NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

DEPARTENT OF ECONOMICS

PROJECT PROPOSAL

**Modeling Station-Level Demand in Bike-Sharing Systems**

Maria Golovina, BEC 145

Academic supervisor:

Stankevich Ivan Pavlovich

Moscow

2018

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. This study will focus on predicting station-level demand for bicycles and docks using both data mining techniques and stochastic modelling. We model spatiotemporal arrival and departure rates and suggest a way of estimating these rates under conditions of censoring caused by finite station capacity based on available historical data. We then use estimated rates to calculate the number of bicycles at each station for each day in the training set as if there were no limitations of station capacity. The new target variable that reflects actual demand for bikes as well as slots at each station is used to train a multivariate actual demand forecasting model. Historical data for Dublin and weather observations scraped from dublinbikes API and Weather.com, respectively, is used to create a feature dataset. Various machine learning techniques are used and compared before choosing the best performing model.

**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 Chapter 1), 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. The problem of finding an appropriate optimal or near-optimal route for bicycle relocation using box trucks or trailers has been frequently addressed in the literature in the past years. However, our study seeks to address the problem of predicting demand for bicycles, which is an important step that anticipates choosing an efficient repositioning strategy.

**Problem statement.** This work will focus on predicting station-level demand for bicycles and empty docks in a station-based bike-sharing system using machine learning techniques and data-driven approach. One of the challenges of this problem that will be addressed is estimation of demand under the limitations of censoring, i.e. when target variable (demand) is observed only if there is no over-demand at the station. Data-driven probabilistic modelling will be used to estimate bike arrival and departure rates of the system as independent Poisson processes in a Markovian time-inhomogeneous queueing model.

**Delimitations of the study.**

The following interim objectives of the study can be stated:

- Data collection and preprocessing

- Modeling arrival and departure rates of bicycles

- Prediction of the actual demand under the limitations of censoring, i.e. when the target variable (demand) is observed correctly if there is no over-demand at the station, but is set to either 0, or the maximum capacity of the station if over-demand takes place.

- Using the obtained demand to train short-run forecasting models using suitable machine learning techniques

- Performance analysis on validation data and comparison with baseline models

**Professional significance.** 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 number of demanded bicycles or empty docks is 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: unsatisfied demand either for bikes, i.e. the station is empty, or for docks, i.e. the station is full and a customer can’t park and has to ride to a nearby station

Arrival and departure rates: the 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.

Предыдущие абзацы подрезать и доработать. Добавить фактические результаты.

**Chapter 1.**

**The history 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. 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.

**Literature review**

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 usage patterns of stations 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 usage.

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.

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 arrival and departure 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. employ a piecewise 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.

**Methodology and data**

This study will combine probabilistic modeling and machine learning techniques to predict actual station-level demand for bikes as well as empty docks using the case of Dublin, Ireland. The real-time data is published by JCDecaux, the French advertisement company that operates dublinbikes as well as many other bike-sharing systems in the cities of Europe. The data was collected during the period from 23 January 2017 to 28 August 2017 and consists of the number of bikes at each of the 102 stations observed every 2 minutes.

The methodology of this study comprises of the following steps.

First, an appropriate probabilistic model to forecast departure and arrival rates will be chosen. We adopt a Markovian time-inhomogeneous queueing model that views departure rate and arrival rate as independent time-inhomogeneous Poisson processes. It requires two assumptions to be made. One states that stations are independent in the asymptotic regime and is justified by theoretical results of other studies, so this approximation will be used in this study. Another assumption implies memory-less property and is justified in (Feng et al., 2016). The intensity rate of such processes is a function of time which is obvious for station modelling, as the demand for bikes or for docks is higher in the rush hours. In particular, we consider stepwise approximation of intensity, i.e. modelling intensity rate as a set of homogeneous Poisson process with different granularity, or time windows. In the study results obtained with different time windows will be compared and the best intensity function will be chosen according to adequate metrics.

