NHH

Norwegian School of Economics

Bergen, Fall 2020

Station-Level Demand Forecasting for Effective Repositioning for Bergen's Bike-Sharing System

Modeling Bike and Slot Demand utilizing Forecasting and Short-Term Repositioning Strategy with Optimization Methodology

Aida Suleimenova and Ping-Yi Chiang

Supervisor: Stein W. Wallace

Master Thesis in Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Problem Description

The bike-sharing system contributes to reducing frequent traffic congestion and provides an environment-friendly way of mobility for the citizens. In a bike-sharing system, users rent a bike from a station, perform a ride, and then deliver a bike to a different or the original station. Due to the city topography as well as the localization of housing versus educational and office facilities some stations are more popular for starting and ending a ride. This phenomenon commonly leads to the state when some stations become full and others empty, making users unable to find available bicycles or free slots in a station. The imbalance problem has a significant impact on the service level and attractiveness of the service for users. Therefore, repositioning is needed to maintain the appropriate number of bikes at each station. Usually, a vehicle or set of vehicles is utilized to balance the fleet of bikes. Inefficiency of the repositioning processes could lead to increased costs in terms of logistics and customer dissatisfaction. However, using a manual approach to plan and execute the repositioning activities is still common practice for bike-sharing systems. Thus, it is therefore a large potential for increased customer satisfaction and cost reduction with a decision support tool based on operations research methodologies. This thesis's primary purpose is to provide a solution method to determine the number of bikes that should be added or removed at each station and solve the repositioning problem for the stations as well as finding the optimal routes for the service vehicles.

Contents

List of Figures

List of Tables

Abstract

Bike-sharing systems (hereafter BSS) have become popular globally and provide positive changes to congestion and environmental concerns in cities. Recently, almost all big cities adopted a bike-sharing system. These systems allow users to access bicycles and return them almost everywhere in the city without thinking about maintenance. However, the increase in popularity and stochastic demand make the planning of operational processes for BSS's operating companies challenging. In particular, the uncertainty of the demand could be linked to the unavailability of bicycles or empty slots at some stations. Modeling bike-sharing demand has been a major research question in the scientific community. However, most of the studies have tried to predict the global (system-level) demand. This in most cases is not sufficient for the purpose of optimizing the operational planning, considering it requires demand prediction for each station. This thesis examines the repositioning of bicycles in bike-sharing systems. The major objective of the thesis is to model the bike-sharing demand and the optimal repositioning strategy.

The BSS in Bergen is chosen as a sample case in the thesis. One of the major problems which Bergen's BSS faces is the imbalance of bikes. To solve the imbalance problem, Bergen City Bike, the company operating the bike-sharing system in Bergen, performs repositioning intuitively. However, repositioning only based on human experience might lead to a choice of inefficient routes or the wrong number of bikes at a station, causing higher transportation costs and/or unsatisfied demand.

In this research, an applicable repositioning tool is proposed for Bergen's BSS. First, a regression with ARIMA errors model and the random forest model are developed to model the station-level bike demand on a rolling basis, considering the seasonalities, weather and weekend as dummy variables. The parameters of the regression with ARIMA errors model are determined by performing the Hyndman-Khandakar algorithm. Subsequently, a model with better performance is determined by using the time-series cross validation. Second, the target number of bicycles to be repositioned can be decided, by adopting the real-time data collected by Bergen City Bike and the point forecasts developed in the first part, which then are applied as parameters to the repositioning route optimization model. A Mixed-Integer Nonlinear Programming model to optimize the repositioning route is developed statically and deterministically, with the consideration of the transportation cost, unsatisfied demand cost, and the capacity of the vehicle and bike stations. Finally, the results of the model are used to define a repositioning strategy.

1. Introduction

1.1 The Concept of BSS

BSS is a service where bicycles are made available for short-term rentals. BSSs aspire to provide the public with a sustainable and convenient mode of transport in urban areas. Many BSSs allow users to borrow a bike from a station and return it at another station belonging to the same system. Stations are special bike racks distributed within the city that lock the bikes. Each station has a finite number of slots where the bikes can be locked. The users enter their information to unlock the bike and return the bike by placing it in the slot. Accessing the bike or returning it to the station is only possible if there is an available bike or slot. However, due to uncertain demand, there are often either no available bikes or no available slots. To avoid these situations, the bike-sharing companies utilize vehicles to redistribute bikes between stations.

While some BSSs can be free of charge, most require a subscription fee or a rental length-dependent fee. Many BSSs encourage short trips by offering subscriptions that make the first 15-45 minutes free of charge. In most cases, BSS is a more affordable option than using public transport or driving a private car.

1.2 History of BSS

The idea of a BSS was first proposed by Luud Schimmelpennink, a well-known Dutch innovator and public figure. He was one of the main inspirers of the "White Bicycle" plan (dutch: Wittefietsenplan), proposed in the mid-1960s in Amsterdam. The program's goal was to reduce the intensity of automobile traffic, which resulted in citizens being trapped in hours-long traffic jams in the city streets. The plan called for 20,000 white-painted bicycles to be installed at special stations around the city. It was assumed that one could use these bikes for free. Together with the other "White plans," this project was sent to the municipality of Amsterdam but was rejected. In response, supporters of the plan, members of the youth group Provo tried to place some white bicycles around the city, but this action was unsuccessful. The police immediately removed them from the streets, referring to municipal legislation, according to which citizens were forbidden to leave private property unattended. Although Luud Schimmelpenninck was elected to the Amsterdam Municipal Council in 1967, he could never get his plan approved (Gauthier et al., 2013).

In 1999, in Amsterdam, again at the suggestion of Schimmelpennink, the first technology using smart cards to access a Bicycle sharing system was implemented as part of a pilot project. There were 250 bicycles in the five-station system. The Dutch Bank Postbank provided smart cards. Subsequently, Postbank lost interest in this project. Unfortunately, the project was closed (Van der Zee, 2016).

Despite the failure, this plan was a driving force to further implement similar systems around the world. In 2002, the JCDecaux Corporation engaged Luud Schimmelpenninck to design a similar system in Vienna. This project proved to be successful. As a result of its implementation, the Citybike Wien rental system was launched in 2003. Then, JCDecaux Corporation, which has owned the Cyclocity brand since 2003, put into operation the Vélo'v rental system in Lyon, and in 2007 the Vélib' system in Paris (Le Figaro, 2011).

After 2007, the BSSs began to develop actively in many major cities in Europe and the United States. China has become the world leader with the highest total number of bicycles. (Gray, 2017). The BSSs' adoption of technology has developed with time; this is especially true for mobile applications, the use of Global Position System (GPS), smart card systems, and the use of machine learning.

1.3 BSS in Bergen, Norway

Norway has a developed and well established public transport system, where bike-sharing appears to be a natural choice of how people in the cities move. The shared bikes are available when one needs them and where one wants to use them. There is also trust in digital solutions within the population accustomed to new technologies (European Commission, 2017). Moreover, the robust network connectivity throughout the vast majority of Norway encourages cycle-sharing schemes. All these factors make Norway's BSSs have one of the best shared-bike usage in the world (The Local, 2019).

Urban Infrastructure Partner Group (UIPG) is a Norwegian company which finances, operates, and provides technology for shared urban infrastructure. UIPG currently operates three BSSs in Oslo, Bergen, and Trondheim.

Bergen City Bike bike-sharing service was launched in 2018 and is a service aiming to reduce traffic jams and greenhouse emissions. Users can find a bicycle near them, unlock the bicycle, ride up to 45 minutes, and return the bike to the station with the use of the mobile application after membership registration. Figure 1.1 illustrates the design of bicycles. Figure 1.2 shows the interface of the mobile application Bergen Bysykkel, where the user can access real-time information about available bicycles and slots. Figure 1.3 demonstrates the map of all active stations in Bergen. A total of 936,453 trips was made in 2019.

Bergen City Bike offers three types of subscription:

- Day pass: unlimited number of trips of 45 minutes for 24 hours for NOK 49.
- Monthly pass: unlimited number of trips of 45 minutes for 30 days, the first month is for NOK 49, auto-renewed for NOK 149.
- Annual pass: unlimited number of trips of 45 minutes for 365 days for NOK 399.

The subscription starts with the first trip and includes unlocking the bicycle and a 45-minute trip. After 45 minutes to 6 hours, the user is charged NOK 5 per 15 minutes, which is cheaper than a day pass for the Skyss bus for a price of NOK 100 and 30 days pass for NOK 800 (Bergen City Bike, 2020). In 2019, there were 99 stations in total, opened from 5 AM until 1 AM.

Figure 1.1 Picture of bicycles at a station in Bergen. Source: (Bergen City Bike, 2020)

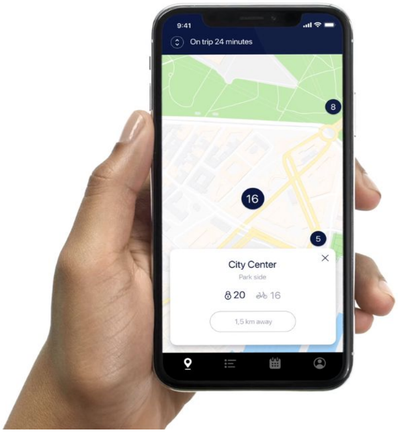


Figure 1.2 The Bergen Bysykkel mobile application. Source: (Bergen City Bike, 2020)

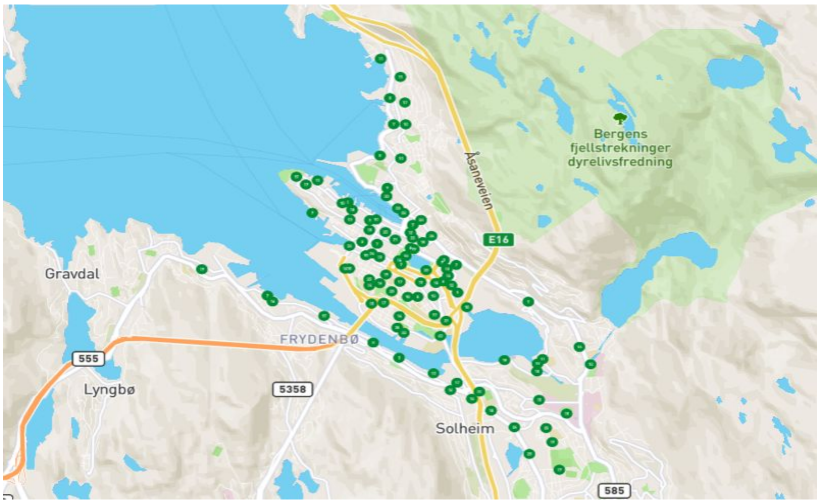


Figure 1.3 Spatial distribution of stations in Bergen. Source: (Bergen City Bike, 2020)

1.4 The Main Challenges Operating BSS

Companies operating BSS encounter many difficult decisions considering the location of the station, the number of slots in the station, the number of bicycles per capita, and the system's central area. Many companies have faced difficulties in deciding the fleet size and the number of stations, which resulted in a higher chance of failure (Sun et al., 2018).

For most BSSs, the capital cost, which includes bicycle investments, the installment of stations, and the establishment of IT infrastructure, are substantial. There are also costs associated with the maintenance of stations, IT systems, and bicycles. Most of the time, initial investment, maintenance and operations cost could not be covered by revenue sources. Usually, BSSs' financing is maintained by the combination of fees and government subsidies (DeMaio, 2008). Many BSSs are supported by charity fundraising. Moreover, the repositioning operation is one of the most significant components of operating expenditures (Andersen, 2016). This makes the planning of repositioning one of the most important problems the BSS' operating company could face.

The major problem for bike-sharing companies is to guarantee the availability of bicycles and empty slots. Avoiding customer dissatisfaction, associated with the inability to rent a bicycle at the desired station or return the bike near to the final destination, has been the biggest challenge. The customers emphasize the importance of availability at peak hours, as a significant share of users rely on BSS as a standard commuting option (Hughes, 2017). Thus, the decision regarding how to approach repositioning the bikes is critical both in terms of operating cost and service level.

The problem with repositioning is to forecast the demand for bicycles and empty slots at each station, as the repositioning strategies are highly dependent on the demand. However, it is not possible for BSS operating companies to know real historical demand, since there is no way to record unsatisfied and lost demand. The forecasting upon historical rentals data is commonly used in practice instead, considering that the lost demand is usually neglectable (Hulot et al., 2018).

According to the operation manager of Bergen City Bike, one of the biggest challenges for the BSS in Bergen is that the stations, which are located uphill, get emptied as people rent the bike from these stations but choose not to return bikes there. Therefore, the imbalance problem has a significant impact on the service level and attractiveness. Consequently, repositioning needs to be executed to maintain the appropriate number of bikes at each station. Another challenge for BSS in Bergen is that Bergen City Bike utilizes an intuitive model for bike repositioning that might lead to an inefficient route. The team moves at around 100 bikes a day. Hence, the primary focus of this thesis is to collect outflow and inflow information and develop an appropriate repositioning model.

2. Literature Review

Numerous studies have analyzed BSS; in particular, academics and practitioners have investigated the effect of BSS as a mode of transport and conducted detailed reviews of the bike-sharing schemas and business-management models.

Referring to BSS as a mode of transport, many studies have found that bicycle trips are mainly substituting a bus trip and walking rather than a trip by private vehicle (Bullock et al., 2017; O’Neil & Cuilfield, 2012). There is a positive synergy between BSS and the public transport networks, in case BSSs are well integrated within the bus routes and rail systems. Typically, users use bicycles to reach the areas not covered by other public modes (O'Neil & Caulfield, 2012). The analysis BSS in Lyon found that the demand for bicycles doubles when other public transport modes are on strike (Jensen et al., 2010).

Although BSS’s adoption is continually growing and can be a way to approach concerns associated with global climate change, the future demand and popularity for BSS are still uncertain. The scarcity of bicycles may impede the popularity of BSS due to unsatisfied demand. On another hand, the significantly higher number of bicycles in areas with a low utilization rate is financially unsustainable. Moreover, more research is needed to understand the effects of the business model and operational decisions on BSS’s benefits in terms of its long-term sustainability. (Shaheen et al., 2010).

