of information technology in bike-sharing systems by the end of 2000s allowed bike-sharing services to track user and trip information, which drew attention to bike-sharing systems in the academic research. There is a considerable amount of publications aimed at predicting demand for bicycles in bike-sharing systems. An early study by Froehlich, Neumann, Oliver (2009) outlined using digital footprint to understand temporal patterns of movements between different stations using the case of Bicing, the bicycle-sharing system inaugurated in Barcelona in 2007. The authors apply hierarchical clustering to group the stations based on their usage rate and occupancy. The geospatial analysis of the results shows that clustering captures that patterns of use in the city core are different from the outskirts. In this work machine learning techniques, such as Bayesian Network, were successfully applied to make real-time prediction of the number of available bikes and station availability, outperforming historical mean.

This study set off a variety of analytical publications about bike-sharing. The majority of early studies focus on time series models. Bognat et al (2009) model departure rate as using statistical signal processing methods. In a more recent study Bognat, Abry, Flandrin (2011) conducted clusterization similar to an already mentioned work of Froehlich et al., but using a rich dataset published by Vélo’v, bike-sharing service in Lyon, France.

As more different data becomes available online, econometric modeling and especially machine learning methods of predicting demand for bikes gain popularity. The conclusions drawn in the work of Maurer (2011) after a log-linear regression model is trained uncovers that income, job density, car ownership, station capacity, modes of commuting and other factors determine bike-sharing system usage. Rixey (2013) adopts a multivariate linear regression model to investigate how various demographic and built environment characteristics, such as education, income and availability of stations in neighborhoods affected total monthly demand for bicycles. A Faghih-Imani (2014) proposes a station-specific linear mixed model and includes time, weather and land use factors, and Hampshire (2011) among other findings shows that nearby places of interest have a significant impact on station use.

A significant body of literature addresses the problem of over-demand, offering both machine learning and probabilistic methods of predicting whether over-demand will take place. However, these works aim at predicting the fact of over-demand rather than the exact quantity of unobserved demand during this period. Chen et al. (2016) publish an exhaustive study proposing a dynamic cluster-based prediction model of over-demand situations and comparing its performance to baseline time series and machine learning models, such as ARIMA, Bayesian Network and Artificial Neural Network. It shows that predicting the probability of over-demand cases improves classification metrics compared to baseline models. However, the work does not address the problem of predicting quantity of demand.

There is a small number of predominantly theoretical articles introducing probabilistic modeling of arrival and departure rates at the stations. Feng et al. (2016) justify modeling of pick-up and drop-off rates in bike-sharing system as independent Poisson processes. The study of Gast et al. (2015) validates on historical data of bicycle availability in the Vélib bike-sharing system that arrival and departure rates can be precisely represented by time-inhomogeneous Poisson arrival and departure processes. Gast et al. (2015) employ a piecewise-constant form of Poisson intensity rates to simplify the model and conduct an empirical study. This article also is the only one to our knowledge to address explicitly evaluation of rates under censorship constraints, i.e. in situations of over-demand. They modify the formula for expected number of arrivals in the Poisson process to include only the parts of each time interval when the station was not in over-demand state, i.e. when the station was not completely full or empty. The obtained predictor outperformed other point predictors for prediction horizons of 2 to 5 hours.

# Stochastic modelling of bicycle-sharing systems

In this section we discuss a single-station stochastic model of a bike-sharing system that can be used to obtain estimates of departure and arrival rates at given time. While the main concept has already been used by Gast et al. (2015), we provide intuition on how this probabilistic model can be used to get point estimates as well as properties of these estimates. Let us introduce some basic designations and definitions.

* A renewal process is a stochastic process described as:

, ,   
where almost surely.

* A homogeneous Poisson process is a renewal process such that the inter-arrival time between events are distributed exponentially, i.e. for some real has an exponential distribution, i.e. a distribution with a density for . The parameter is called the rate of a homogeneous Poisson process. A time-inhomogeneous process has an intensity function .
* A counting process counts the number of time an event has occurred by the time t and , i.e.

, so

=

for any time interval of size t.

Consider a station currently is in state , meaning that out of a finite number of K slots are occupied by bicycles. Customers arrive at the station to pick up a bicycle according to a time-inhomogeneous Poisson process with a certain intensity rate , moving the station from state to state . At the same time cyclists drop off their bicycles according to another Poisson process with intensity rate , moving the station from state to state . These processes of arrival and departure are time-inhomogeneous Poisson distributed, given that the demand for bikes and empty slots both change with time independently from each other. An oriented graph depicting this model is shown in Figure 1.

Figure 1. Graph of a station of capacity K and time-inhomogeneous occupancy-independent arrival and departure rates

Several assumptions are made to model bicycle stations this way.

* Firstly, processes of arrival and departure are Poisson distributed. The hypothesis that arrival and departure at bicycle stations fit Poisson distribution was validated by Gast et al. (2015) on a dataset from Vélib bike-sharing system. The authors justified this assumption by testing the hypothesis that for each time interval where intensity rates are fixed the distribution of arrivals and departures follow a Poisson distribution, i.e. that

=

if for each for some , where are arrival and departure processes, by measuring K-S test statistic. They found out that stations with many observations have a better fit with the Poisson distribution.

* Secondly, it is assumed that the states of different stations are independent. This assumption could be incorrect for situations of over-demand. When one station is empty, arrival rates at other stations reduce. Later in this section we will propose heuristics for estimating arrival and departure rates under condition of over-demand. However, this effect fades in a dense bike-sharing system where customers have several stations close enough to them to make this ride, and asymptotically stations become independent (Fricker et al., 2012).

To simplify the model but still capture variation of intensity rates over time we will adopt a piecewise constant form of rates at discrete time intervals instead of a continuous function of time, thus simplifying each time-inhomogeneous Poisson process to a sequence of homogeneous processes, which makes it feasible to evaluate model parameters from historical data. The inter-arrival times at each time interval are distributed exponentially.

There are several other limitations to this method:

* While considering a process for each trip from station j to station k with arrival and departure rates specific to each process would make a more realistic queueing model of a network of docking stations as a whole, this complication is not computationally tractable, especially for BSSs with many stations.
* Intensity rates of a station depend on other factors apart from the time. As we have already mentioned, behavior of a certain station is asymptotically independent from other stations (Fricker et al., 2012). However, it definitely depends on weather conditions, season and whether it is a working day or a holiday. Thus, applying this method to real BSSs would require a lot of observations under all the possible scenarios to estimate arrival rates for each station under at least several scenarios for it to be universal.
* For the empirical study of this paper we will narrow down observed period to working days of summer. We will focus on an estimate for good weather without precipitation but neglect temperature variations assuming that changes of weather in summer do not account for change in intensity rates.

# Estimation of arrival and departure rates

For a process of the number of events if events are observed in the interval , then the maximum likelihood estimate of intensity rate at station j is

.

Using observations of station availability throughout the day we can divide the time span of the day into intervals with homogeneous Poisson processes of arrival and departure on each of them and estimate intensity rates for each time interval for each day and then take their average. This estimate is unbiased:

.

The variance is . Then by averaging estimates for many realizations we will get more accurate results.

There are multiple ways to obtain a confidence interval for resulting estimated intensity rates. Patil, Kulkarni (2012) compare different methods and recommend different methods of computing a confidence interval depending on how big the estimated intensity is, i.e. how frequently events occur (how many times arrival or departure of cyclists is observed during an hour in our cases). For a high intensity like ours a recommended formula of a confidence interval for is

( where is -quantile for Pearson distribution with v degrees of freedom.

# Estimation of arrival and departure rates during over-demand periods

To estimate arrival and departure rates in peak hours for a particular station we need to keep in mind that at peak hours one of the rates is unobserved, so we need to adjust our formula to include only part of time interval when the station was not completely full or empty for each day.

,

,

where D is a set of days used for estimation and denominator counts the number of observations in given interval for which occupancy was not equal to maximum occupancy or to 0 for departure rate and arrival rate, respectively.

After the section where we describe the data and discuss limitations and issues connected with rebalancing that is already performed but not markered in the data we will use the formulas introduced above to estimate intensity rates of a station.

# Data description

In this paper performance analysis of the implemented model will be realised on a simulated station. However, investigating real data is important to understand patterns of use and justify importance of predicting actual demand.

The real-time data of dublinbikes, the bike-sharing system of Dublin, Ireland is published by JCDecaux, the French advertisement company that operates dublinbikes as well as many other bike-sharing systems in the cities of Europe.

With more than 22 million trips of an average duration of 14 minutes made since its launch in 2009, dublinbikes is considered one of the most successful cases of integrating a BSS with public transport system. The number of people cycling to work rose from 3.8% in 2006 to 10.3% in 2017[[5]](#footnote-5) and cycling accounts for 12.9% of total vehicular traffic in central Dublin in 2017, while in 2009 it accounted for less than 5% of traffic. Despite this significant rise, road deaths to cyclists have not increased, which means that cycling has become safer as well.[[6]](#footnote-6)

The data covers the period from 24 January 2017 to 14 August 2017 and consists of the number of bikes at each of the 102 stations[[7]](#footnote-7) observed every 2 minutes[[8]](#footnote-8). We then narrow down the period to May and summer months and select only working days. We drop all the days for which there are at least 20 missing values for one station. We handle other missing values by substituting them with means of neighbouring values.