Second, arrival rates during periods of over-demand will be approximated using historical data on station availability. What makes it possible is the rich dataset of station availability collected for this study. Collecting and averaging all the information about intensity rates that irregular rebalancing and fluctuations near over-demand border can give us, we will be able to estimate average intensity rates by station and by time. Furthermore, other features, e.g. temperature, precipitation, dummy variable equal to 1 on weekends and 0 on weekdays, etc. can be use when estimating intensity rates.

Finally, arrival and departure rates will be used to predict the number of bicycles at the station in a certain time horizon as if there were no censoring limits. These results will be used as target values to train short-run forecasting models using suitable machine learning techniques and then compared to results achieved by the same machine learning algorithms, but with actual number of bikes used as target values.

**Modelling arrival and departure rates. Introducing historical data based approach to estimate unobserved rates.**

Markov process is widely used in modeling the dynamics of stochastic systems and the state transitions of complex stochastic systems. A Markov process  is a stochastic process with the property that, given the system state at time , , for a time , the system state  is not influenced by the system states,  for , that is, prior to the time .  is a transfer matrix; the matrix elements are not negative, and the sum of all the various elements is equal to 1, expressed in the probability of each element. The element in the matrix is the probability that the bicycle will be retained, acquired, or lost.  represents the  step transfer matrix.

<https://academic.oup.com/ajae/article-abstract/47/3/742/161327?redirectedFrom=PDF>

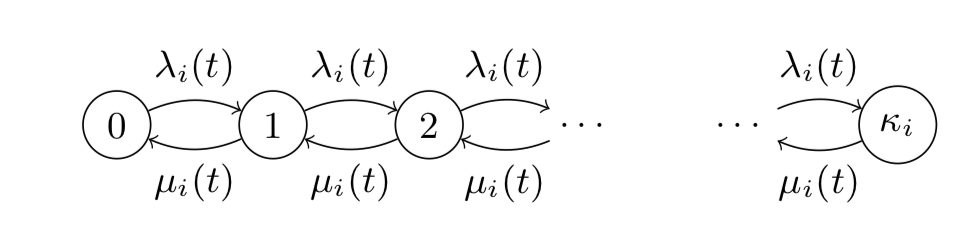
In this section we’ll introduce a single-station stochastic model of a bike-sharing system that can be used to obtain estimates of departure and arrival rates at given time. While the main concept has already been proposed by Gast et al. (2015), we provide more intuition on how this probabilistic model can be used to get point estimations. 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 interarrival 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.
* Определение 4.1. Пусть Λ(t) - некоторая дифференцируемая монотонно- возрастающая функция такая, что Λ(0) = 0. Неоднородный процесс Пуас- сона с функцией интенсивности λ(t) := Λ′(t) - это целочисленный неубыва- ющий процесс Nt, t ≥ 0 (“неубывающий процесс” означает, что Nt+h −Nt ≥ 0, ∀t≥0, ∀h>0),такойчто
* N0 = 0 п.н.;
* Nt имеет независимые приращения;
* (Nt+h − Nt) имеет распределение Пуассона с параметром (Λ(t + h) −  Λ(t)).
* 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 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.



Several assumptions are made to model bicycle stations this way.

* Firstly, processes of arrival and departure are Poisson distributed. This approximation 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. 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 its dynamic over time we settle on a piecewise constant form of intensity rates at discrete time frames, thus presenting each station as a sequence of homogeneous Poisson processes with intensity rates constant over certain time intervals. These assumptions allow us to be able to evaluate model parameters from historical data. The inter-arrival times at each time frame are distributed exponentially. (переформулировать, уж очень повторяю работу киаротти, юзнуть док-ва из статьи по ссылке ниже