The thesis focuses on techniques and tools that could provide sound decision-making tools on the demand prediction and the strategy for repositioning. Therefore, the literature review will mainly highlight the studies of demand prediction and repositioning methods.

2.1 Demand Modeling

Modeling the demand for bicycles and empty slots can help BSS operating companies to allocate bicycles better by providing support to strategy makers and managers in search of optimized decisions. Thus, the modeling of bike-sharing demand has recently received significant attention among researchers.

Many papers analyze the relationships between the demand and factors which possibly could affect it. The first conducted studies suggest that bike-sharing demand is dependent on the month, the weekday, the hour. Other than time, it has been proven that there is a dependency on the temperature, humidity, wind speed, and neighborhood of the stations (Borgnat et al., 2011; Gebhart & Noland, 2014; Mahmoud et al.; Vogel et al., 2011; Vogel and Mattfeld, 2010).

Yin et al. (2012) predict the system-level demand of BSS in Washington using time and weather data as independent variables. The paper indicates that the problem is highly nonlinear. Therefore, the gradient boosted tree method is utilized to predict demand. Li et al. (2015) also predicted the demand with the use of the gradient boosted tree method by clustering the stations based on geographical distance. Yoon et al. (2012) provide a model to predict the demand using the ARMA (AutoRegressive Moving Average) method. Rudloff and Lackner (2014) propose a model to predict a station-level demand with the use of neighboring information to improve its prediction.

Wang (2016) presents the analysis of BSS in New York City and the prediction of its demand with the use of weather and time features as predictors. The paper proposes a random forest regressor method to erase the missing weather data problem. It also has been found that the log transformation of the number of trips significantly improves the model's performance.

Zhang et al. (2016) predict the final destination station and arriving time for users using the information about the departure station and time. The research proves the time dependency of trips.

The majority of conducted studies focus on global (system-level) demand. However, such models are able to capture only global behavior and patterns and are prone to underfitting the data. Moreover, in real-life situations, the BSS operating companies usually need a prediction of demand per station. The stochastic nature of the demand makes the task of developing the prediction model per station challenging, as the uncertainty is more present at this level. To address this problem, some proposed methods are to group together the stations geographically (Li et al., 2015). The cluster-level predictions are usually accurate and easy to interpret. Nevertheless, it is not applicable to the situations when the terrain affects the attractiveness of a station as an origin or a final destination, and for the systems, where the major share of the rentals are from very few stations.

Several papers focus on predicting the state of the system. Cagliero et al. (2017) predict if the station gets full or empty using Bayesian classifier, decision trees, and SVM. The hour, the day, and a dummy variable indicating whether the day is a working day are used to predict the state of the station. Yoon et al. (2012) present a model to predict the changes in the network state using the real-time system status. They use the clustering approach and propose an ARIMA (AutoRegressive Integrated Moving Average) model which makes predictions based on clusters of stations, time and weather features. Clustering is performed with a KNN (K-nearest neighbors) method.

Rudloff and Lackner (2014) introduce a model to predict the station-level demand. The presented model gives predictions of the demand per station for an hour using time and weather features. The linear regression method is utilized, with the use of time, weather, season, week day, temperature as categorical variables. The state of a close station is also considered in the model. The research also tries to answer whether there is a dependency between the demand and a critical state (empty or full) of the neighboring station, however, a clear influence was not confirmed. Also, the paper concludes that the Poisson distribution is not always the most appropriate distribution for the historical data and that the negative binomial and zero-inflated distributions could provide similar results. However, the model was only tested for BSS in Vienna and the conclusions could be different for smaller-sized BSS.

Yin et al. (2012) apply ridge linear regression, SVR (Support Vector Regression), random forest, and gradient boosted trees methods to the demand predictions. The study indicates that random forest is the most promising method to achieve the best performance. However, the problem was simplified to system-level prediction, while the full problem would be predicting the station-level demand, considering that the main task is to match the demand at each station by repositioning bikes.

The prediction of the demand in a major share of papers is based on historical rentals data. The historical rentals data illustrates only satisfied demand, but due to the nonavailability of bikes and empty slots, part of the demand is lost. Several papers address the lost demand by assuming that the rentals follow a Poisson distribution (Brinkmann et al., 2015; Shu et al., 2010; Alvarez-Valdes et al., 2016). However, it is not applicable to every BSS's historical data.

Typically, the lost part of the demand is minor compared to the total demand (Hulot et al., 2018). The problem is also partly addressed by introducing mobile applications by BSS. The lost demand is neglected in Vogel and Mattfeld (2010); Caggiani and Ottomanelli (2012); Mahmoud et al. (2017); Yin et al. (2012); Schuijbroek et al. (2013).

2.2 Repositioning

The repositioning of bicycles problem is among the most addressed problems with operating BSS. While the demand prediction is mainly used to plan the repositioning, the repositioning strategies are applied in real-time.

The repositioning problem can be approached as an optimization problem. The goals can be to minimize the transportation cost and/or the cost of unsatisfied demand. It is the pick-ups and returns problem with a fixed number of vehicles used to perform the redistribution and stochastic demand. The problem needs to consider the number of stops, the state of stations, the number of vehicles, the number of bikes to redistribute, and the time step.

Benchimol et al. (2012) provide a model to minimize company cost with one truck with finite capacity and neglecting the time capacity. Chemla (2012) presents a model to minimize the traveled distance with the use of two trucks and assuming that there is an infinite amount of time. Raviv and Kolka (2013) propose a solution which minimizes the lost demand and maximizes the service level. Several papers conclude that the repositioning performed by the operator could not be replaced by offering incentives to users to return bicycles to certain stations (Chemla, 2012; Fricker and Gast, 2013; Waserhole and Jost, 2012).

The most common approach is to generate the demand from a Poisson distribution (Raviv and Kolka, 2013; Vogel et al., 2016; Vogel and Mattfeld, 2010). The Poisson distribution is suitable most of the time, given that it explains independent pick-ups and returns in each station. However, these models do not consider the dependency between the demand the external factors such as time and weather.

Schuijbroek et al. (2013) propose a model with intervals to decide when to perform the repositioning. Caggiani and Ottomanelli (2012) model the urgency of rebalancing a certain station with the use of a fuzzy logic algorithm to model the urgency of rebalancing a given station. These approaches are more applicable, given that they consider the variability of the demand. Nevertheless, the time and weather influence on the demand is not captured in proposed solutions. Despite the fact that the bike-sharing demand nature is stochastic, a deterministic model could be applicable for the repositioning strategy with the input from the demand prediction model which explains the variability and the dependency on time and weather features. Moreover, in real-life scenarios, the computational feasibility concern is present, making the problem unsolvable without introducing assumptions. Whereas the deterministic approach with predicted demand input is a feasible method for commercial software.

3. Data Explanation and Processing

The first step to build a model is to clean and preprocess the data. This chapter explains the trip data used to build models for Bergen's BSS. It analyses features of the data and the irrelevant data to be removed.

The trip data information consists of transactions information from January to December 2019 (a year). Each sample of information includes the timestamp of when the trip started, the timestamp of when the trip ended, duration of the trip in seconds (further converted into minutes for the sake of simplicity), the unique ID for start station, name of start station, description of where start station is located, the latitude of start station, the longitude of start station, the unique ID for end station, name of end station, description of where end station is located, the latitude of end station, the longitude of end station. The format of the features is illustrated in Table 3.1. The original data set consists of 923,923 records, excluding canceled trips and bikes moved by Bergen City Bike Team. After data screening, we removed 2% of transactions from the original data set, which included trips with a duration of more than 6 hours (12911 records) and trips that were started and ended at the same station with a duration of fewer than 2 minutes (6402 records). Given that, the trips with duration longer than 6 hours are violations of terms of use. The trips with the same departure and return station with a duration of fewer than 2 minutes are possibly trips which users failed to cancel in the mobile application (the users who could not cancel the trip in mobile application would lock the bike at the same station where they unlock it). Moreover, these trips do not represent movements and change in the state of a station. Thus, the trips with duration of fewer than 2 minutes which started and ended at the same stations are removed from the data set.

Information	Format
Timestamp of when the trip started	Timestamp
Timestamp of when the trip ended	Timestamp
Duration of trip in seconds	Integer
Unique ID for start station	String
Name of start station	String
Description of where start station is located	String
Latitude of start station	Decimal degrees
Longitude of start station	Decimal degrees

Information	Format
Unique ID for end station	String
Name of end station	String
Description of where end station is located	String
Latitude of end station	Decimal degrees
Longitude of end station	Decimal degrees

Table 3.1 Transaction information and its attributes (Source: Bergen City Bike, 2020)

The mean trip duration is 10.83 minutes, and the standard deviation is 12.55 minutes. Figure 3.1 shows the distribution of trip duration.

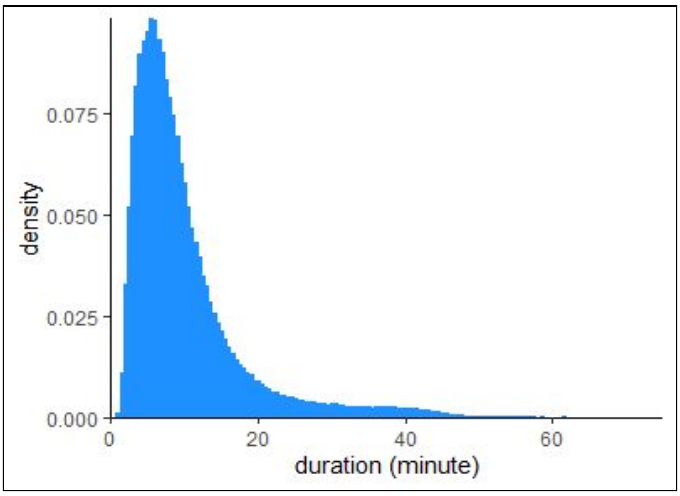


Figure 3. 1 The distribution of trip duration

There is an evident linear relationship between the number of pickups and returns regarding the stations, and the number of pickups and returns are similar for most stations, according to Figure 3.2. However, there are some outliers whose number of pickups and returns are far from to be equal. The observation responds to the statement in Section 1.4: imbalance problem does exist in the BSS in Bergen. Also, it implies that the pattern of pickups and returns might not be similar for every station, and the patterns for each station are not identical.

Moreover, ten of the most popular bike stations as a start point, which are the labeled points in Figure 3.2, are also the most popular bike stations for a return point. Most of these stations are centrally located.

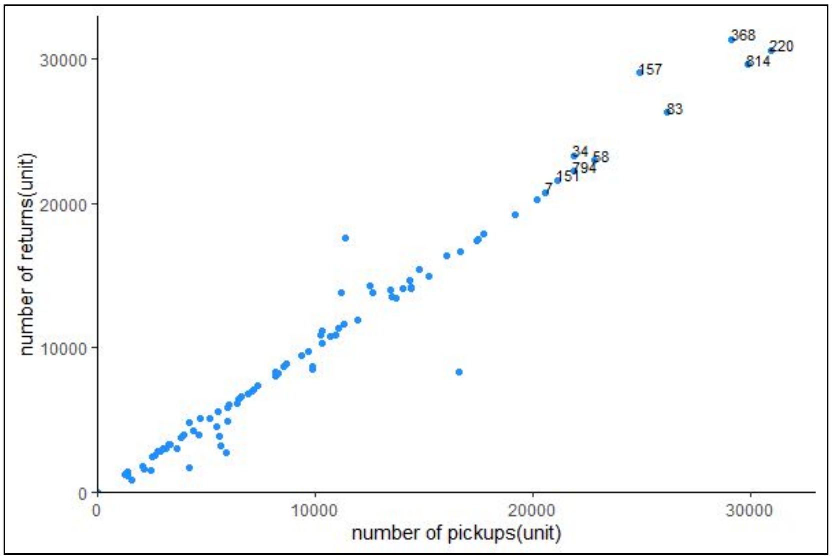


Figure 3. 2 Bivariate distribution of pickup and return number

Further, according to Figure 3.3, there is a negative correlation between the week's day's popularity and its mean trip duration. Thus, users tend to travel more during the weekdays than on weekends. However, the mean duration of trips is higher during weekends.