As dublinbikes, like most of public bike-sharing systems, does not publish trip-specific data, there is a possibility that part of the arrival and departure traffic is not captured in the data as it cancels out between the 2-minute intervals when observations are made. Using trip-specific data could improve the accuracy of results.

# Introduction of heuristic for detecting rebalancing acts

We believe that dublinbikes already does rebalancing of some of the stations throughout the day, but as journey data is not available and this rebalancing is changing the occupancy of the station we will have to manually detect and eliminate at least some of these cases. It should be pointed out that this is a heuristic that is just aimed at making data closer to reality and it would not be necessary if dublinbikes published trip information which it probably tracks. However, the presence of these rebalancing peaks is quite useful for arrival and departure rates estimation. Combined with controlling for the origin of the event, i.e. was it a trip made by a customer, or station occupancy rose because of a rebalancing act, rebalancing popular stations during their peak hours might be used to research the amount of unobserved demand in order to determine whether the capacity of the station should be expanded or a new station should be opened near the one that is being studied.

Let us call a pattern in consequent observations of a station a peak if it starts with an abrupt change in station occupancy in one direction and then is followed by a section of a graph with a prominent trend in opposite direction. Let us call a peak upward if it leads to an abrupt rise of occupancy of the station, or downward if it leads to an abrupt fall of occupancy. A peak is high if it results in a significant change of occupancy, and low otherwise.

We will now provide several statements and assumptions that will be used to justify decision on whether or not the peak should be considered a rebalancing act. A small case study of Portobello stations aimed at justifying these assumptions is provided afterwards.

Statement 1. Each station that is prone to over-demand has a certain pattern of use that is usually shaped by direction of traffic of cyclists going to work.

To prove it we conducted K-means clusterization of all the stations in the BSS and chose the appropriate number of clusters according to the graphs of within cluster sum of square errors using elbow method, which is shown in Figure 2.

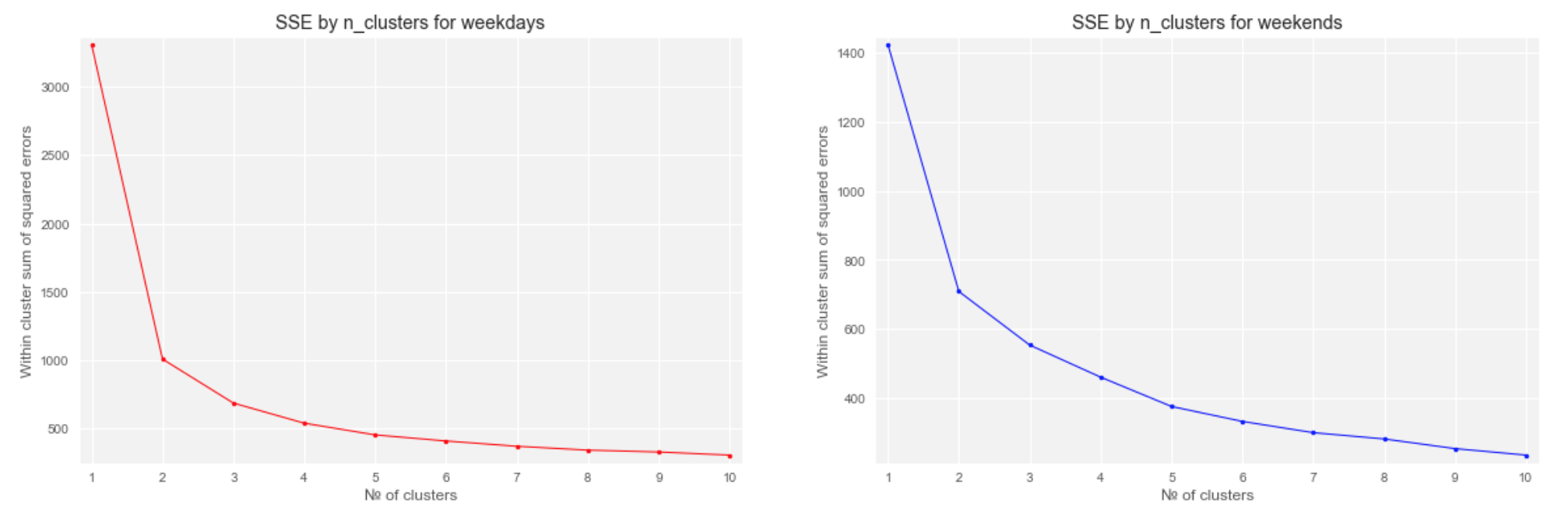


Figure 2. Dependence of the sum of square errors on # of clusters for weekdays and weekends

Figure 3 illustrates that the two peaks of bicycle use at weekdays are during the start and the end of working hours whereas peak use at weekends is shifted towards lunchtime.



Figure 3. Clusters of the patterns of bicycle usage at weekdays and weekends

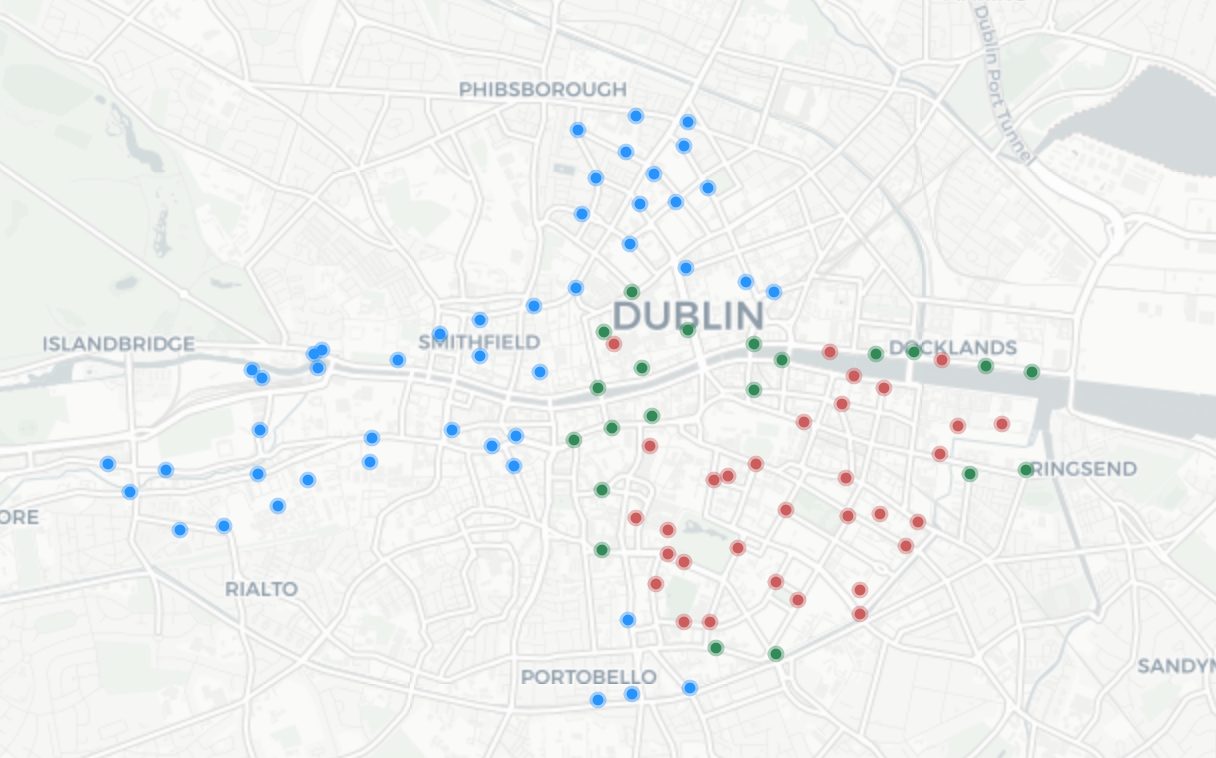


Figure 4. Dublin BSS with stations colored depending on their usage pattern cluster

Assumption 1. A peak with a change of at least 7 bicycles that lasts 2-6 minutes and occurs during the time when this station usually starts getting to an over-demand state but moves in the opposite direction is a result of a rebalancing act.

To show that big peaks occur due to rebalancing and not for other reasons let us consider behavior of occupancy of the Portobello stations. There are four stations in Portobello area of Dublin. Three of the most southern stations in Dublin are situated in Portobello, Portobello road is the most southern one.

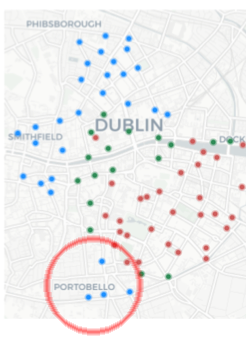
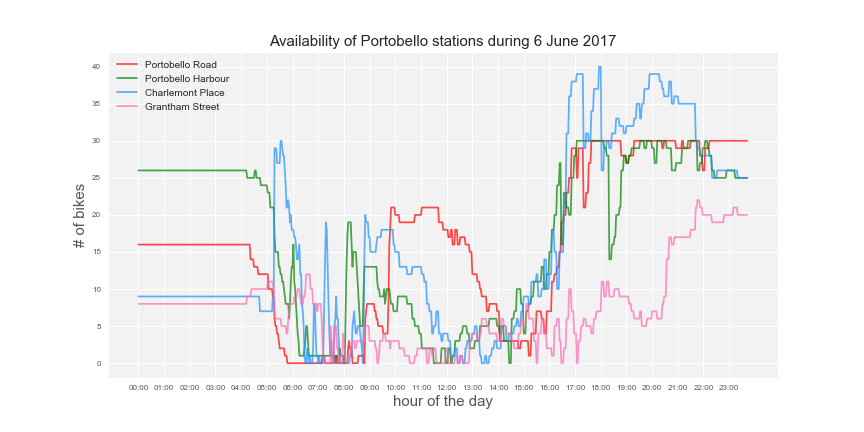


Figure 5. Portobello stations location & occupancy on 6 June 2017

The pattern of use of these stations on weekday mornings is typical for stations which are used as a starting point to commute to work in the center of Dublin. Consider Portobello road station. From approximately 6 in the morning its occupancy starts to fall rapidly. Since working hours in Dublin generally start at 9 in the morning, it is expected that from 8 to 9 in the morning departure rate will be much bigger than arrival rate, causing the station to end in an over-demand state at this time. However, as it can be seen in a plot of daily occupancy of Portobello road in June (Figure 2), single high peaks can be observed at this time.