* Хотя скопипастить его формулу транзишен пробабилити можно было бы

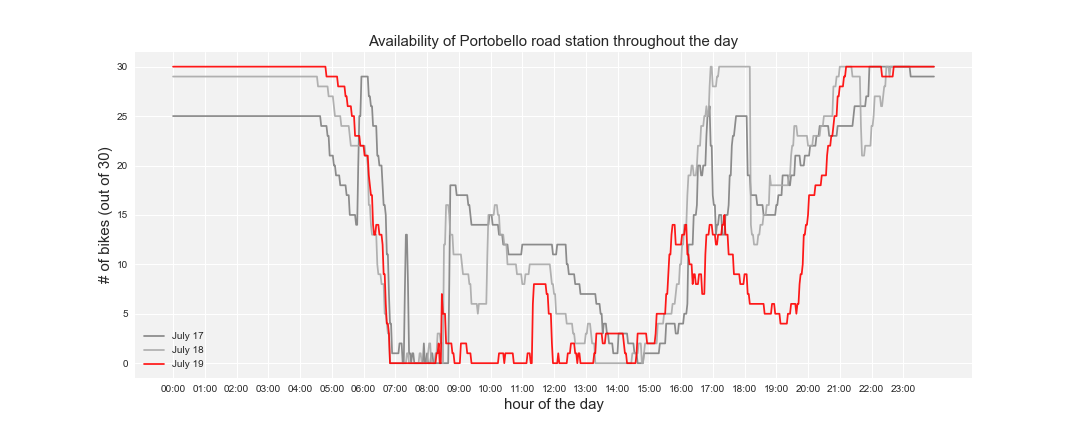
Потом эвристика оценки на over-demand platoes в контексте реальных данных с неуказанными поставками новых велосипедов)

* My Bible: [http://mescal.imag.fr/membres/nicolas.gast/papers/cikm2015.pdf and Sensors 2018 p.6](http://mescal.imag.fr/membres/nicolas.gast/papers/cikm2015.pdf%20and%20Sensors%202018%20p.6) a dynamic approach to rebalancing bike-sharing systems Chiariotti

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 the 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’ve already seen, behavior of a certain station is asymptotically independent from other stations. However, it definitely depends on weather conditions, season and whether it is a working day. 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. Or maybe test a hypothesis that they have the same shape but different constant coef ?
* For the empirical study of this paper we will narrow our observation period to working days of summer. We will estimate different intensity rates for two weather states depending on whether or not it was expected to rain at that time interval. We will neglect temperature variations assuming they do not account for dramatic change in intensity rates as most of trips during workdays are still made by people moving to work.