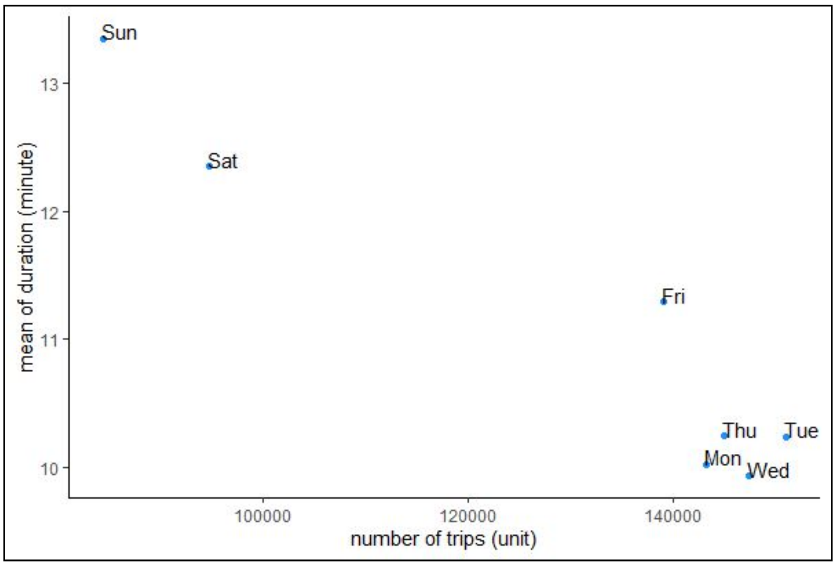


Figure 3. 3 Bivariate distribution of popularity of the day of the week and its mean trip duration

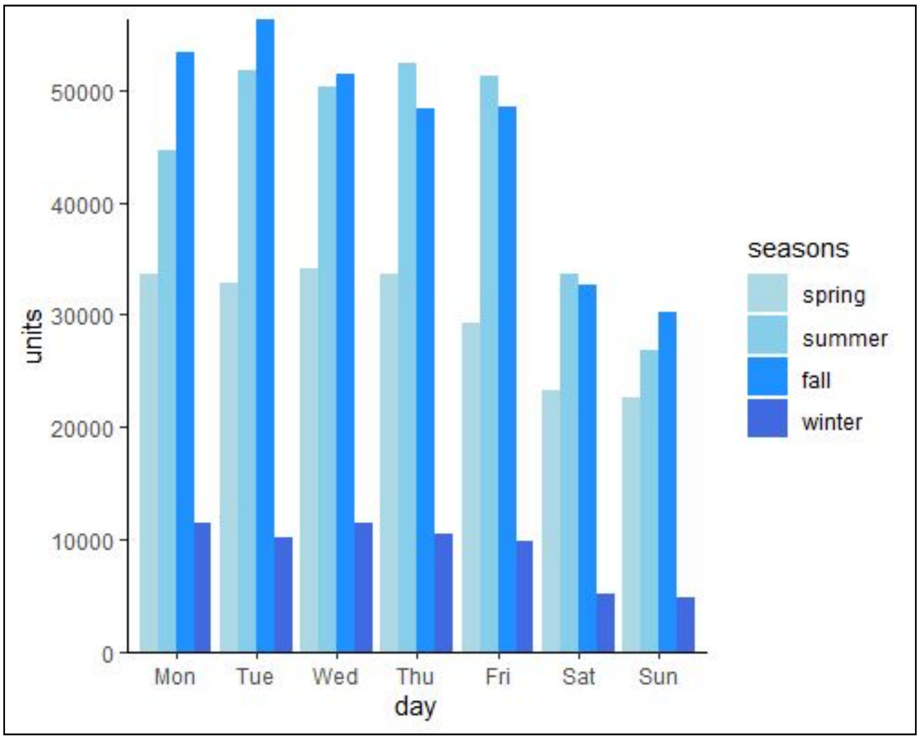


Figure 3. 4 Total bike rentals by days over seasons in 2019

Figure 3.4 presents the bike rental was higher on weekdays than that on weekends in all seasons, implying the bike rental pattern on weekdays and that on weekends are different. The finding is consistent with what we find in Figure 3.3. Also, Figure 3.4 illustrates that the rentals vary a lot among the four seasons. Therefore, different weather conditions might also be important factors influencing bike rentals.

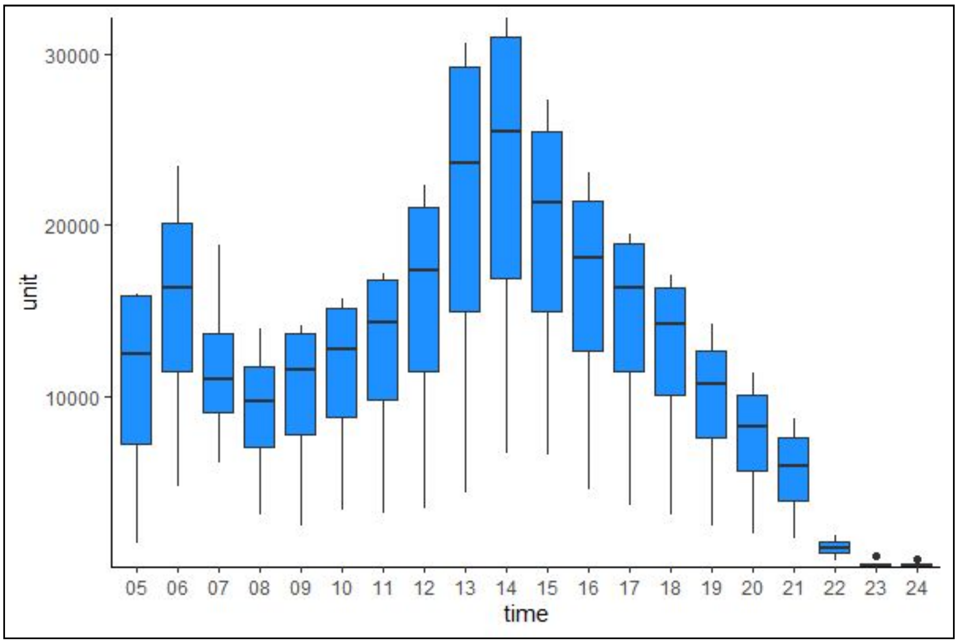


Figure 3. 5 The daily rentals over time in 2019

According to Figure 3.5, there were two peaks in a day: from 6 AM to 7 AM, and from 13 PM to 15 PM. Therefore, there might be daily patterns for bike rentals. We plot the bike rentals over time in May 2019 and decompose the time series to observe detailed facts and retrieve information to make a more reliable conclusion on the daily patterns.

Many techniques, such as classical decomposition, X11 decomposition, SEATS decomposition, and STL decomposition, are frequently used to decompose time series data. Among all the decomposition methods mentioned above, STL decomposition is adopted in this case, given that, unlike X11 decomposition and SEATS decomposition, the STL decomposition method can deal with different types of seasonality and handle hourly data (Hyndman & Athanasopoulos, 2020). Moreover, it is robust to the outliers in the data, which means that unusual patterns will not be included in trend or seasonal patterns and the pattern interpretation of a time series will not be based on non-frequently happened events. The STL decomposition of bike rentals time series in May 2019 is shown in Figure 3.6.

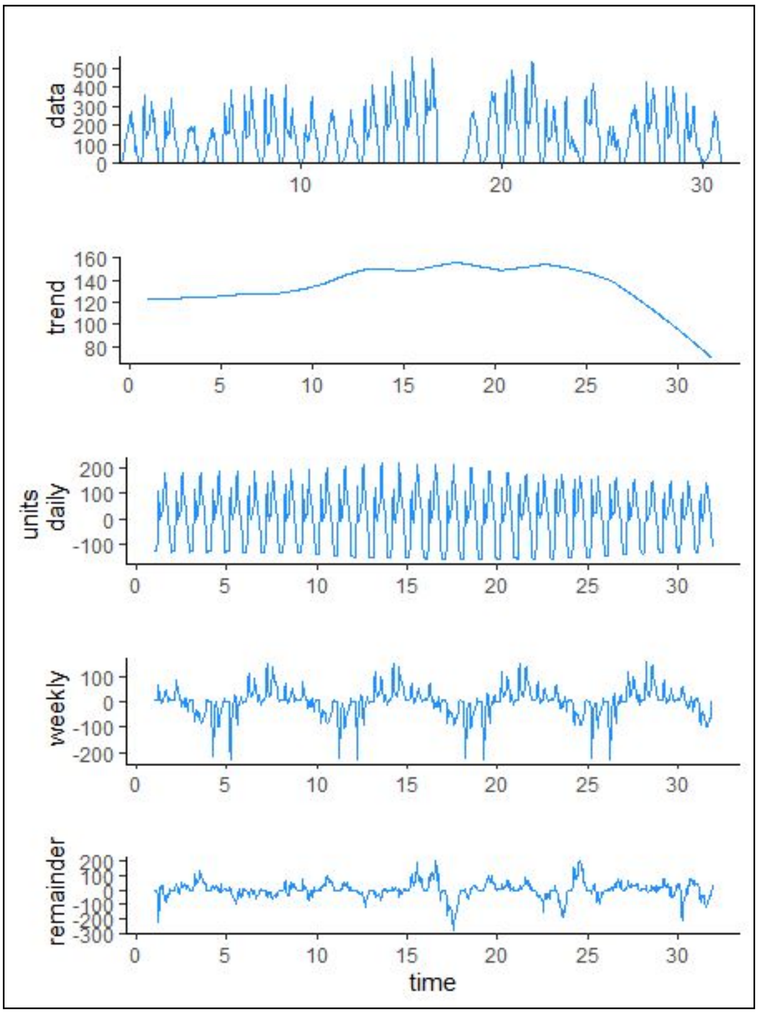


Figure 3. 6 STL decomposition of bike rentals in May 2019

Figure 3.6 indicates that the total bike rentals in May 2019 have daily and weekly seasonality. According to Figure 3.6, the bike rentals time series is decomposed into four parts: trend, daily pattern, weekly pattern, and remainder. There are two peaks in every weekday from the third panel in Figure 3.6. This finding is consistent with what we observe in Figure 3.5. From the fourth panel in Figure 3.6, the number of bike rentals remains at a similar level on weekdays and decreases on weekends. The weekly seasonality is also in line to the observation in Figures 3.3 and 3.4.

In conclusion, most of the popular stations are centrally located. The rentals of Bergen's BSS have daily and weekly patterns, and the weather might be a crucial factor influencing bike rentals. Furthermore, the patterns of pickups and returns are different for each bike station, and the patterns of each bike station are not identical to each other.

4. Methodology

Based on the literature review and data explanation, the choice of the methodology is explained in Section 4. Section 4 outlines the framework of the proposed model in Section 4.1. The methodologies for modeling the demand, the target number of bikes to be repositioned, and the optimal repositioning route are discussed in Section 4.2, 4.3, and 4.4, respectively. The data input is introduced in Section 4.5. Finally, the technical tools to be used in the thesis are presented in Section 4.6.

4.1 Framework

As demonstrated in Figure 4.1, there are three primary parts of the repositioning model developing process: the demand forecasting model, the formulation of the target number of bicycles to be repositioned, and the repositioning route optimization model. A fundamental problem of bike repositioning is determining the target number of bikes to be added or removed for each station. To solve this problem, the regressions with ARIMA errors (hereafter ARIMA model) and the random forest model are developed in our research to forecast the net station-level demand. The point forecast of the net station-level demand presents the number of bikes and empty slots needed at a station. Afterward, the point forecast from the model with better performance and the real-time information of the available bikes and empty slots are applied as input to determine the target number of bikes to be repositioned for a station. Finally, the repositioning route optimization model is developed based on the target number of bikes to be repositioned and each station's real-time data.

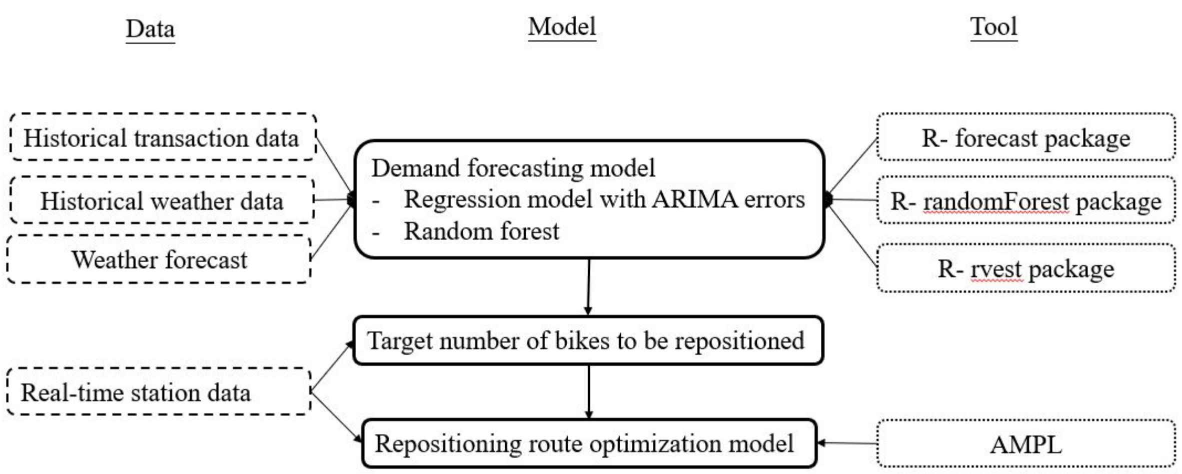


Figure 4. 1 The framework

A similar framework of the repositioning model was also adopted in Regue & Recker (2014). However, the model proposed by Regue & Recker (2014) does not consider real-time station data. The inventory level of bikes for each station was calculated by the model in their work. For Bergen's BSS, real-time station data is available due to current technology improvement, and it is closer to reality. Therefore, we developed the solution with the consideration of real-time station data, as mentioned above.

4.2 The Demand Forecasting Model

The models for forecasting station-level bike demand are trained separately for each station by implementing the ARIMA and the random forest methods. The introductions of these two models are in Section 4.2.1 and 4.2.2, respectively. The patterns of each bike station demand might differ from one another, according to the existence of outliers demonstrated in Figure 3.2. Although developing the model for forecasting system-level bike demand is a simpler task, there might be bias when adopting a system-level demand forecasting model to forecast the station-level demand due to the fact that the station-level pattern differences would be ignored (Lin et al., 2018). Therefore, to train forecasting models for each station separately is a more suitable choice, rather than train a system-level demand forecasting model for the BSS and apply it to forecast the station-level demand.

The net demand for bike stations consists of two parts, bike pickups, and bike returns. Figure 3.2 illustrates that not all stations had similar numbers of pickups and returns during 2019. As concluded in Section 3, there are different patterns for pickups and returns for some stations. Consequently, training the forecasting model for the bike pickups and returns separately is more accurate than aggregating pickups and returns to train them together. Therefore, the number of pickups (demand for bicycles) and returns (demand for empty slots) forecasting models are developed separately in the thesis.

The demand forecasting model is trained on a rolling basis, which means the data used to train the model is updated over time. For example, in the thesis, the developed model trains the bike pickups and returns data from May 1, 2019, to May 7, 2019, to forecast the bike demand on May 8, 2019. Bergen City Bike collects trip data for each station continuously. Therefore, the model can be trained with the most recent data, which makes the model more applicable. The bike demand forecasting model built on a rolling basis was also adopted by Froehlich et al. (2009).

In the following three sections, we introduce the demand forecasting models adopted in the thesis, the ARIMA and random forests models, and their validation. The reasons why these two methods are chosen are also stated below.