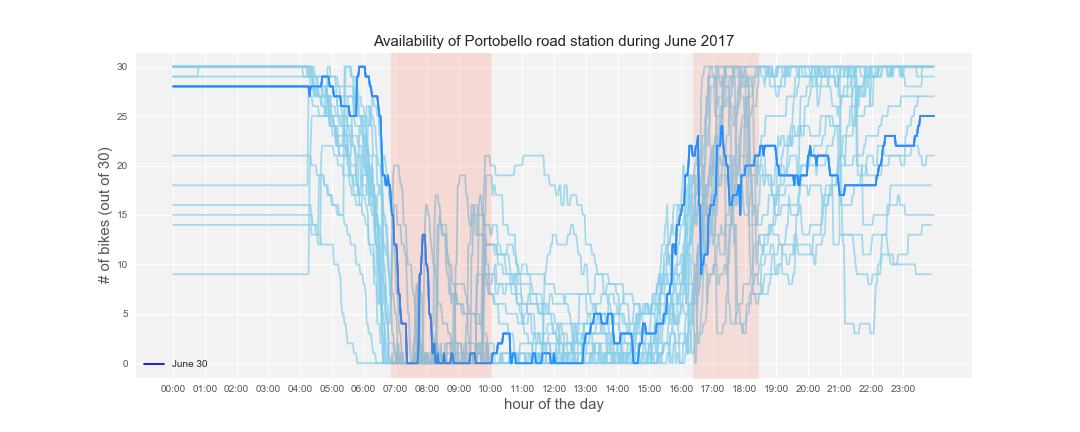


Figure 6. Graph of availability of Portobello road station during June 2017

There are several possible reasons for a high upward peak:

* If it accompanied by the change of properties of Poisson processes that generated this sequence of arrivals and departures of bikes, then it is due to either a rebalancing act, or a sudden temporal rise of demand for slots. While a high downward peak could occur due to arrival of another mode of public transportation (e.g. train) or due to the end of working hours, the only probable reason for a high upward peak would be arrival of a group of tourists or locals that were on a bicycle tour. However, as it can be seen in the graph depicting behavior of Portobello road during June these peaks occurred almost every day at different time, which is highly improbable for a scheduled tour. In addition, Portobello road is not a place of interest for tourists.
* If a peak appeared by chance as a possible order of arrivals and departures in the same homogeneous Poisson processes as the sequences of arrivals and departures to the left or right of it, i.e. in some wide time interval with constant rates, then formal approach would require testing likelihood of such order of arrivals and departures under given distribution and would result in a probability close to zero.

There are several reasons for a high downward peak:

* A sudden temporal rise of demand for bikes. This could occur due to arrival of another mode of public transportation (e.g. train). However, our research of train and bus stations in close proximity to Portobello road bicycle station showed that the closest ones are situated in the eastern part of Portobello area closer to Charlemont Place station. If a vehicle full of people demanding for bicycles arrived, they would all head to the closest station which is Charlemont Place, and if it didn’t have enough bikes, customers who didn’t get a bike would go to the next closest station, and so on. However, as it can be seen in the graph depicting behavior of all Portobello stations on the 6th of June, this was not the case. Another reason for a sudden temporal rise of demand for bikes could be end of working hours. Suppose, at the end of the working day in a particular office A all the workers leave their workplace simultaneously. Then on different working days it would generate spikes in approximately the same time. As we can see from the graph showing Portobello road behavior in June, the spikes occured on an at least hour-long time interval, which would not be the case if it were for end of work in a particular place.
* Again, probability that such a high peak appeared by chance in the same homogeneous Poisson processes would result be close to zero.

As we can conclude from this mostly informal case study, the only probable possibility for high peaks at Portobello stations would be a rebalancing act. Proposed heuristic does not allow us to detect correctly all cases of rebalancing, especially if we chose a low threshold for a peak height. However, this solution allows us to still improve the accuracy of our dataset.

# Realization of peak detection heuristic

To detect rebalancing peaks defined earlier we used rolling mean with a window of 3 observations of first differences. After peaks of a length of 7+ bikes over 2-6 minutes are selected, we check that after this peak occupancy dynamic will change its direction. This ensures that surges and falls of occupancy due to actually high difference between arrival and departure rates are not considered a peak. The algorithm realized in Python is presented in the Appendix. The graphs below show an example of peak detection at Portobello road station on 13 June 2017.

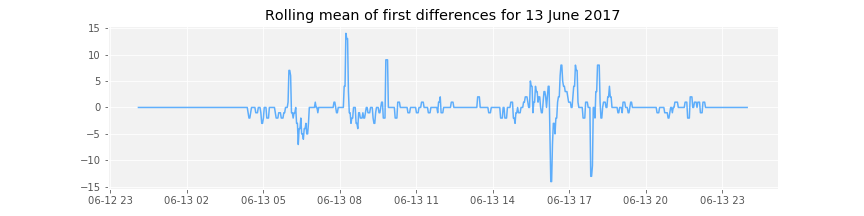


Figure 7. Rolling mean of first differences for 13 June 2013. Portobello road station

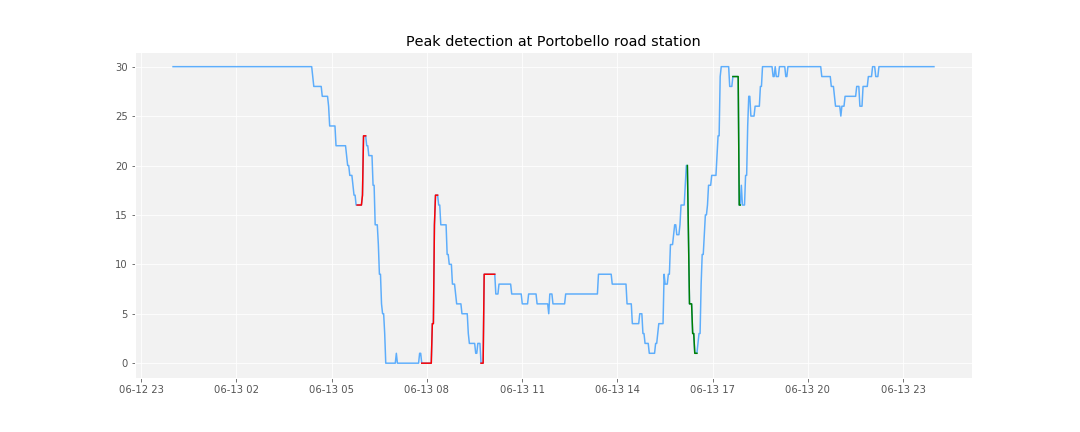


Figure 8. Peak detection for 13 June 2013. Portobello road station

Now that we detected which bicycles appeared at the station because of rebalancing we can make sure that occupancy changes due to rebalancing are not counted when estimating arrival and departure rates.

It should be mentioned that after dropping rebalancing acts Portobello station ends up being over-demand during the working hours. While Portobello road is a busy station away from city center that is regularly rebalanced, many of the stations in the center are not rebalanced through the day. Over-demand periods occuring during working hours can be observed at most of the stations. The graph below aggregates by week the number of stations out of 102 that were full or empty for at least 1.5 hours a day. hours. During the observed period from January to August about 60% of dublinbikes stations remained colmpletely full for at least 1.5 hours a day.