**Heuristics for estimating intensity rates during overdemand periods.**



While presented with the data

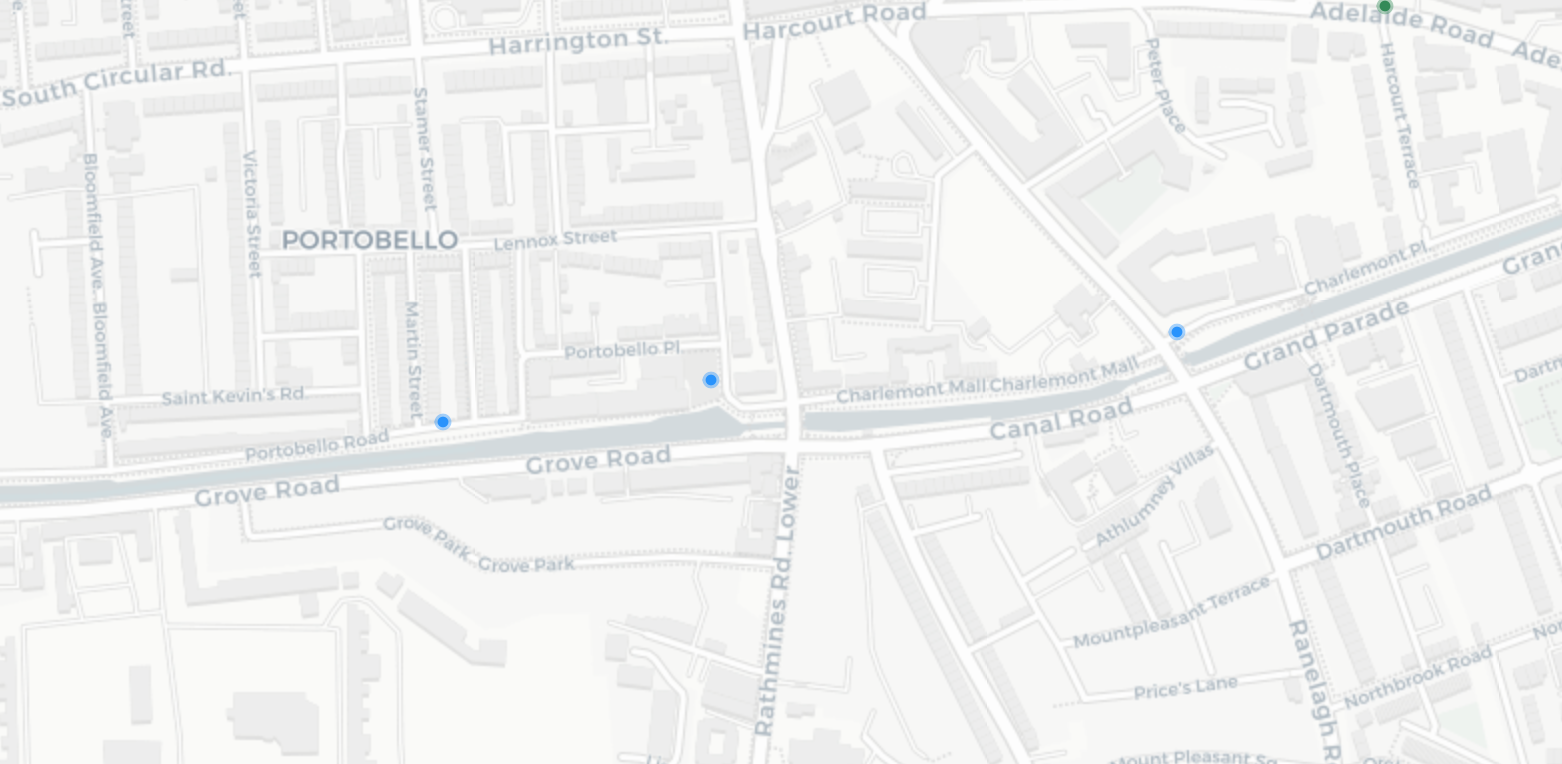
Portobello road and strange peaks in availability.

* Charlemont is the closest ‘metro’ <https://saletur.ru/Ирландия/Дублин/maps/> or actually the closest public transport stop.





* At least one station is closer to metro than Portobello road!



**Anticipated results**

When approximating arrival and departure rates by a stepwise function, it is expected that we will be able to choose an optimal time window that would preserve information about the variation of arrival and departure rates, but also have enough generality. As far as the second step in the proposed methodology is concerned, the dataset for dublinbikes consists of about 720 observations for 217 days, so we expect to have enough data to conduct adequate averaging of the rates on the intervals where over-demand occurs frequently.

All in all, we expect to train a set of models for each station that will get real-time status of the stations as an input and predict the number of bicycles that will be demanded at this station in several hours. To compare the predictive power of our models to other baseline models, we will develop a metric that takes only intervals of observed demand into consideration.

**Conclusion**

Replacing the observed numbers of bikes at the station with the numbers reflecting actual demand as if the stations had no upper or lower limitations of capacity would 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. This model could then be used to provide input demand values for efficient route designing models solving pickup and delivery problem. However, comparing this model to baseline ones could be difficult or not in favor of the proposed model. As we cannot assess how accurate unobserved demand predictions for over-demand situations are, the models could only be compared by their performance inside the capacity limits. So even if this model were accurate for predicting peak values of over-demand, which is still suitable as it is its main purpose, it might show lower quality of results on the interval within capacity limits of the station. The validation of these results could be done by actually expanding the capacity of any of the stations prone to over-demand.

**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.

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).

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.

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)