4.2.1 The Regression Model with ARIMA Errors (the ARIMA Model)

Unlike regression models, the ARIMA model allows autocorrelation in the error term of regressions. It assumes the error term from regression follows an ARIMA model, where the error term is a white noise series. The ARIMA model, namely the AutoRegressive Integrated Moving Average model, forecasts the future variable of interest by using a linear combination of past values of that variable and its past forecast errors with consideration of differencing (Hyndman & Athanasopoulos, 2020). Therefore, it can rigorously capture the seasonality in the data.

The Hyndman-Khandakar algorithm is utilized to determine the order of the ARIMA model in the thesis. This method is widely applied in previous literature to determine the ARIMA's parameters (Andrysiak et al., 2014; Krishna et al., 2015). The Hyndman-Khandakar algorithm for automatic ARIMA modeling is shown in Table 4.1. P and p are the orders of auto-regressive factors, D and d are the orders of differencing, and Q and q are the orders of moving average. The upper case denotes the order for seasonality parts, and the lower case displays that for non-seasonality parts.

Hyndman-Khandakar algorithm for automatic ARIMA modelling ARIMA $(p, d, q)(P, D, Q)$
1. First, the order of differencing from 0 to 2, which are d and D in the model, are decided by using Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests.
2. Hyndman-Khandakar algorithm uses a stepwise search to determine the orders of p, P, q and Q by minimizing the AICc (AICc, the abbreviation of corrected Akikae's Information Criterion, is an estimator of prediction error with consideration of sample sizes and inclusion of a penalty term of number of parameters.). There are four models mentioned below that are fitted first. A constant term is considered unless d=2, and ARIMA(0,d,0)(0,d,0) without a constant is fitted when d≤1. • ARIMA(0,d,0)(0,d,0) • ARIMA(2,d,2)(2,d,2) • ARIMA(1,d,0)(1,d,0) • ARIMA(0,d,1)(0,d,1)
3. The model with the minimal AICc in step 2 is considered as the "current model".

4. Hyndman-Khandakar algorithm varies p, P, q and Q from the current model by ±1, and also includes or excludes constant terms if there are any. If the AICc for the varied model is lower than the current model, the model will become the new current model.
5. Hyndman-Khandakar algorithm repeats step 4 and till the model with the lowest AICc is discovered.

Table 4. 1 Hyndman-Khandakar algorithm for automatic ARIMA modelling (Hyndman & Athanasopoulos, 2020)

The ARIMA model can deal with the time series data with seasonality and is capable of including other covariates in the model (Hyndman & Athanasopoulos, 2020). According to Section 3, strong daily seasonality is found in the bike demand in Bergen City Bike. Additionally, weather factors and weekend dummy should be included in the model. There is a difference in rental numbers over seasons, as demonstrated in Figure 3.4, implying weather conditions might affect bike demand. Some previous studies also pointed out that the weather conditions could be significant independent variables to predict the bike demand, as mentioned in Section 2.1. Moreover, Figures 3.3 and 3.4 illustrate that patterns in weekdays and weekends might be different. Therefore, weekend dummy variables need to be considered in the demand forecasting model. Moreover, some previous studies use the ARIMA model to forecast bike demand, as mentioned in Section 2.1. Consequently, with the need of capturing seasonality and including other covariates, we choose the ARIMA model to develop the station-level demand forecasting model.

4.2.2 The Random Forest Model

The random forest takes many training sets from the population, trains the model by a subset of predictors separately, and averages the prediction result. It improves bagged trees, which trains the model by all predictors, given that it decorrelates the prediction and leads to a smaller variance of the average predictions (James et al., 2013).

The random forest method is also able to address data with lots of covariates, and many previous studies adopted it for forecasting the bike demand, as mentioned in Section 2.1. Moreover, Yin et al. (2012) applied ridge linear regression, SVR (Support Vector Regression), random forest, and gradient boosted trees methods to the demand predictions. The study indicated that random forest is the most promising method among the machine learning methods they tested. Thus, the random forest model is also selected to develop station-level demand forecasting models in our thesis.

4.2.3 Model validation

Residual diagnostics is widely implemented to validate time series analysis methods, as mentioned in Hyndman & Athanasopoulos (2020). The ARIMA method is a time series analysis technique, so residual diagnostics is adopted in our thesis. The residual diagnostics mainly consists of two parts: checking whether the residuals have zero mean and observing whether it is uncorrelated by adopting the Ljung-Box test. Residuals with zero mean are necessary criteria of an unbiased forecasting result. Non-zero mean residuals can be adjusted by simply adding the mean of residuals in all forecasts (Hyndman & Athanasopoulos, 2020). Therefore, the mean for residuals in forecasting each station would be added to our point forecast to overcome the bias problem. On the other hand, auto-correlated residuals do not necessarily indicate that the forecasting result is biased. Autocorrelated residuals rather show that there is still some information in the data not being used, and the model can be improved (Hyndman & Athanasopoulos, 2020).

Additionally to residual diagnostics, the time-series cross-validation is also commonly used to validate the time series model, such as in Bergmeir et al. (2012). As displayed in Figure 4.2, the time-series cross-validation firstly splits the data into a training data set and test data set. The training data set is used to train the model, and the test data set is used to measure the forecast error of the trained model for performance validation. By performing this process several times, the average forecast error can be obtained.

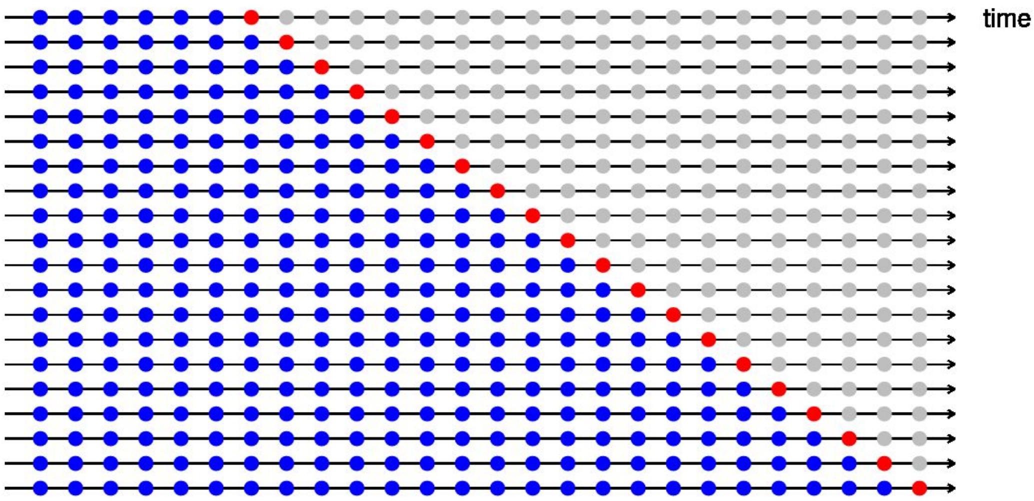


Figure 4. 2 Time series cross validation. The blue spot represents the training data set and the red spot represents the test data set. (Source: Hyndman & Athanasopoulos, 2020)

The forecast error adopted in this thesis is the mean squared error (hereafter MSE). Unlike mean absolute deviation (hereafter MAD) and mean absolute percentage error (hereafter MAPE), MSE emphasizes the extreme errors in the model (Chopra et al., 2013). Although MSE is scale-dependent, we develop the rolling basis model based on a short term data set and the scale of bike pickups and returns' time series varies little over the short term. Hence, MSE is a valid measurement to represent the forecast error in the thesis.

Normally, the lower the forecast error of the model, the more accurate the model is. However, we comprehensively consider the seasonalities' correctness captured by the model and forecast error to determine the model with the best performance in this thesis.

In general, the lengths of training data sets are different when implementing time-series cross-validation, according to Figure 4.2. However, as mentioned in Section 4.2, the station-level demand forecasting model is developed on a rolling basis by using a certain length of hourly data, so we use fixed-length training data set before the test date of interest when performing time-series cross validation to achieve developing a model suitable for real-life application.

In conclusion, first, we check if the regression model with the ARIMA model passes residual diagnostics. Subsequently, we select the best-performed model with lower MSE according to the time-series cross validation and with the correct seasonality.

4.3 The Target Number of Bikes to be Repositioned

As mentioned in Section 4.1, a similar framework of the repositioning model was also adopted in Regue & Recker (2014), but that study does not include real-time station data, and proposes a model to forecast the bike inventory level. Considering the data availability of real-time state for each station (inventory level) for Bergen's BSS, we develop the model with the consideration of real-time data.

The acronyms in Figure 4.3 are explained as follows. The acronym HP represents the historical number of pickups, HR represents the historical number of returns, HW stands for historical weather data, FW represents the weather forecast, and W stands for the weekend dummy. FD, FP, and FR denote the station-level point forecast of the net demand, the number of pickups, and the number of returns, respectively. TA, TR, RAB, and RAS stand for the target number of bikes to be added, the targeted number of bikes to be removed, the real-time number of available bikes, and the real-time number of available slots, respectively.

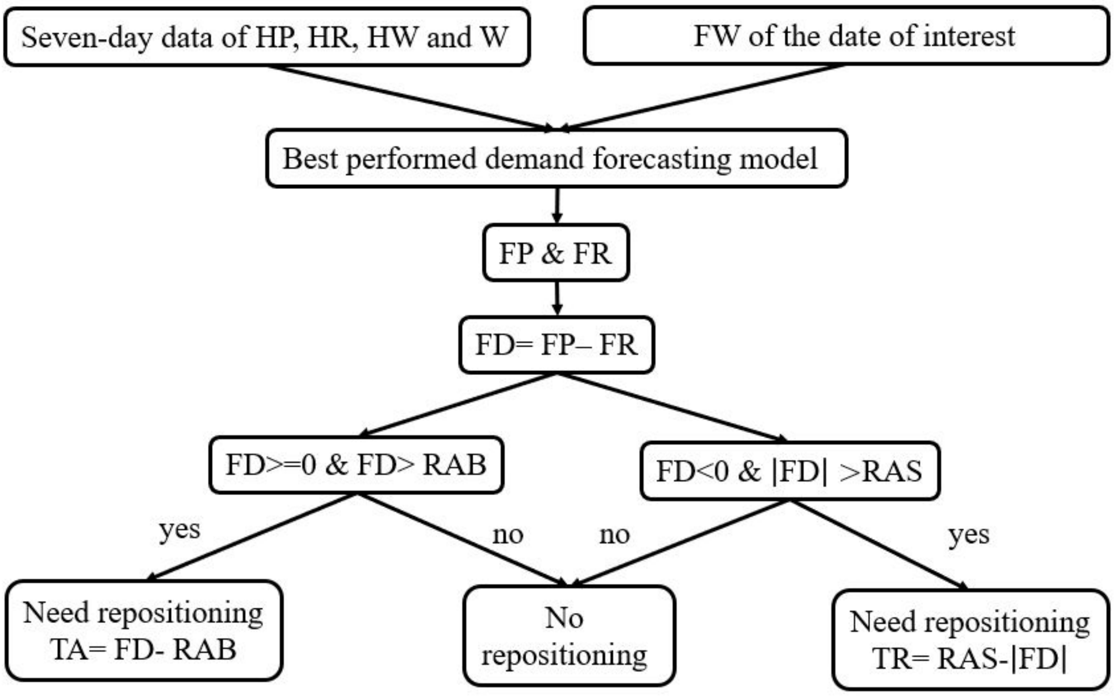


Figure 4. 3 The process of determining the target number of bikes to be repositioned for each bike station.

Figure 4.3 demonstrates the process of determining the target number of bikes to be added (hereafter TA) or removed (hereafter TR) for each station. First, historical bike pickups and returns data of a certain time length before the date of interest, weekend dummy variable, and the weather forecast data of the date of interest are obtained to train the station-level pickups and returns forecasting model. After training the model by using the best-performing method selected previously and getting the result of point forecast of bike pickups and returns (hereafter FP and FR) of every hour in each station, the forecasted net demand (hereafter FD), is computed by subtracting FR from FP. If the bike station's FD is positive and greater than the real-time number of available bikes (hereafter RAB) in that station, or if the FD is negative and its absolute value is greater than the real-time number of available slots (hereafter RAS), those bike stations are selected to be repositioned. The difference between the FD and RAB or RAS is the TA or TR for each station. Otherwise, no repositioning is needed for those specific bike stations that do not meet the criteria mentioned above.

4.4 Repositioning Route Optimization Model

A Mixed-Integer Nonlinear Programming model is developed considering the transportation cost and the unsatisfied demand cost in the objective function. Also, we include restrictions of vehicle and bike stations' capacity to address the case of Bergen's BSS. Moreover, the station-level number of bikes to be repositioned (TA and TR), derived from the process mentioned in Section 4.3, is the parameter Ts in the repositioning route optimization model. TA represents the positive values of Ts, and TR stands for the negative values of that.

There are two main discussions for the model of bikes repositioning. One is the adoption of a deterministic or stochastic model. The deterministic model considers bike demand parameters as constant. On the other hand, a stochastic model considers bike demand uncertainty and fits a distribution for bike demand as parameters in the model. The other discussion is whether to employ a static or dynamic model. A static model assumes the bikes do not move during repositioning; it is opposite for a dynamic model (Gleditsch & Hagen, 2018).

A deterministic model is adopted in our thesis. This choice contradicts with the data variation explored in Section 3. However, handling the stochasticity could be done by improving forecasting of the demand (Chopra et al., 2013), and a more complex model is used to forecast the station-level bike demand in this thesis, as mentioned in Section 4.2. Therefore, the uncertainty of demand is addressed, and the value of adopting a stochastic model is reduced. Therefore, a deterministic model is implemented in this thesis.

Furthermore, according to Figure 1.3, the most popular bike stations and the depot are located in the central area of Bergen. Therefore, Bergen City Bike workers do not need to spend a long time on bike repositioning, and few bikes are moving for a short period of repositioning. Consequently, a static model is conducted by assuming few moving bikes during short repositioning time and therefore these movements are neglectable.