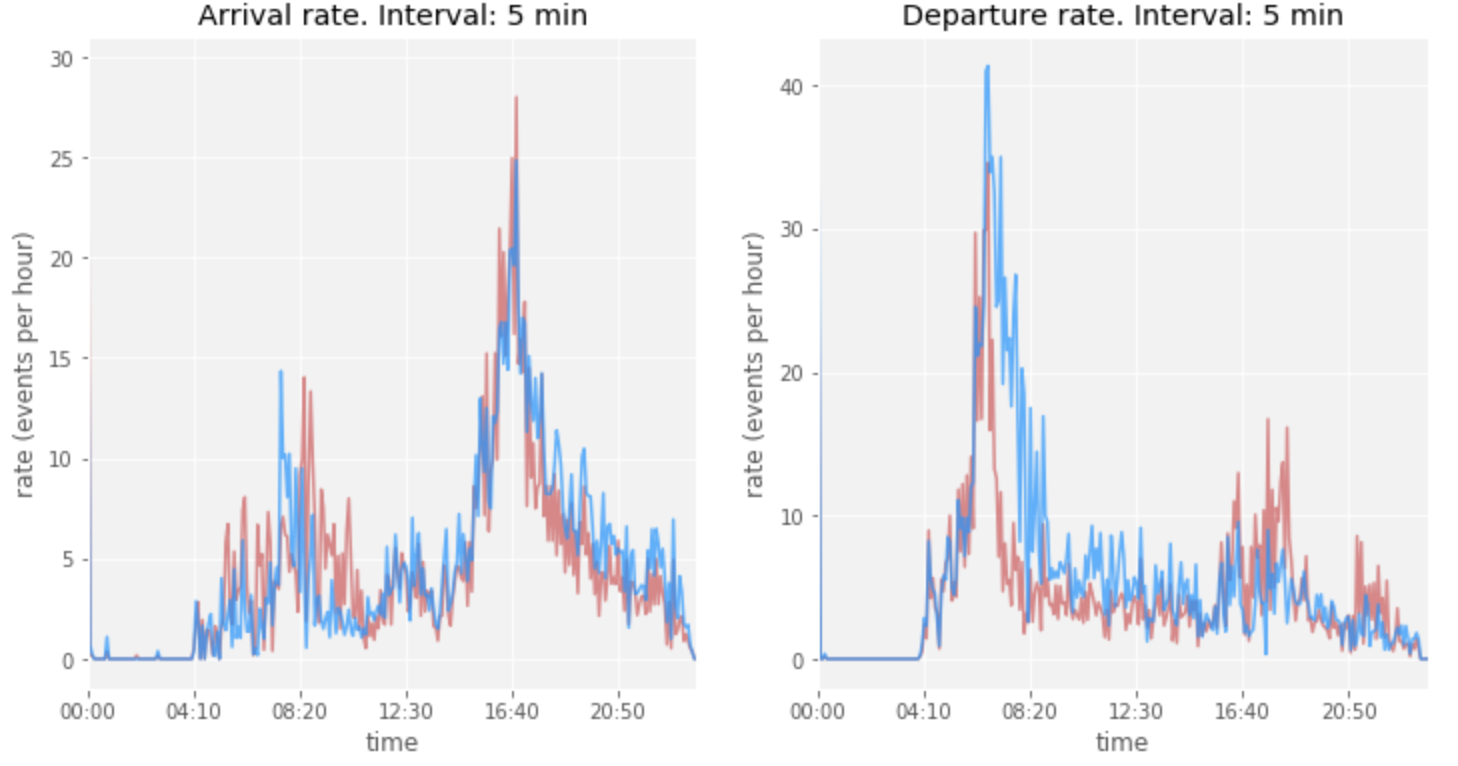
# Estimation of arrival and departure rates

We used step sizes of 5, 15, 30 and 60 minutes to estimate arrival and departure rates using previously introduced formulas:

,

,

Figure 9 portrays comparison of a basic maximum likelihood estimates (in red) and adjusted estimates (in blue) of intensity rates of Portobello road station arrival and departure processes. As expected, adjusted formula captures intensity rates better both at peak and moderate values:



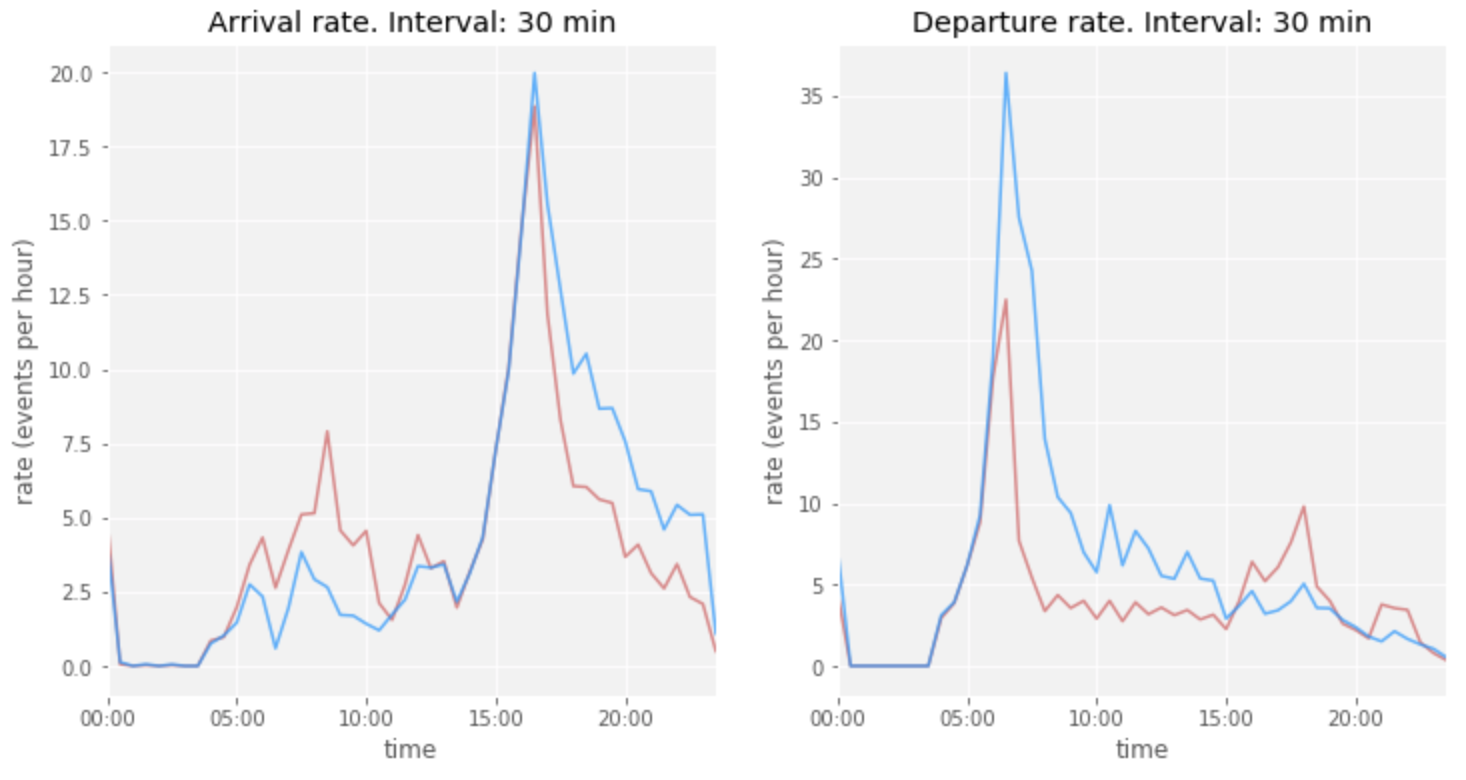


Figure 9. Estimated intensity rates with different steps for Portobello station.

During working hours when Portobello station is usually close to being empty adjusted formula captures a higher departure rate. What is more, basic estimate did not fully capture the fact that the quantity of demand for bikes in the morning is higher than the quantity of demand for empty docks in the evening: for instance, estimated with 30-minute steps adjusted departure rate peaks at 1.82 times the maximum height of adjusted arrival rate.

The step size of 30 minutes allows us to keep important variations of the data but

not overspecify the rates. A full set of graphs of estimated intensity rates with different steps can be found in Appendix 1.

# Simulation

To estimate performance of the model on the intervals of unobserved demand we conduct a simulation of an isolated station.

|  |
| --- |
| **Table 1. Algorithm of getting one realization of an arrival process** |
| select piecewise-constant arrival rate with steps of the same length  for a time interval of length of a step do:  ....generate # of arrivals in this interval as a Poisson value with an according intensity rate  ....for each arrival in generated # of arrivals in that interval:  ........select arrival time uniformly in that interval |

Assumptions:

* Arrival and departure rates of a station are Poisson processes. As it was already stated, the fact that Poisson process is an adequate model for station occupancy was shown by Gast et al. (2015).
* All observations are made under similar weather conditions and on weekdays. It means that for the full modelling of a real station we would need to estimate rates in several scenarios and alternate between estimated rates for predictions in different weather.
* If the station is in over-demand, customer do not wait at the station. This assumption is very plausible, because there is no scheduled rebalancing in Dublin. What is more, users have access to real-time information about the number of bikes at each station published in the app and on the website.

The algorithm of obtaining simulations is described in Table 1. It uses the fact that inter-arrival times of a Poisson process are exponentially distributed, and exponential distribution has the memory-less property. Because of that we can model arrival times as uniformly distributed within each step with a fixed rate.

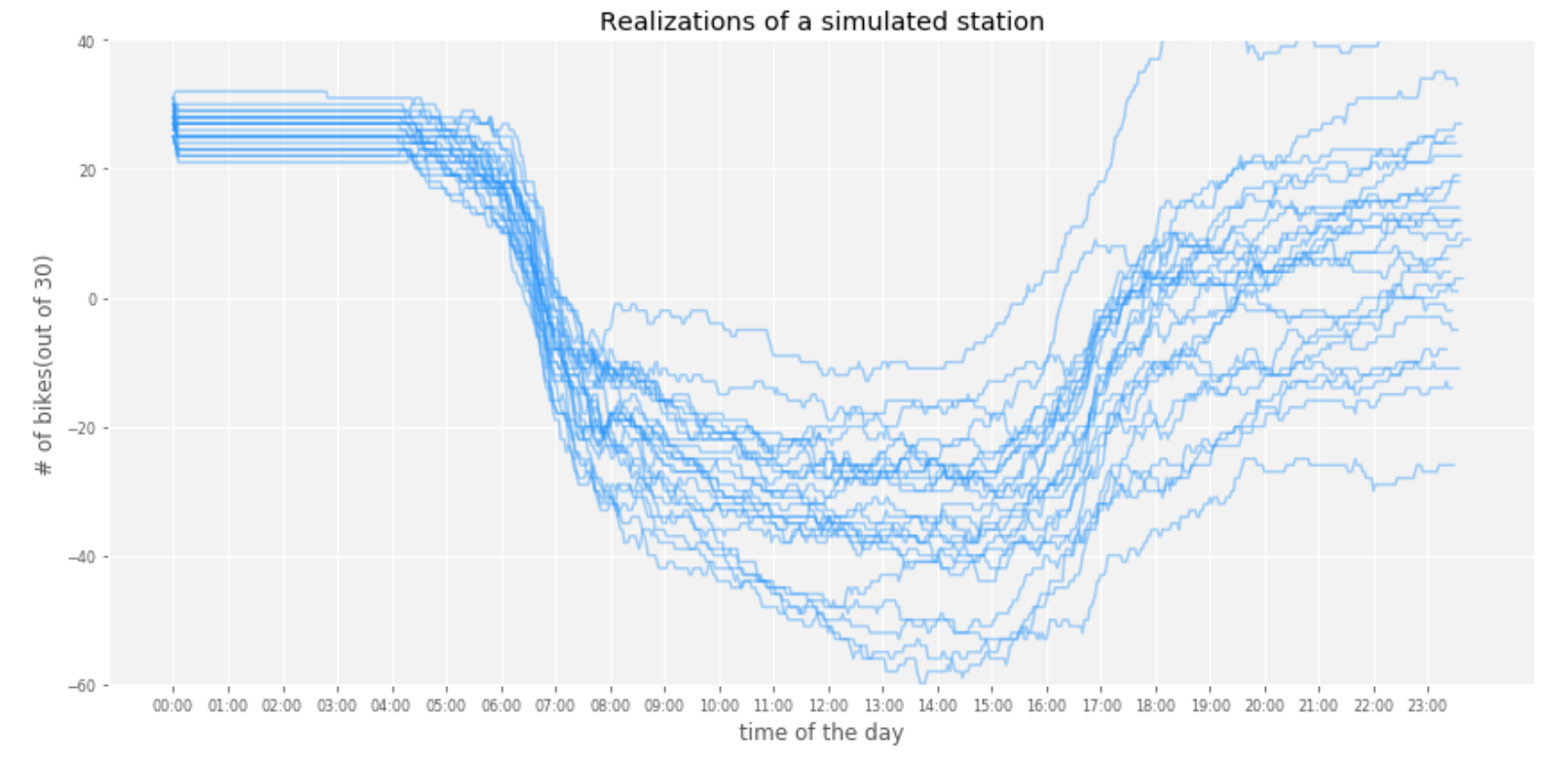
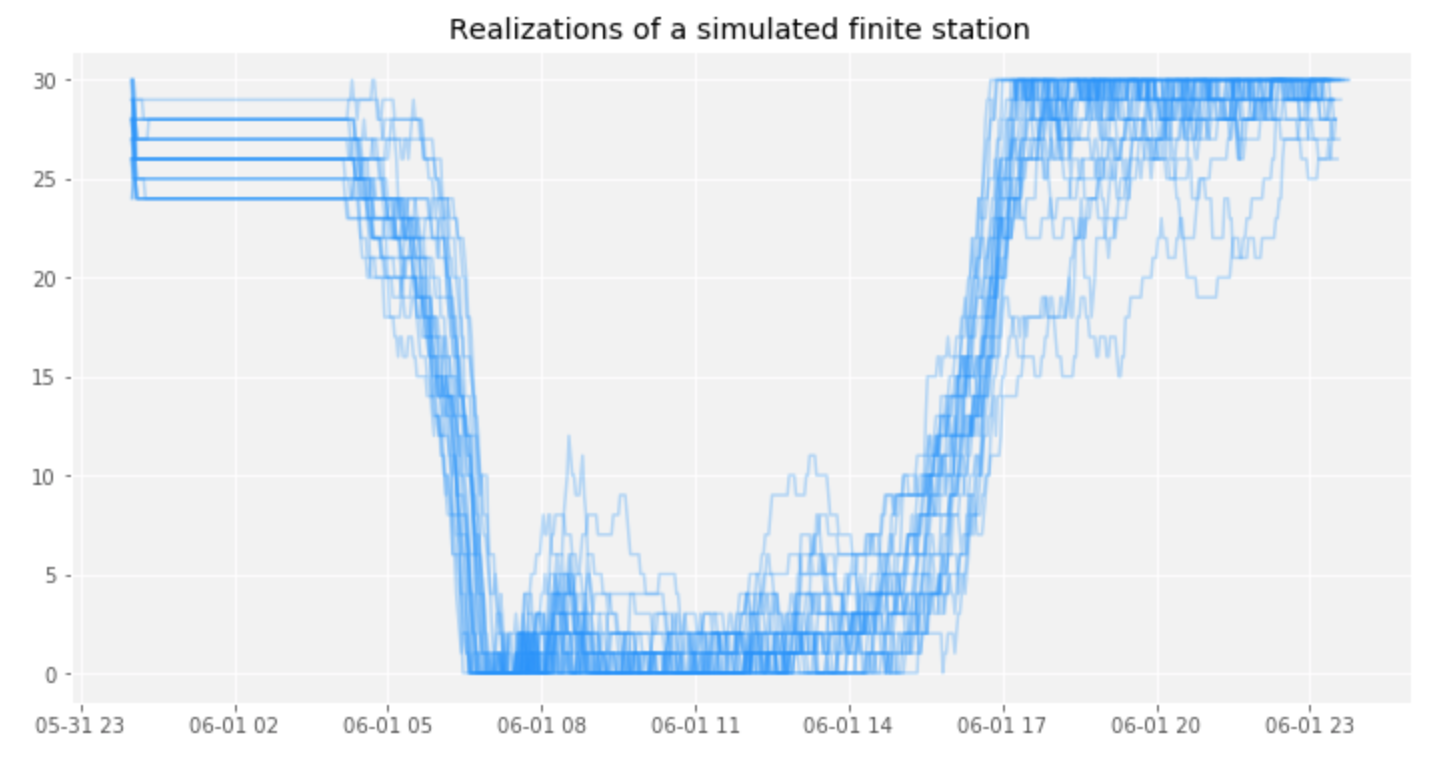
****

Figure 10. Realizations of a simulated station with no capacity limit for 29 days.

To simulate working with observed data arrival rate we will now neglect modelled times of departure if at the time there are no bicycles at the station.

****Figure 11. Realizations of a simulated station with capacity limit for 29 days.

When testing predictive power of resulting models, several different step sizes of a piece-wise function were considered. After estimating the results on up to 3000 simulations the 15-minute step yielded lower mean squared error when predicting observed and actual demand both on small and large datasets.

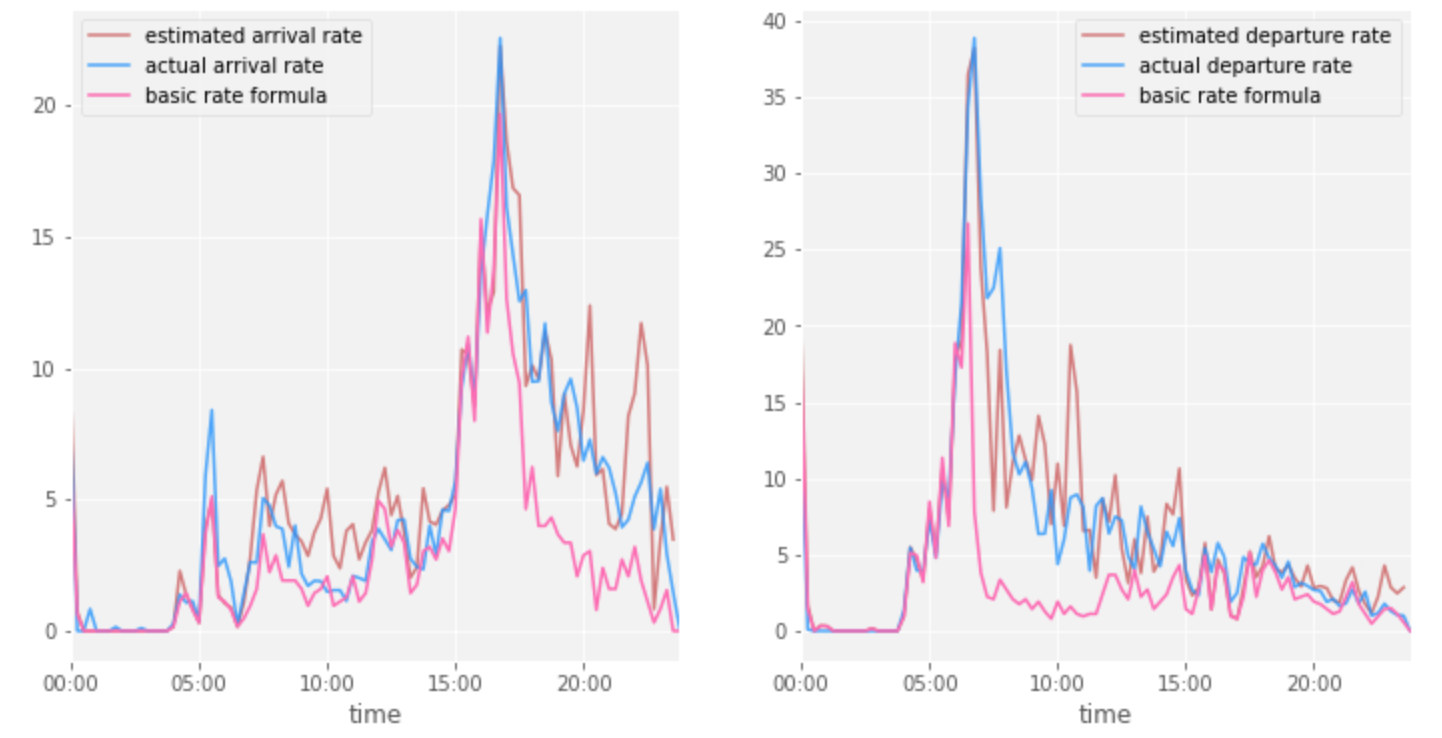


Figure 12. Intensity rates of a simulated stations for 29 days.