Although some other articles assume that capacity of a vehicle is unlimited, there is only one vehicle performing repositioning for Bergen City Bike. Therefore, considering the vehicle capacity is important in this thesis. Wang & Szeto (2018) propose a Mixed-Integer Linear Programming model to minimize carbon dioxide emission. The proposed model tracks the number of bikes on the vehicle in each stop. The ability of tracking the number of bikes on the vehicle enables to set a capacity constraint in the optimization model to ensure the number of bikes the vehicle carries will not exceed its capacity. However, the repositioning model in

Wang & Szeto (2018) does not consider each bike station's capacity and the effect of unsatisfied demand. Therefore, we formulate the Mixed-Integer Nonlinear Programming model that considers the capacity of bike stations and the effect of unsatisfied demand in Section 5.3, referring to the model proposed by Wang & Szeto (2018).

4.5 Data Input

The historical number of pickups and returns data and real-time station data, including stations' locations, stations' capacity, real-time number of available bikes and slots in each station, are obtained from the Bergen City Bike open data website (2020). Historical hourly weather data applied to train the demand forecasting model and weather forecast data employed in forecasting are obtained from yr.no website (2020) by web crawling.

As mentioned in Section 1.4, it is not possible for Bergen City Bike to know real historical demand since there is no way to record unsatisfied and lost demand. Given the lost demand is usually minor (Hulot et al., 2018), historical numbers of pickups and returns data are used to forecast the station-level demand for the repositioning model.

4.6 Tools to be Used

In this section, tools to be used in the thesis will be introduced.

As mentioned in Section 4.5, historical hourly weather data needs to be obtained from yr.no website. The historical hourly weather data is separated into different web pages and need to be obtained by searching day by day. Rvest package in R enables to perform web crawling on yr.no website (2020) and to store the historical hourly weather data at once without manually searching the data for each date.

One of the main crucial parts of this thesis is modeling the demand. As introduced in Section 4.2.1, the ARIMA and random forest model are implemented to train the demand forecasting model. Forecast package in R is applied in previous studies when conducting ARIMA modelings (Alghamdi et al. 2019; Chujai et al. 2013). Therefore, R's forecast package is implemented in this thesis for developing the regression model with ARIMA errors. Auto.arima() function in the forecast package returns the best ARIMA model with the lowest corrected Akaike information criterion (hereafter AICc) based on the Hyndman-Khandakar algorithm, mentioned in Section 4.2.1. The randomForest package in R is conducted to develop the random forest model in the thesis, and it is also employed in the research of Chen et al. (2018).

AMPL and the BONMIN solver are used to develop a static deterministic model for the repositioning route optimization model. AMPL is widely utilized in different optimization models in bike-sharing research. Maggioni et al. (2019) employed AMPL to formulate and solve the model for a bike-sharing problem with transshipment. The BONMIN solver was also applied to solve Mixed-Integer Nonlinear Programmings proposed by Lukáš & Branda (2016).

Finally, to obtain the driving distance between each pair of two stations as the parameter in the repositioning route optimization model, the R’s gmapsdistance package is utilized. This package computes the shortest driving distance between two places based on Google map and is also used in the research of Heaney et al. (2019).

5. Proposed Model

5.1 The Demand Forecasting Model

5.1.1 The Regression Model with ARIMA Errors and The Random Forest Model

First, the time length of the data used to train the rolling-basis model should be decided. In the work of Froehlich et al. (2009), five-day data before the date of interest was used. The results show that the more observations used, the lower the prediction error is, and five-day data is enough to develop a forecasting model (Froehlich et al., 2009). In our thesis, seven-day (168 hours) hourly data before the date of interest is applied to train the model, given that seven-day hourly data is sufficient to include the daily pattern and the weekly pattern, which are major patterns for training the model found in Section 3. Also, the ARIMA model cannot deal with time series with too complex seasonality. Hence, we shorten the training data set for training the model to seven days due to the ARIMA model's limitation. Therefore, we conclude that the seven-day period data for training the model is suitable for this case.

Subsequently, we first apply the data from May 1 to May 7, 2019, to train the model and forecast the station-level net demand on May 8, 2019, by employing the ARIMA model and the random forest model. Although there were 99 bike stations in Bergen in 2019, not all of them were frequently used. For computational feasibility, the demand of the stations whose average daily demand in 2019 was more than its capacity, is forecasted in this thesis, which assumes that if the demand of these stations is high, the need for repositioning will be relatively high as well. The data exploration indicates that 20 out of 99 stations in Bergen are frequently used. Thus, the station-level net demand of those stations is forecasted in the thesis.

When applying the ARIMA model, some other techniques are implemented regarding the model and data characteristics. For the ARIMA model, the forecasting result might be negative. Therefore, log form is employed for every variable in the ARIMA model to ensure the forecasting result to be equal or greater than zero. Also, there are some extreme values either for the dependent variable and the independent variables. Applying log form can adjust the extreme values. However, there are zeros in the time series data for hourly bike pickups and returns, but log 0 is not valid. Consequently, 0.001 is added to every data in the number of pickups and returns data to overcome this problem.

The independent variables used to train the model are discussed. As mentioned in Section 3, there is a significant difference in the number of bike rentals between weekdays and weekends. Therefore, the weekend dummy variable is used for distinguishing weekdays and weekends. Weather conditions are also relevant to bike pickups and returns, as demonstrated in Section 3. Thomas et al. (2009) highlight the effect of weather on net demand. Consequently, we include independent variables available on yr.no website, which are average temperature, wind speed, and precipitation.

The model derived from the ARIMA model for forecasting station-level net demand on May 8, 2019, is displayed in Appendix A. Figure 5.1 demonstrates the point forecast of net demand with consideration of the mean value of residuals adjustment. As mentioned in Section 4.3, each station's net demand is calculated by subtracting the point forecast of the number of returns from the point forecast of the number of pickups, with consideration of biased mean adjustment.

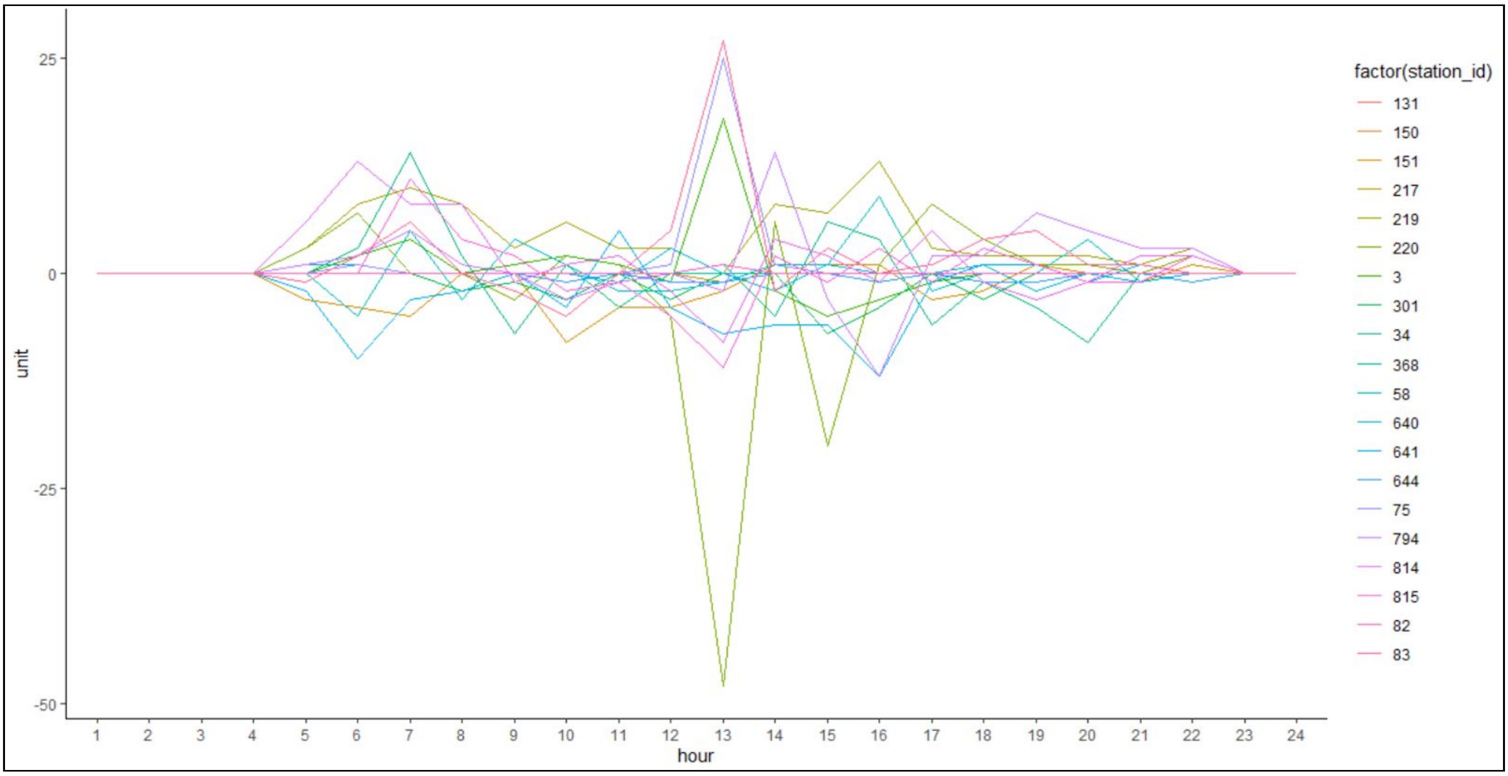


Figure 5. 1 The point forecast of station-level net demand on May 8, 2019, from the ARIMA model

According to Figure 5.1, there are evident peaks at the 6-7 AM and 13-15 PM for several stations. It means that the point forecast of net demand is consistent with what we have observed in Section 3: there are two peaks at around 6-7 AM and 13-15 PM. The MSE of testing data is 788.92. It means that, on average, there are around 28.09 unit differences between actual data and the predicted value in total per hour. There are around 1.40 units difference between the actual value and the predicted value on average for each bike station in each hour.

In addition to the ARIMA model, we also implement the random forests model to forecast the station-level net demand for the same date. For the sake of comparability, the training dataset and independent variables are the same as those applied in the ARIMA model. However, the random forests model cannot consider the daily and seasonality by default, so dummy variables are applied for every weekday, weekend, and hour to overcome this problem.

Besides the variables to be included in the random forest model, the number of forests used to train the model is also discussed for the random forest model. A system-level demand forecasting model is developed using the random forest method beforehand. The most suitable number of trees is considered in the benchmark of developing a station-level model. The forecast error gets stable after the number of trees reaches 30. Therefore, the number of forests is set as 30 to train the station-level pickups and returns model.

The point forecast of station-level net demand from the random forest model is illustrated in Figure 5.2. As mentioned in Section 4.3, the point forecast of station-level net demand is calculated by subtracting the point forecast of the number of returns and the point forecast of the number of pickups.

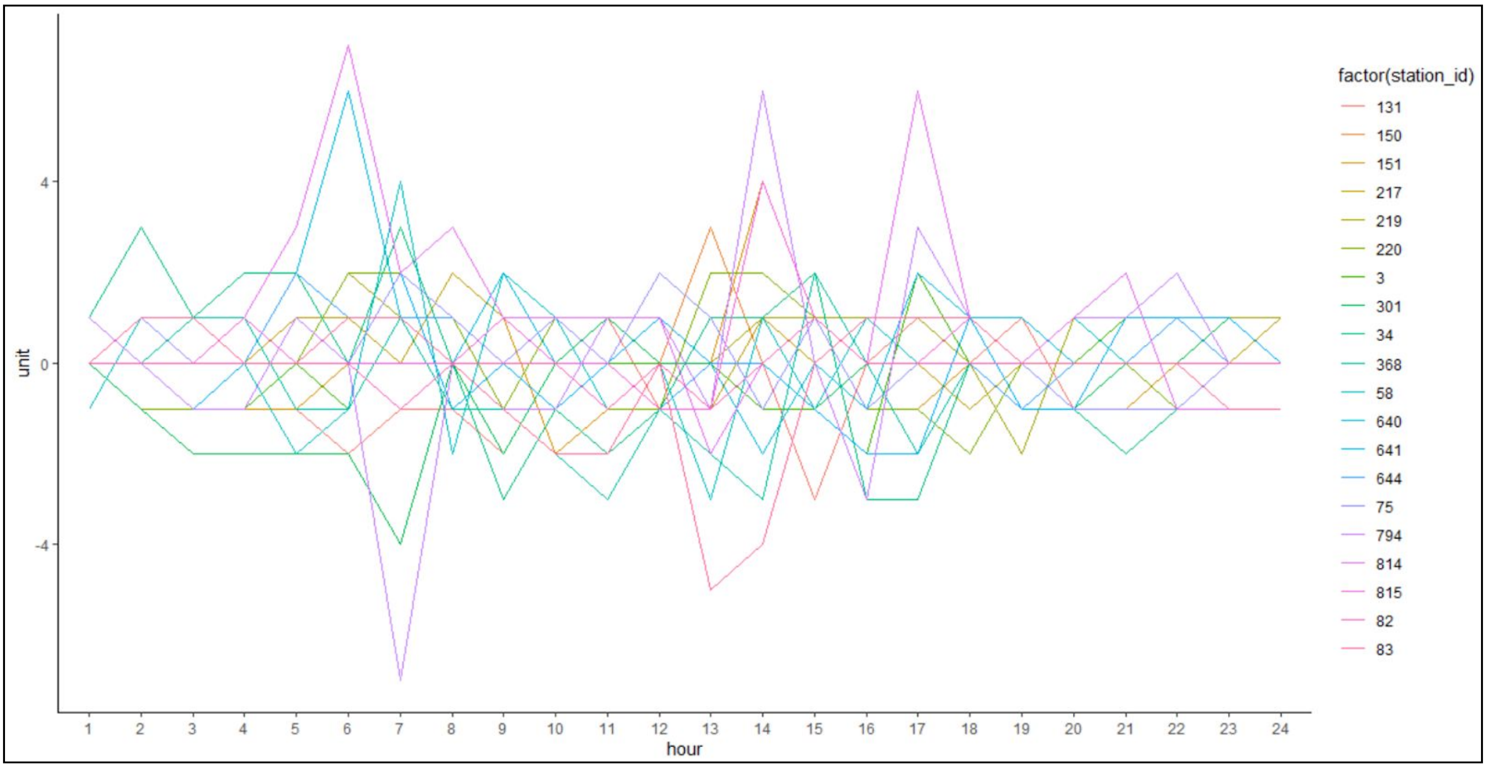


Figure 5. 2 The point forecast of station-level net demand on May 8, 2019, from the random forest model

According to Figure 5.2, for several stations, there are evident peaks at the 6-7 AM and 13-15 PM, which shows the consistency in peak hours with what we have observed in Section 3 and the point forecast result from the ARIMA model. However, the random forest model's point forecast displays incorrect forecasting in the non-operating hour: it forecasts there is demand at midnight. The MSE of testing data is 485.54. It means that, on average, there are around 22.03 unit differences between actual data and the point forecast per hour. For each bike station in each hour, there is around 1.10 units difference between actual data and point forecast on average.