Figure 12 compares actual rates and both basic and adjusted estimates made on 29 days. It is clear that the basic estimate undervalues peaking rates after over-demand state is reached, while adjusted rate seems to overestimate it due to the fact that only a small set of data was used for these estimations. Overall fit of the adjusted rates has a smaller error.

It is important to mention that we focused on testing stations that are in high demand at some time interval, so the intensity rates were that chose lead to over-demand cases. Several different shapes of arrival and departure processes were considered to control for possible differences in simulation results. The one that is demonstrated in this section is intentionally taken from Portobello road intensity rate estimation in previous section.

To test predicting power of the model and compare it to other popular predicting algorithms, 30-minute, 1-hour and 2-hour time horizons were chosen. As the main goal of the model is to provide recommendations for operators about the number of bikes to relocate to or from the station, a time horizon should leave enough time for the relocation to be made. Then random points in time were selected as time of prediction and both actual values and predictions for a simulated station in 30 minutes, 1 hour or 2 hours from those points, respectively, were used to calculate root mean square error:

For comparison we chose ridge and lasso linear regressions and random forest algorithm. Alpha parameter for Ridge and Lasso regressions was tuned to choose among several values. We applied cross-validation to get stable results. It is expected that random forest yields the best results as it is a strong ensemble algorithm that is resistant to overfitting and including useless variables. One set of target values was set equal to station occupancy in one hour from the moment of prediction, another set of target values reflected station occupancy in two hours. The following features were selected for these models: # of bicycles at the moment of prediction, # of bicycles 1 hour ago, # of bicycles yesterday, minute and hour at the moment of prediction. This set of variables makes it comparable to our model in the way that it is trained with assumption that weather conditions are the same for the whole period of prediction. It allows us to compare predictive power of these models with proposed model. The results are shown in Table 2:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2. RMSE for predictions on simulated data** | | | |
| Time horizon | 30 minutes | 1 hour | 2 hours |
| Proposed model | **2,96** | 3,53 | 5,99 |
| RF | 3,44 | **3,46** | **4,77** |
| Lasso | 6,54 | 6,63 | 10,95 |
| Ridge | 6,54 | 6,74 | 11,05 |

Proposed model outperformed Random Forest on a time horizon of 30 minutes, but Random Forest showed better results for predictions on horizons of 1 and 2 hours.

These results applies to the case of similar weather conditions, whereas in reality occupancy depends on weather conditions, as it was found by authors of some articles listed in the literature review. Nevertheless, we believe that using weather features can improve our model as well: having an extensive dataset and weather information, one can manually divide the data into parts with similar weather conditions to estimate different intensity rates on a number of scenarios and then alternate between estimated rate functions when predicting occupancy depending on what the weather forecast is.

The model we used, however, cannot be substituted by Random Forest for all purposes, because our model makes predicting unobserved demand possible, which cannot be achieved training a model using only observed data as target variables.

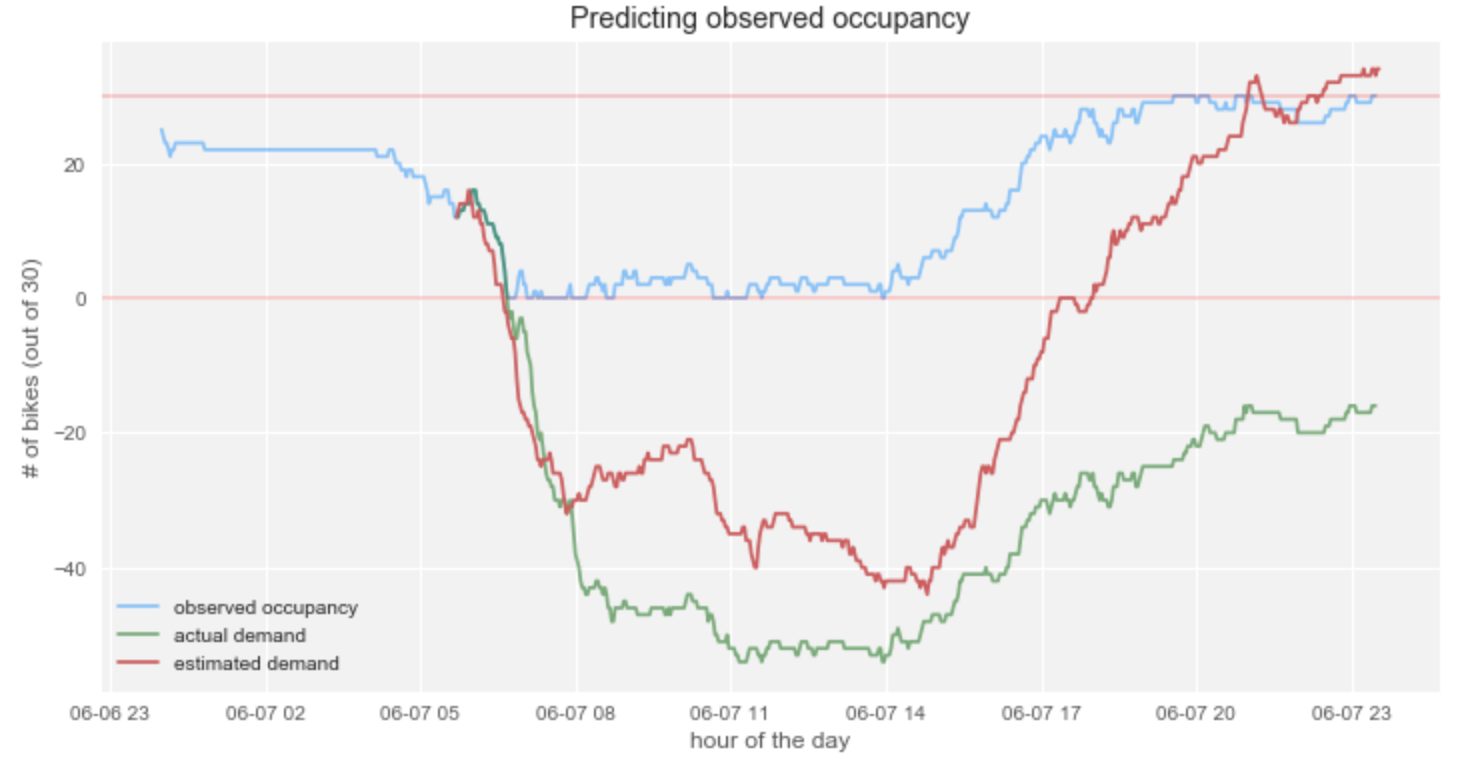


Figure 13. An example of modelling demand using proposed model.

# Conclusion

The results obtained in this study show how modelling station occupancy as a combination of independent Poisson processes of arrival and departure can be used to predict actual demand for bicycles despite the fact that it cannot be observed in over-demand states. We show that estimating intensity rates on historical data as a piece-wise constant function using only points in time when both of the rates are observed helps avoid underestimating the magnitude of these rates. We estimate these parameters for a simulated bibycle station and use them to predict future values of actual demand at the station. Performance analysis of this model in comparison to results yielded by several popular machine learning algorithms show that proposed model outperforms them on a time horizon of 30 minutes, but does not beat Random Forest algorithm on time horizons of 1 hour and 2 hours in terms of RMSE.

The model we introduced is different from aforementioned algorithms in that it makes prediction of actual demand possible. Knowing actual demand could be insightful for bicycle-sharing operators making decisions about expansion of their system: comparing the stations’ average quantity of unsatisfied demand shows the area where new stations should be opened or the number of docks should be changed in the first place. Forecast of actual demand can improve the accuracy of input into systems that find optimal route of bicycle relocation and thus efficiency of its work.

The direction for further research may lie in building a closer to real conditions model that accounts for weather conditions. It can be achieved by manually dividing the data into parts with similar weather conditions and with or without precipitations to estimate different intensity rates on a number of scenarios and then alternate between estimated rate functions when predicting occupancy depending on what the weather forecast is. Performance of this model can then be compared to algorithms that use weather variables apart from the ones that were used in this paper, further investigating its potential for implementation in real-life bicycle-sharing systems.

# References

Borgnat, P., Abry, P., Flandrin, P. and Rouquier, J.B., 2009, September. Studying Lyon's Vélo'v: a statistical cyclic model. In ECCS'09. Complex System Society.

Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.B. and Fleury, E., 2011. Shared bicycles in a city: A signal processing and data analysis perspective. Advances in Complex Systems, 14(03), pp.415-438.

Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., Wu, Z., Pan, G., Nguyen, T.M.T. and Jakubowicz, J., 2016, September. Dynamic cluster-based over-demand prediction in bike sharing systems. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 841-852). ACM.

Chiariotti, F., Pielli, C., Zanella, A. and Zorzi, M., 2018. A dynamic approach to rebalancing bike-sharing systems. *Sensors*, *18*(2), p.512.

Faghih-Imani, A., Eluru, N., El-Geneidy, A.M., Rabbat, M. and Haq, U., 2014. How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. Journal of Transport Geography, 41, pp.306-314.

Feng, C., Hillston, J. and Reijsbergen, D., 2016, August. Moment-based probabilistic prediction of bike availability for bike-sharing systems. In International Conference on Quantitative Evaluation of Systems (pp. 139-155). Springer, Cham.

C. Fricker, N. Gast, and A. Mohamed. Mean field analysis for inhomogeneous bike sharing systems. In Aofa 2012, International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, 2012.

Froehlich, J., Neumann, J. and Oliver, N., 2009, July. Sensing and predicting the pulse of the city through shared bicycling. In IJCAI (Vol. 9, pp. 1420-1426).

Gallager, R.G., 2013. *Stochastic processes: theory for applications*. Cambridge University Press. <http://www.rle.mit.edu/rgallager/documents/6.262lateweb2.pdf>

Gast, N., Massonnet, G., Reijsbergen, D. and Tribastone, M., 2015, October. Probabilistic forecasts of bike-sharing systems for journey planning. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 703-712). ACM.

Hampshire, R.C. and Marla, L., 2012, January. An analysis of bike sharing usage: Explaining trip generation and attraction from observed demand. In 91st Annual meeting of the transportation research board, Washington, DC (pp. 12-2099).

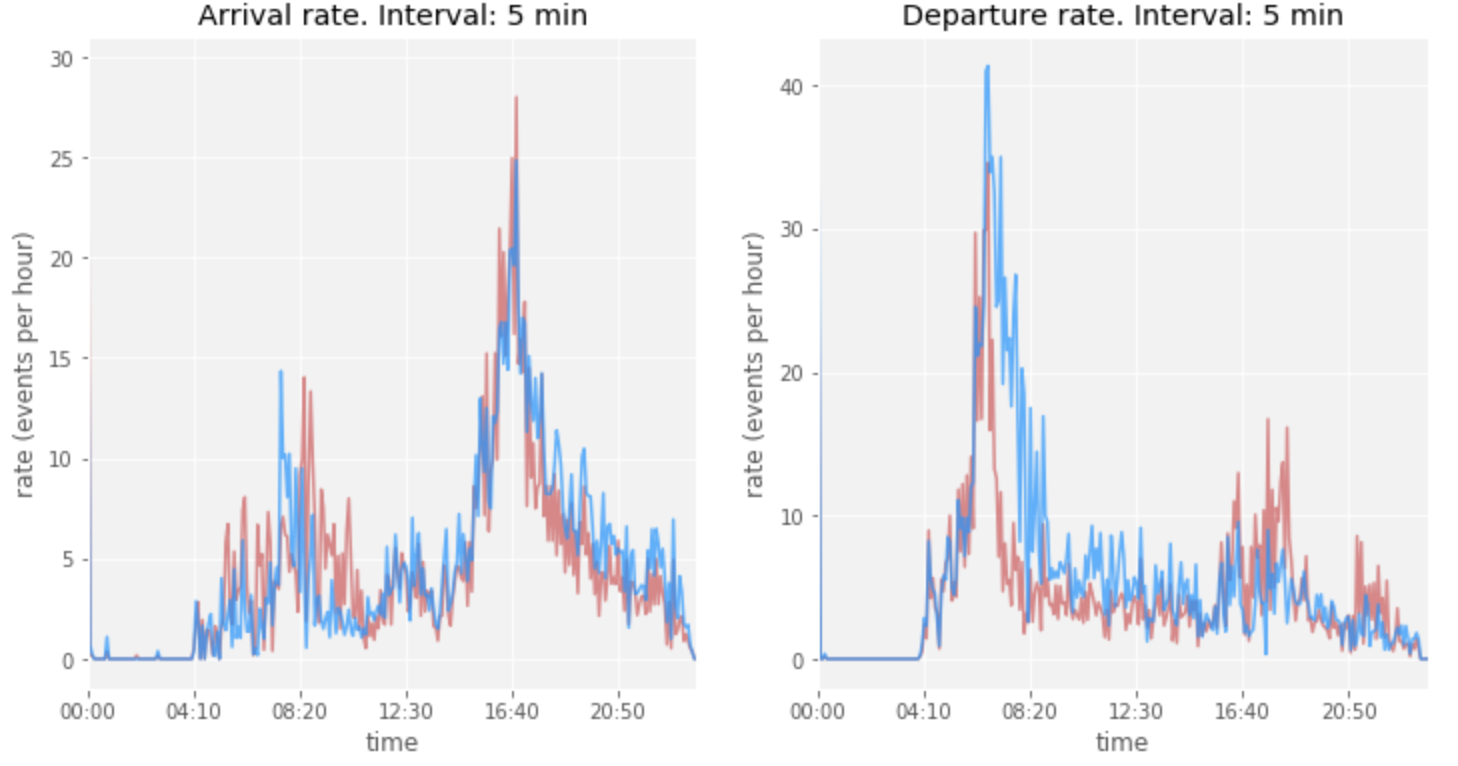
Patil, V.V. and Kulkarni, H.V., 2012. Comparison of confidence intervals for the Poisson mean: some new aspects. *REVSTAT–Statistical Journal*, *10*(2), pp.211-227.

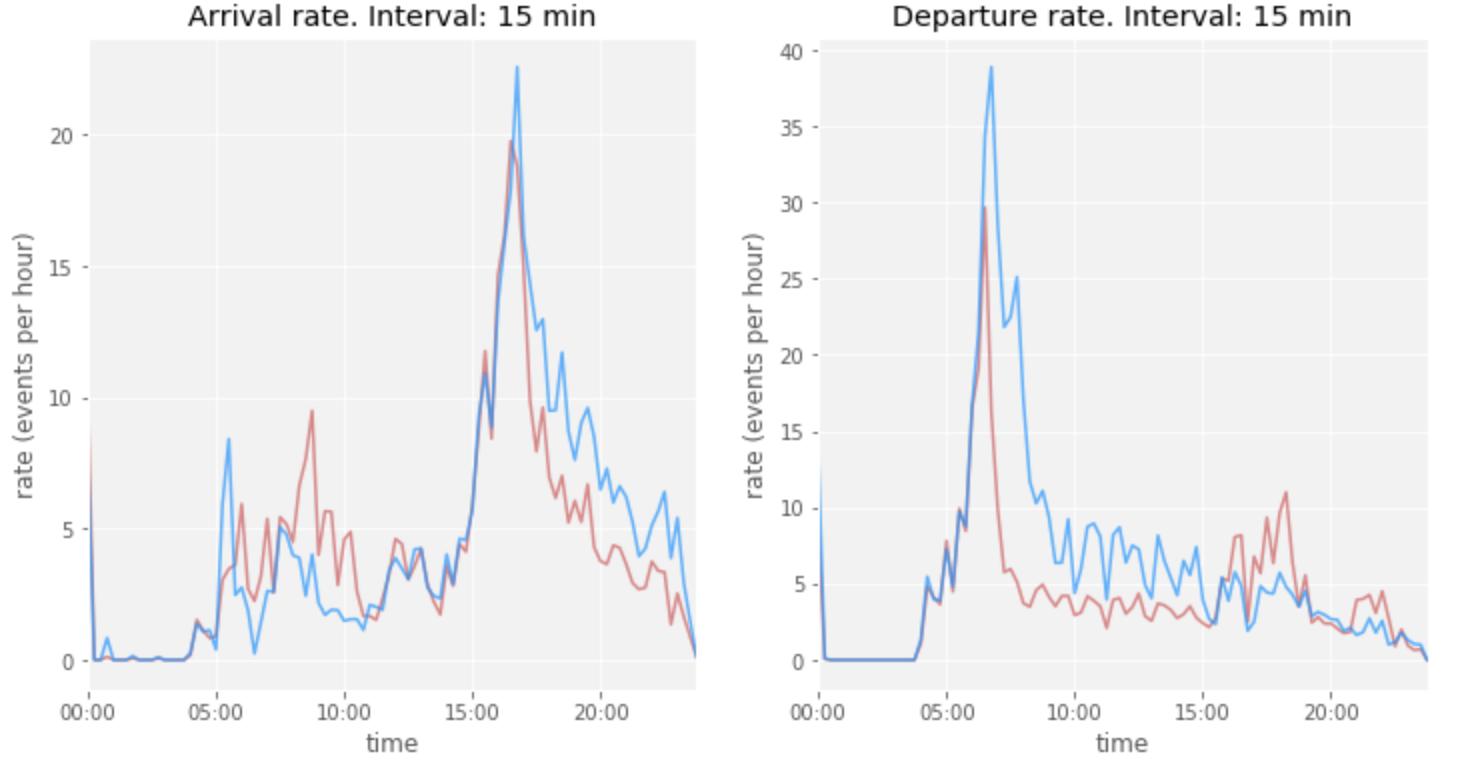
Rixey, R., 2013. Station-level forecasting of bikesharing ridership: Station Network Effects in Three US Systems. Transportation Research Record: Journal of the Transportation Research Board, (2387), pp.46-55.