5.1.2 Model Validation

Residual diagnostics are performed for the ARIMA model. The mean value of residuals is added to the point forecast, as mentioned in Section 4.2.2, to prevent biased results. Therefore, the point forecast in the thesis has met the criteria of passing residual diagnostics. Additionally, the Ljung-Box test is performed to see whether the correlations in the residual series are statistically significant. The p-value of the Ljung-Box test is demonstrated in Figure 5.3.

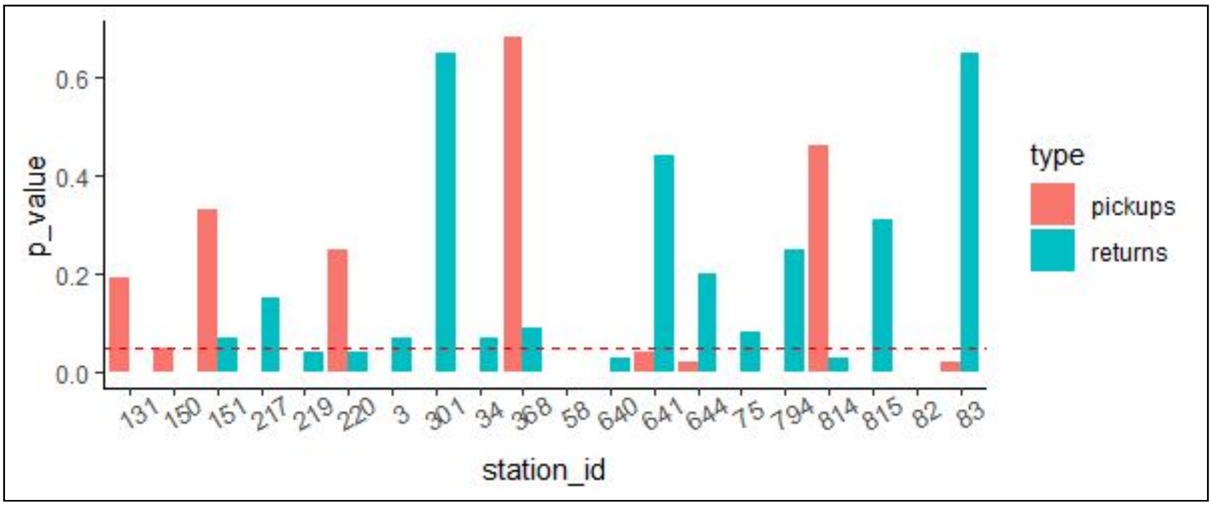


Figure 5. 3 P-value of the Ljung-Box test for each station. The red line presents the 5% significance level.

Figure 5.3 demonstrates that many p-values are greater than the 5% significance level, which means that most of the information in the data has been collected. The ones whose p-values are lower than the 5% significance level do not indicate the point forecast is biased but rather suggest that the model can be improved.

Other than residual diagnostics, the time-series cross validation is also implemented to validate both the ARIMA and the random forest model. Seven-day (168 hours) training data set and one-day (24 hours) test data set are implemented in time-series cross-validation in our thesis. Seven-day data includes the daily pattern and the weekly pattern in the bike demand, as mentioned in Section 4.2, so seven-day data is the appropriate time length for the training data set. Moreover, repositioning is performed several times daily, which implies the station-level demand forecasting needs to be performed more than once a day, and the forecasting time length is shorter than one day in this case. Consequently, the test data set's length is set to be one day because it is the most extended forecasting time length of interest for every repositioning.

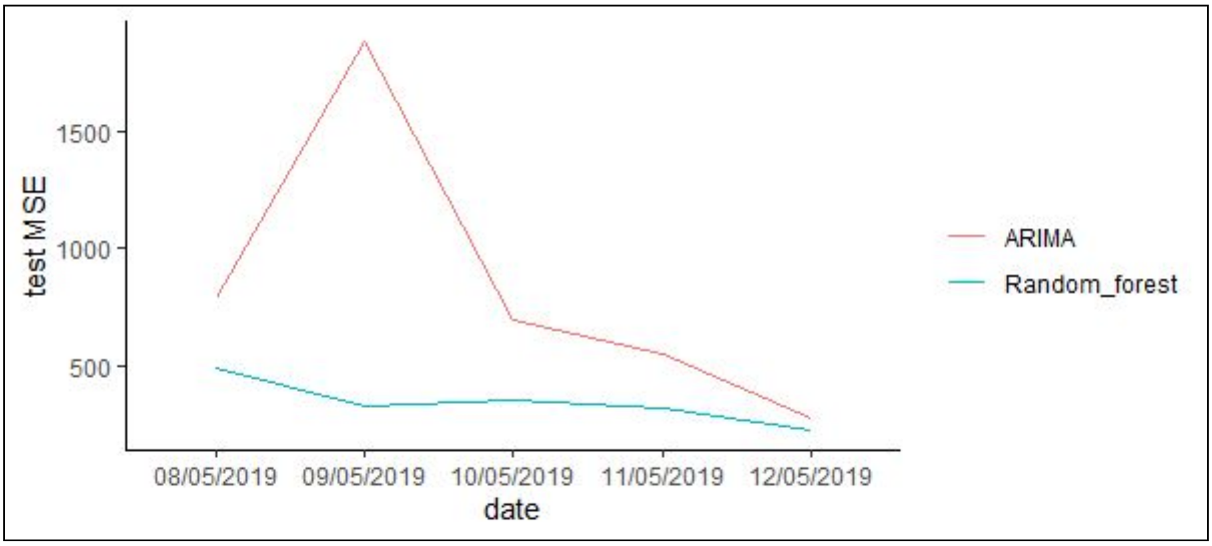


Figure 5. 4 The test MSE for the regression with ARIMA errors and the random forest model

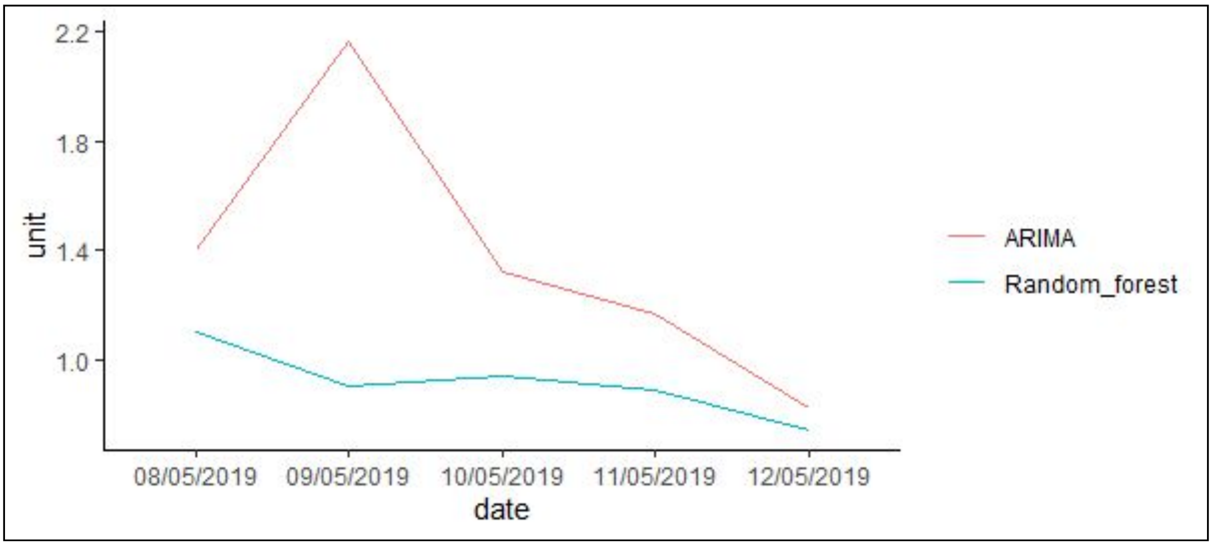


Figure 5. 5 The average forecast error per hour and station for the ARIMA and the random forest model

Figure 5.5 demonstrates that both models' average forecast error is minor because all are smaller than 2.2 units of bikes per station each hour. Therefore, both models are applicable in forecasting station-level demand.

Figure 5.4 and 5.5 demonstrate that in terms of lower test MSE and average forecast error per hour and station, the random forest model performs better than the ARIMA model. However, Figure 5.2 indicates that the random forest model captures wrong seasonality during the non-operating time: it forecasts that there is bike demand at midnight. Therefore, we conclude that the ARIMA model is the most accurate model in this thesis, given it passes the residual diagnostics, has low average forecast error per station and hour according to time-series cross-validation, and captures seasonality more correctly than the random forest model.

5.2 The Target Number of Bikes to be Repositioned

We first estimate the target number to be repositioned in the first peak hour on May 8, 2019, which is 6-7 AM in this case, because the users emphasize the importance of availability at peak hours, as a significant share of users rely on BSS as a standard commuting option (Hughes, 2017).

We apply the point forecast from the best-performed model, which is FD derived from the ARIMA model in this thesis, to determine the target number of bikes to be repositioned at a specific station by going through the process mentioned in Section 4.3. According to Section 4.3, the real-time available bikes and slots information, RAB and RAS, can be obtained from Bergen City Bike's website. However, the corresponding real-time data for the sample we used in the thesis was not stored, and, therefore, not available. Therefore, we set the RAB and RAS to be the average number of bikes or slots at 10 PM on October 23, 2020 instead, which are three bikes and seventeen slots per station in this case.

After going through the process mentioned in Section 4.3, the result is set to be Ts in the repositioning route optimization model in Section 5.3. The positive value represents the target number of bikes to be added, and the negative value indicates the number of bikes to be removed.

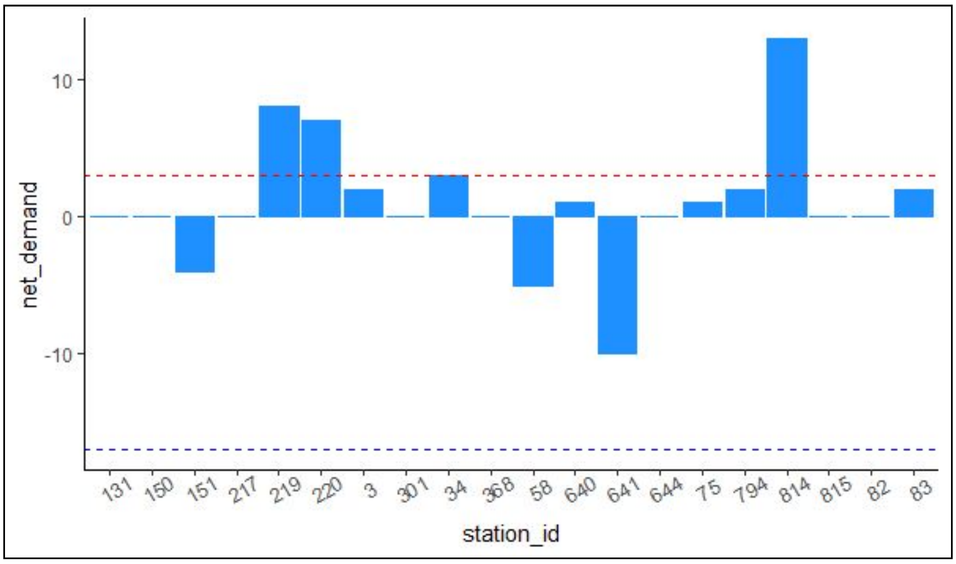


Figure 5. 6 The stations need to be repositioned. The red line presents the number of available bikes (RAB), and the blue line stands for the number of available slots (RAS).

Figure 5.6 shows that only three stations need to be repositioned: five, four, and ten bikes need to be added for station 219, station 220, and station 814, respectively.

5.3 Repositioning Route Optimization Model

The constraints (1) to (7) are referred to Wang & Szeto (2018). As mentioned in Section 4.4, the model presented by Wang and Szeto (2018) records the number of bikes in a vehicle at each stop. However, the repositioning model in Wang & Szeto (2018) does not consider each bike station's capacity and the effect of unsatisfied demand. Therefore, we formulate the model considering the capacity of bike stations and the effect of unsatisfied demand in the following passage.

There are only one vehicle and one depot available for transporting bikes in Bergen. Therefore, the model above is developed based on assuming that there are only one vehicle and one depot in the BSS.

Sets

S : the set of stations

S_0 : the set of stations and the depot

A: the set of stops. 1, 2,..., N

Variables

x_{sa} : equals to 1 if station s is the ath stop for the vehicle, 0 otherwise. Binary variable.

y_{sa} : the number of bikes added or removed at station s at ath stop. Integer.

n_a : the number of bikes on the vehicle after visiting a specific station at ath stop. Non-negative integer.

i_{sa} : the number of bikes at station s at ath stop. Non-negative integer.

b_s : total number of bikes added or removed at station s. Integer.