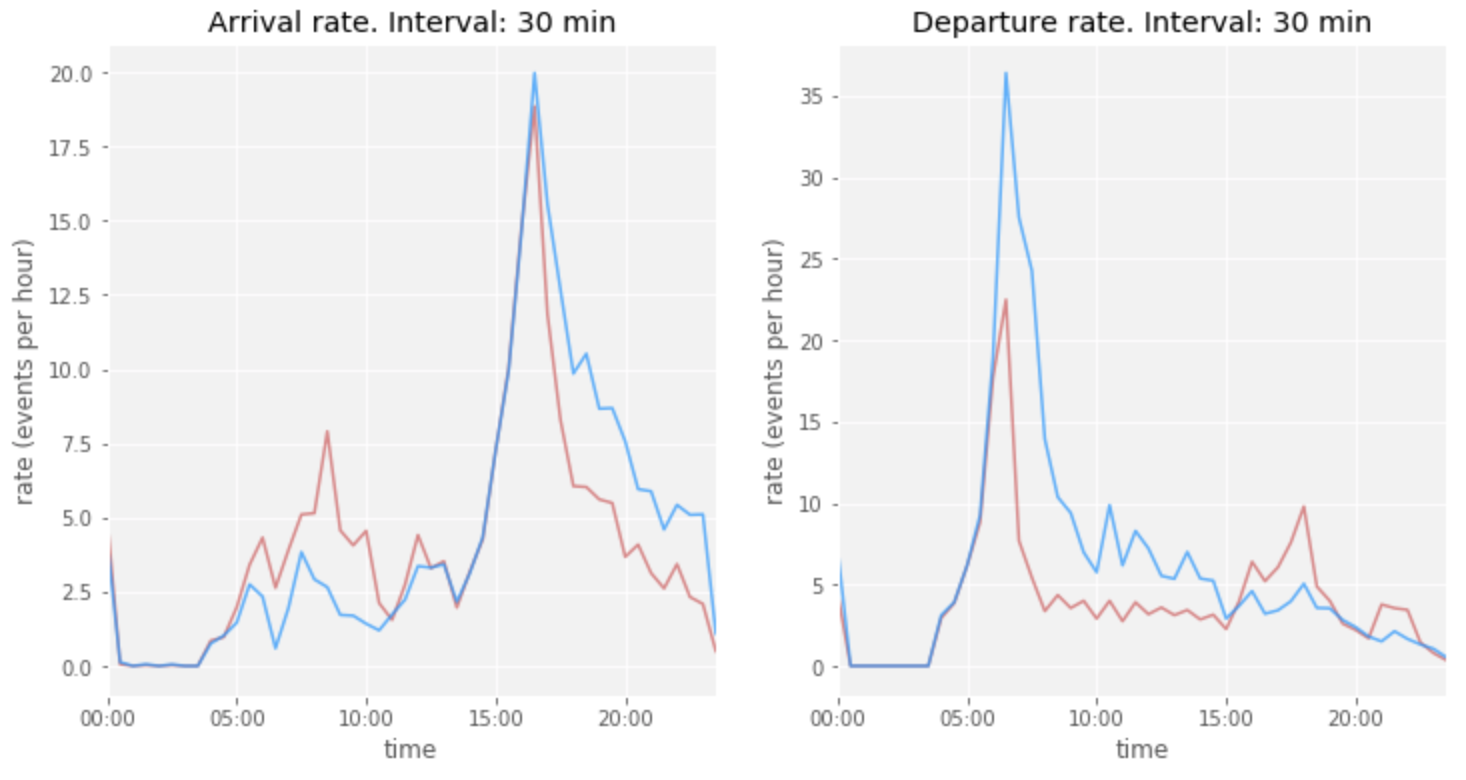
Schuijbroek, J., Hampshire, R.C. and Van Hoeve, W.J., 2017. Inventory rebalancing and vehicle routing in bike sharing systems. European Journal of Operational Research, 257(3), pp.992-1004.

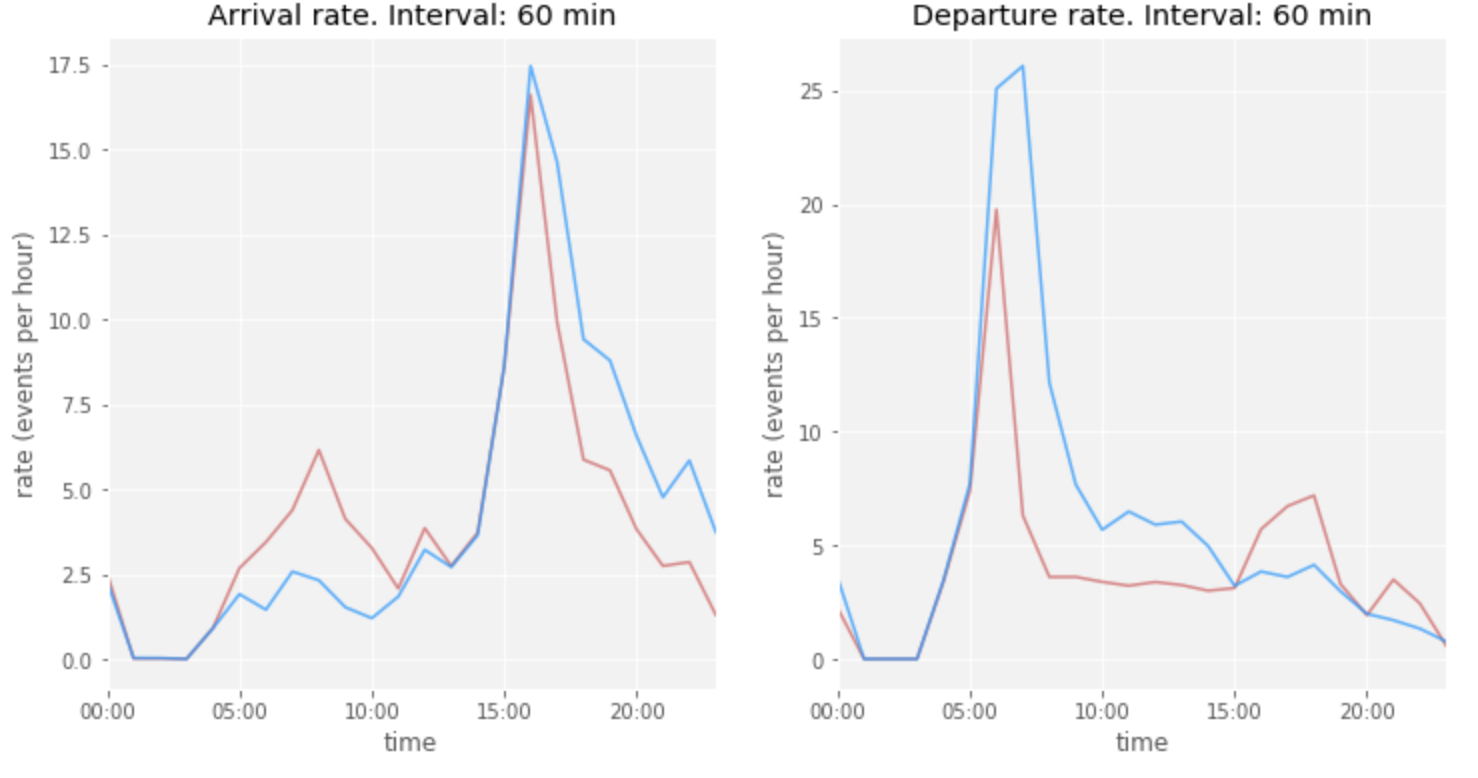
Appendix **1.** Estimated intensity rates with different steps.

Portobello road station









1. https://www.dallasnews.com/news/dallas-city-council/2018/03/21/dallas-will-soon-stop-wild-west-bike-sharing [↑](#footnote-ref-1)
2. https://www.dallasnews.com/news/dallas-city-council/2018/04/09/city-council-worries-proposed-permit-fees-will-chase-bike-share-companies-dallas [↑](#footnote-ref-2)
3. https://www.theguardian.com/world/2018/feb/25/gobeebike-france-mass-destruction-dockless-bikes [↑](#footnote-ref-3)
4. <https://www.citylab.com/transportation/2013/09/paris-thefts-and-vandalism-could-force-bike-share-shrink/7014/> https://www.whatdotheyknow.com/request/stolen\_barclays\_cycle\_hire\_cycle#incoming-393612 [↑](#footnote-ref-4)
5. https://www.bmj.com/content/360/bmj.k94/rr-2 [↑](#footnote-ref-5)
6. http://irishcycle.com/2018/01/22/pedestrians-and-cyclists-nearly-50-of-traffic-in-dublin-city-centre-counts/ [↑](#footnote-ref-6)
7. stations that were launched later than 24 January 2017 and 3 stations with more than 20 missing values were dropped [↑](#footnote-ref-7)
8. The data published through API is free for use; collection and storage of the data used in the study is courtesy of James Lawlor ([jalawlor@tcd.ie)](mailto:jalawlor@tcd.ie)) and is used with his permission [↑](#footnote-ref-8)