Parameters

P_s : The initial inventory of bikes at station s, $s \in S$

T_s : The target number of bikes added or removed at station s, $s \in S$

D_{sj} : The travel distance from station s to station j on Google Map, $s,j \in S_0$

Q : The capacity of the vehicle

C_s : The capacity of station s, $s \in S$

U : The cost per unit of unsatisfied net demand

K : The cost per meter for the vehicle

Objective function:

Minimize the transportation and the unsatisfied demand cost

$$\sum_{s \in S_0} \sum_{j \in S_0} \sum_{a=2}^{N} K\, D_{sj}\, x_{s,a-1}\, x_{j,a} + \sum_{s \in S} U(T_s - b_s)^2 \qquad [1]$$

Constraints:

(1) The vehicle can only visit one station at one stop:

$$\sum_{s \in S_0} x_{sa} = 1\ ,\ a = 1, \ldots, N \qquad [2]$$

(2) The first stop must be the depot (station 0):

$$x_{01} = 1 \qquad [3]$$

(3) The last stop must be the depot (station 0):

$$x_{0N} = 1 \qquad [4]$$

(4) No consecutive stops at the same station:

$$x_{sa} + x_{s,a+1} \le 1\ ,\ \forall s \in S_0,\ a = 1, \ldots, N-1 \qquad [5]$$

(5) No station will have removed/ added bikes when no vehicle visiting:

$$-Qx_{sa} \le y_{sa} \le Qx_{sa},\ \forall s \in S_0,\ a = 1, \ldots, N \qquad [6]$$

(6) Balance of the number of bikes on the vehicle:

$$n_a = n_{a-1} - \sum_{s \in S_0} y_{sa},\ a = 2, \ldots, N \qquad [7]$$

(7) Initial number of bikes on the vehicle:

$$n_1 = y_{01} \qquad [8]$$

(8) Balance of the number of bikes at a specific station:

$$i_{sa} = i_{s,\, a-1} + y_{sa},\ \ \forall s \in S_0, a = 2, \ldots,\ N \qquad [9]$$

(9) Initial number of bikes of each station:

$$i_{s1} = P_s\ ,\ \ \forall s \in S_0 \qquad [10]$$

(10) Total number of bikes to be repositioned:

$$b_s = \sum_{a \in A} y_{sa},\ \ \forall s \in S_0 \qquad [11]$$

(11) Vehicle capacity:

$$0 \le n_a\ \le Q,\ \ \forall a = 1, \ldots,\ N \qquad [12]$$

(12) Station capacity:

$$0 \le i_{sa} \le c_s,\ \ \forall s \in S_0,\ a = 1, \ldots,\ N \qquad [13]$$

(13) If the sum of the target number to be repositioned is greater or equal to 0, there will be no returned bikes at the last stop:

$$y_{0N} = 0 \qquad [14]$$

Otherwise, returned bikes are allowed for the last stop:

$$y_{0N} \ge\ 0 \qquad [15]$$

(14) Variable nature:

$$x_{sa}\ \in\ \{0, 1\} \qquad [16]$$

(15) Variable nature:

$$b_s, y_{sa}\ \in\ integer \qquad [17]$$

(16) Variable nature:

$$n_a, i_{sa} \in \textit{nonnegative integer} \qquad [18]$$

The objective function is composed of two parts. The first part is the transportation cost of the vehicle when traveling from one station to another station. The cost depends on the total distance the vehicle travels. When the vehicle travels a meter, the cost will increase by K NOK. The other part of the objective function is the penalty for unsatisfied demand. If the total number of bikes to be repositioned at the stations does not fulfill the target number, then each square of unsatisfied units will get punished by U NOK. Considering the vehicle capacity constraint, we square the unsatisfied demand to make the vehicle visit the stations with higher demand first. For example, if there are three stations that need to be repositioned and that the number of bikes need to be repositioned are (10,8,8) for each station, and if all the stations are within equal travel distance, the vehicle will first visit the first station to add 10, and then will try to satisfy the demand of other stations in this case, given that the vehicle constraint is 20. It also emphasizes that we pay more attention to stations with higher demand, which is close to reality: the stations with higher demand indicate that they are more popular, and it is more critical to satisfy the demand of popular stations rather than less popular ones. The purpose of including the penalty term in the objective function is that the penalty term punishes unsatisfied demand and can reduce the lost demand, which will not be recorded in the system. Consequently, the future rental data will be closer to actual demand, and the lost demand issue mentioned in Section 4.5 can be partially addressed.

Constraint (1) guarantees that there is only one station being visited in one stop. This constraint prevents the vehicle from appearing at different stations in the same stop in the optimal route. Constraint (2) and (3) guarantee that the first stop and the last stop of the vehicle should be the depot, which means the depot is always the start point and the final destination in the route. Constraint (4) ensure there will be no consecutive stops at the same station for the vehicle. For example, when the vehicle visits station 1 as its first stop, it cannot revisit station 1 right afterward. Constraint (5) ensures that no bikes are added or removed at a specific station when the vehicle does not visit that station. Constraint (6) tracks the number of bikes on the vehicle after a specific stop, and constraint (7) sets the number of bikes on the vehicle at the beginning to be the same as the number of bikes removed from the depot at the first stop. Constraint (8) tracks the number of bikes at each station at a specific stop, and constraint (9) sets the initial number of bikes at every stop equal to the real-time data.

Constraint (10) computes the total number of bikes added or removed after the completing repositioning. Constraint (11) and (12) are the capacity constraints for the vehicle and the stations. They restrict the number of bikes on the vehicle or at the station will be lower than their capacity. Without constraint (13), when the total target number of bikes to be repositioned is positive, the optimal solution will sometimes choose to extract more bikes than needed from the depot (the first stop) and then return the residual bikes to the depot (the last stop) at the end. It is not a reasonable choice to carry more bikes than required in reality. Therefore, constraint (13) ensures that if the sum of the targeted number of bikes to be repositioned is positive or equal to 0, then there will be no returned bikes to the depot at the last stop, and the unnecessary bikes would not be picked up. On the other hand, if the sum of the targeted number is negative (the bikes need to be removed from the stations to empty slots), returning bikes are allowed to the depot at the last stop. Constraints from (14) to (16) reflect the nature of each variable.

In the following passage, we explain the parameter setting in the thesis. All the parameters are set based on developing a model of the optimal repositioning route for the bike demand from 6 - 7 AM on May 8, 2019. Originally, P_s should be obtained by the real-time data on Bergen City Bike's website. However, the real-time data for 6-7 AM on May 8, 2019, is already missed. Therefore, we set P_s to be the average number of bikes in selected stations at 10 P< on October 23, 2020, instead, which is three bikes per station in this case. We set T_s as the target number of bikes to be repositioned, as mentioned in Section 5.2. D_{sj} is retrieved by implementing the gmapsdistance package. According to the managerial information of Bergen City Bike, there is only one vehicle utilized in Bergen, and the capacity of the vehicle is 20 bikes. Therefore, 20 units are set as Q in the repositioning model. The capacity of each station, C_s, is also obtained by the real-time data on the Bergen City Bike open source.

There are three types of subscription for Bergen's BSS, day, monthly, and annual passes, as mentioned in Section 1.3. Most of the users who buy monthly or annual passes possibly have a higher tendency to use Bergen's BSS. They could rely on BSS as a transport mode on a daily basis, that could mean that these users will still decide to buy a monthly pass or annual pass, even if they cannot find a slot or a bike when they need it. Therefore, we focus on the users who buy day passes, assuming these users' choices are more affected by service level. The day pass is NOK 49 for Bergen City Bike, so we assume when the users cannot find a slot or bike when they need it, they will not buy the day pass for one time afterward.

Consequently, NOK 49 is set as U, which is the unit cost of unsatisfied net demand in the repositioning model.

We first set the unit transportation cost per meter, the parameter K in the model, to be NOK 0.0014, which is the average transportation cost per meter derived from the data provided by Bergen City Bike.

6. Results Analysis

Section 6 Results analysis consists of five parts. Section 6.1 compares the demand forecasting model of the system-level demand and that of the station-level demand. Section 6.2 is the optimal solution for the first peak hour sample, from 6 AM to 7 AM, on May 8, 2019. The discussion of the change in total cost in the different unit transportation costs comes as Section 6.3. Section 6.4 discusses adopting an interval to decide the target number of bikes to be added or removed. In the last section, Section 6.5, we discuss the situations when adopting the partial objective function.

6.1 Comparison of the System-Level and the Station-Level Demand Forecasting Model

Station correlation factors affect each bike station demand (Lin et al., 2018), but when we forecast the system-level bike demand, the effects from these factors will offset each other and therefore are not relevant. However, we do not consider these factors in the station-level demand forecasting model in Section 5. To examine whether it is suitable not to consider these factors in the model for Bergen's BSS, we develop the model for the system-level net demand forecasting by using the ARIMA model and compare the result from the system-level demand forecasting model and that of the station-level demand forecasting model. If the difference is small, considering these factors when developing the station-level net demand forecasting model is irrelevant.

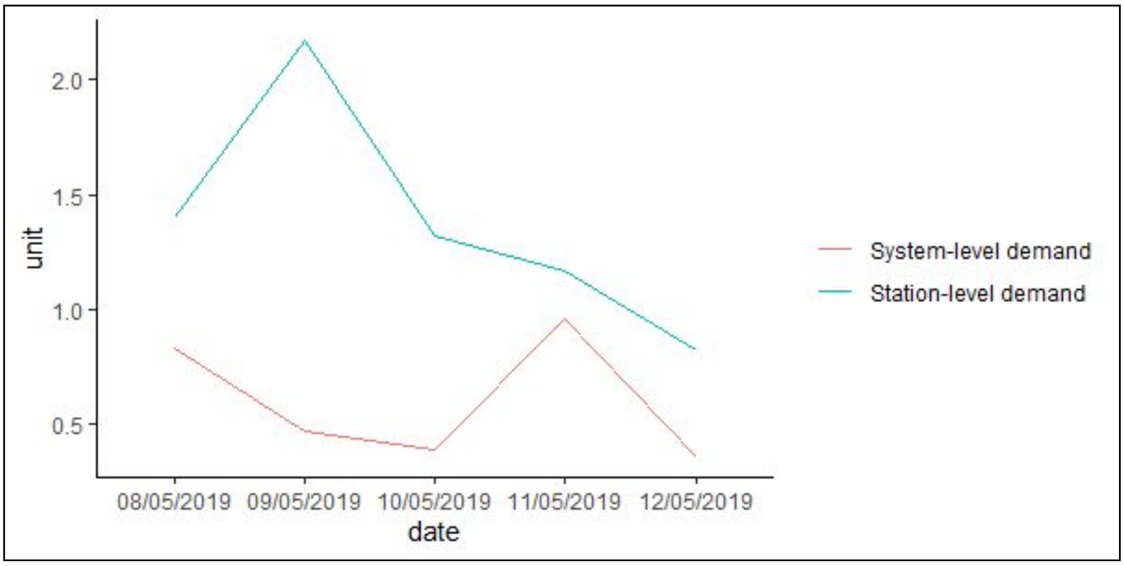


Figure 6. 1 Forecast error per hour and station for system-level demand forecasting model and station-level demand forecasting model

We further compute the test MSE of two models for five days, from May 8 to May 12, 2019. Figure 6.1 displays that the average forecast error per hour and station for the system-level demand forecasting model is always lower than that for the station-level net demand forecasting model in all test periods. The test MSE of the station-level net demand forecasting model is 837.77 on average, and the average forecast error is 1.38 units per station and hour. On the other hand, the test MSE of the system-level demand forecasting model for 99 stations is 4148.78, which indicates the average forecast error is 0.6 units per station and hour. The forecast error from the system-level demand forecasting model is slightly lower than that of the station-level demand forecasting model by 0.78 units. This observation meets our expectation that the forecast error for system-level demand forecasting model is lower than that for station-level demand forecasting model. However, the average difference is lower than one bike. Therefore, we conclude that the effect of considering station correlation factors is irrelevant for BSS in Bergen.

6.2 The Optimal Solution

We first obtain the optimal solution in the first peak hour on May 8, 2019, which is 6-7 AM, because the customers emphasize the importance of availability at peak hours (Hughes, 2017), as mentioned in Section 5.2.

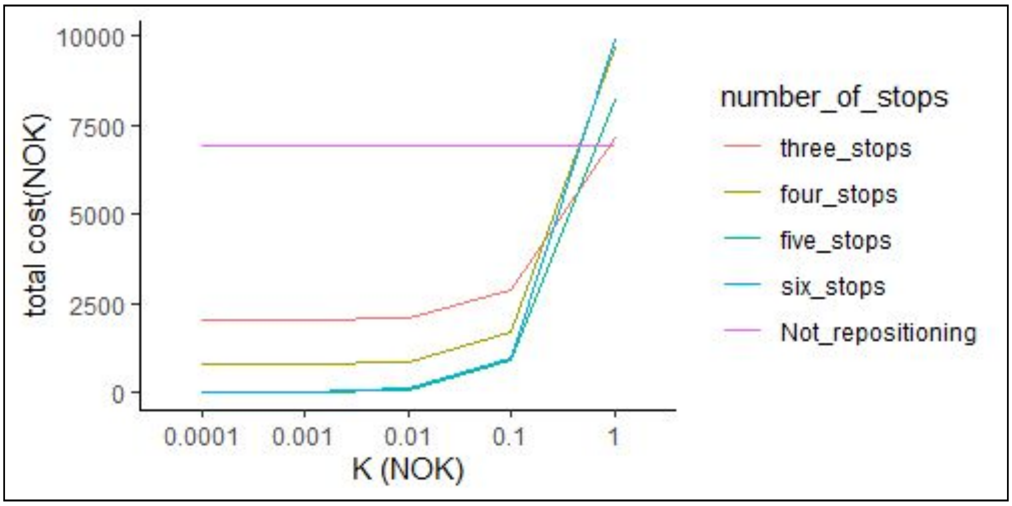
We first set the number of stops as five stops because only three stations needed to be repositioned, and the total number of bikes to be added is 19 bikes. It is lower than the vehicle's capacity. The optimal route is as follows:

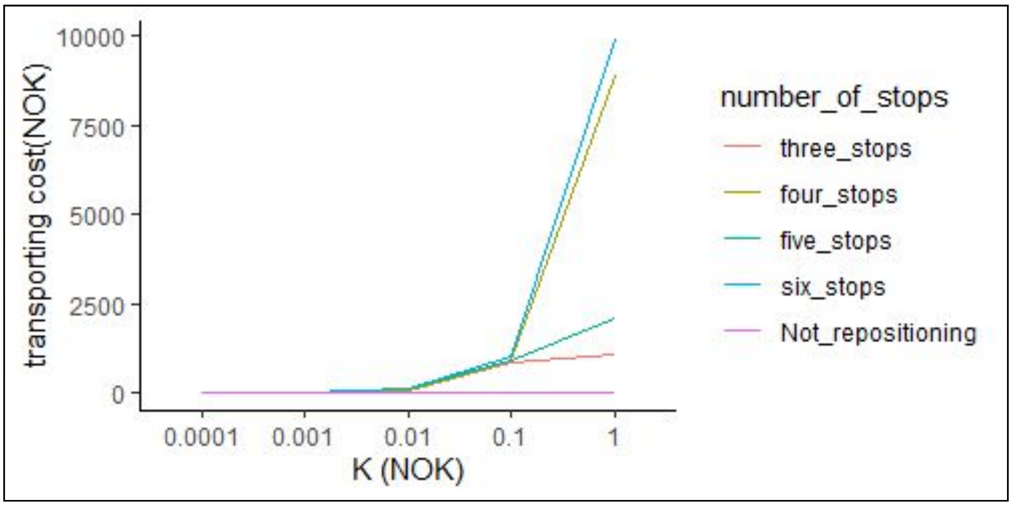
$$0\ (-19) \rightarrow 220(+4) \rightarrow 814(+10) \rightarrow 219(+5) \rightarrow 0(0)$$

Firstly, the vehicle picks up 19 bikes at the depot (station 0), travels to station 220 to add four bikes at station 220, moves on to station 814 to add ten bikes, travels to station 219 to add five bikes at that station, and returns to the depot at the end. Given the unit transportation cost is 0.0014 NOK per meter, the total transportation cost is 12.94 NOK, the unsatisfied demand is 0 NOK, and the total cost is 12.94 NOK. The optimal number to be added or removed from each station is the same as the target number we set.

6.3 Total Cost for Different Unit Transportation costs

To observe the total cost in different situations, we set the unit transportation cost per meter, the parameter K in the model, to be the actual unit transportation cost of NOK 0.0014, obtained from managerial information. Other parameters are set the same as the ones in Section 6.2. We run the model from 3 stops to 6 stops for every different set of parameters K. The total cost, cost of transporting, and cost of unsatisfied demand are illustrated as follows:





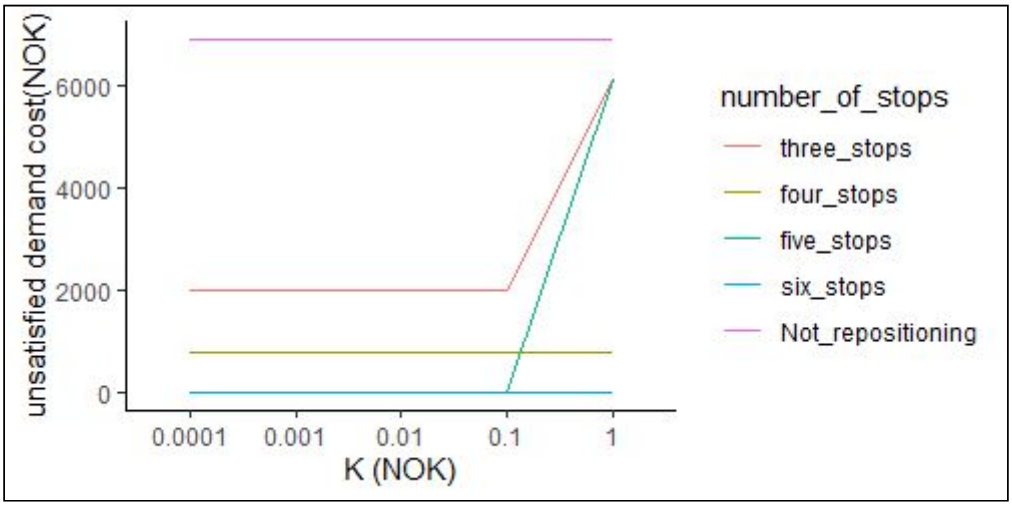


Figure 6. 2 The costs when there are three stops, four stops, five stops, and six stops

Intuitively, we might think having more stops would lead to less cost due to decrement in the unsatisfied demand cost, but the results show that intuition is wrong in this case. According to Figure 6.2, when the unit transportation cost per meter is low, the more stops, the better. When the transportation cost per meter is low, the share of unsatisfied demand cost will be relatively high and therefore affects the total cost more. Consequently, the optimal solution will tend to satisfy as much demand as possible. On the other hand, when the unit transportation cost is high, the share of unsatisfied demand cost will be relatively low. The optimal solution will tend to visit fewer stations and lead to higher unsatisfied demand cost.

Taking the result of the first peak hour on May 8, 2019, for example, according to Figure 6.2, we can see that the unit transportation cost's critical point is around NOK 1 per meter. Any unit transportation cost higher than the critical point will enhance the effect of travel distance in the model, and therefore, the optimal solution will be visiting fewer stations. Any cost smaller than the critical point will reduce travel distance's cost effect, and the optimal solution will be visiting more stations and satisfying as much demand as possible.

The total cost of not repositioning is NOK 6909 from 6 AM to 7 AM on May 8, 2019, which is also the unsatisfied demand cost. According to facet 1 in Figure 6.2, only when the unit transportation cost is lower than around NOK 1 per meter, the repositioning is less costly. Otherwise, if the unit transportation cost is more significant than around NOK 1 per meter, performing reposition will be more costly. The observation is consistent with the critical point mentioned above. Also, NOK 1 is relatively higher compared to the actual unit transportation cost, NOK 0.0014.

In conclusion, when the unit transportation cost is higher than the critical point, which is around NOK 1 per meter in this case, not repositioning is the best solution. On the other hand, if the unit transportation cost is lower than the critical point, more stops can lead to less unsatisfied demand, and the lowest number of stops that can meet all the demand can lead to the lowest total cost. For Bergen City Bike, the average unit transportation cost is 0.0014 NOK per meter. Therefore, as mentioned in Section 6.2, if traveling to five stops can fulfill all the unsatisfied demand, and it is also the lowest number of stops to meet all the demand, the analysis suggests that five stops are appropriate in this case.

6.4 Repositioning Based on Point Forecast and the Interval

In Section 6.4, we only discuss when the company follows the point forecasting result to perform repositioning. However, to cope with the uncertainty, the company might perform repositioning based on an interval, which means the company might want to add more bikes to each station as a safety inventory. In this case, having more bikes in each station can reduce the unsatisfied pickups. On the other hand, the company might prefer to remove more bikes or add fewer bikes from each station to reduce unsatisfied returns. We will discuss these two situations in the following passage based on the example from 6 AM to 7 AM on May 8, 2019.

First, how many additional bikes should be added or removed is discussed. As mentioned in Section 6.1, the average forecast error per hour and station is 1.38 units. Therefore, to offset this forecast error, we assume the company will perform the repositioning model based on two more bikes and two fewer bikes than the target number to be repositioned in Section 5.2.

Also, as mentioned in Section 6.2, the optimal number of stops is five. Consequently, we also discuss these two situations by starting from performing five stops repositioning.

Second, the targeted number to be repositioned when adding additional two bikes in each selected station is seven bikes, six bikes, and twelve bikes for stations 219, 220, and 814, respectively. The optimal solution from adding two more bikes is as follows:

$$0\ (-20) \rightarrow 220(+5) \rightarrow 814(+10) \rightarrow 219(+5) \rightarrow 0(0)$$

The result shows that the optimal solution is to remove 20 bikes at the depot (station 0), add five bikes at station 220, add ten bikes at station 814, add five bikes at station 219, and finally return to the depot at the end. The transportation cost is NOK 12.94, the unsatisfied demand cost is NOK 441, and the total cost is NOK 453.94. Compared to following the point forecast, the total cost increases due to unsatisfied demand cost. It is mainly because the vehicle's capacity constraint restricts the number to be added in each station in the optimal solution.

If we set the number of stops as six stops, the optimal solution is as follows:

$$0\ (-19) \rightarrow 220(+5) \rightarrow 0\ (-5) \rightarrow 814(+12) \rightarrow 219(+7) \rightarrow 0(0)$$

The result shows that the optimal solution is to carry nineteen bikes from the depot at first, add five at station 220, return to the depot to carry five more bikes, go to station 814 to add twelve bikes, move to station 219 to add seven bikes and return to the depot at the end. In this case, the target number to be added or removed for each station can be fulfilled by allowing the vehicle to travel more stations, the vehicle can firstly fulfill the closest station demand at the first place, which is station 220 in this case, and return to the depot to carry more bikes to meet the demand of the following stations needed to be repositioned. The transportation cost is NOK 13.90, with no unsatisfied demand cost, and the total cost is NOK 13.90. Compared to the result of 5 stops, the cost is significantly reduced. Also, compared to the result of following the point forecast mentioned in Section 6.2, where the total cost is NOK 12.94, the total cost of adding two additional bikes increases by NOK 0.96 due to longer travel distance.

Because Bergen City Bike has a low unit transportation cost, which is 0.0014 NOK, traveling to more stops can reduce the unsatisfied demand, which weighs more in the objective function. Simultaneously, the transportation cost will not increase drastically and will lead to a lower total cost.

Following the same process mentioned above, the optimal solution of adding two fewer bikes (or removing more bikes) is as below:

0 (-13)→ 220(+2)→814(+8)→219(+3)→0(0)

The result is similar to that of following the point forecast, and the only difference is that the vehicle adds two fewer bikes in each station. In this case, the target number to be added or removed to each station is still satisfied. The transportation cost is NOK 12.94, with no unsatisfied demand cost, and the total cost is NOK 12.94. The total cost is also the same as that of following the point forecast.

In conclusion, if Bergen City Bike wants to perform the strategy of adding more bikes or removing more bikes, and if the total bikes to be added or removed are more than the vehicle capacity constraint, traveling to more stops is more suitable. Because unsatisfied demand is more costly than transportation costs for Bergen City Bike, satisfying the demand is more critical.

6.5 Testing Different Parts of the Objective Function

As mentioned in the previous discussion in Section 6.3 and 6.4, for Bergen City Bike, the transportation cost is minor, and it is less critical than the unsatisfied demand cost. Therefore, it also implies that for some companies, the full objective function might not be useful. For example, like Bergen City Bike, only keeping the unsatisfied demand cost in the objective function might be enough. Consequently, we are going to test different parts of the objective function and discuss the results afterward.

Firstly, we keep the unsatisfied demand cost term in the objective function and perform the repositioning model based on the same set of parameters in Section 5.3. The revised objective function is shown as follows.

Minimize the unsatisfied demand cost:

$$\sum_{s \in S} U(T_s - b_s)^2 \qquad [19]$$

The result is as follows:

0 (-19)→ 814(+10)→220(+4)→219(+5)→0(0)

The unsatisfied demand cost is NOK 0, the implied transportation cost is NOK 23.10, and the total cost is NOK 23.10 in this case. Compared to the result in Section 6.2, the total cost is higher than that in adopting the complete objective function. In both cases, the unsatisfied demand cost is equal to NOK 0 because five stops are enough for the vehicle to transport all the bikes to meet the demand in each station. Therefore, the unsatisfied demand cost is equal to NOK 0 in both cases. The reason for the higher total cost is that ignoring the transportation cost leads to a longer route and almost twice the total cost. In Section 6.2, the vehicle visits the closest station to prevent the cost increment when visiting farther stations. On the other hand, the optimal solution of adopting the partial objective function here does not consider the travel distance.

Besides performing repositioning without considering the transportation cost, we also try to obtain the optimal solution when only considering the transportation cost. The revised objective function is shown below, and all the other parameter settings are the same as those in Section 5.3.

Minimize transportation cost:

$$\sum_{s \in S_0} \sum_{j \in S_0} \sum_{a=2}^{A} K\, D_{sj}\, x_{s,a-1}\, x_{j,a} \qquad [20]$$

The optimal solution is as follows:

0 (0)→ 220(0)→0(-9)→220(+9)→0(0)

Due to ignoring the unsatisfied demand cost term in the objective function, the optimal route shows that the vehicle goes back and forth between the depot and the closest station, which is station 220 in this case, to reach the lowest transportation cost. The unsatisfied demand cost is NOK 7350, the transportation cost is NOK 2.95, and the total cost is NOK 7352.95. The total cost is higher than that of the optimization route model when considering the full objective function by NOK 7340.

According to the discussion above, for Bergen City Bike, omitting the transportation cost term in the objective function makes the total cost increase twice, even for a single route, compared to adopting the full objective function. Therefore, the high increase percentage might indicate that all routes' total cost will increase significantly if Bergen City Bike chooses to ignore the transportation cost for all routes. Moreover, according to the previous passage's results, ignoring the unsatisfied demand cost seems unrealistic for Bergen City Bike. In conclusion, including both the transportation cost and the unsatisfied demand cost is better for Bergen City Bike.

6.6 Conclusion on the Results Analysis

First, the effect of station correlation factors is not significant for BSS in Bergen. Moreover, for Bergen's BSS, when the unit transportation cost is higher than a certain critical point, not repositioning is the best solution. On the other hand, if the unit transportation cost is lower than the critical point, more stops can lead to decreasing unsatisfied demand, and the lower number of stops that can meet all the demand leads to the lowest total cost.

If Bergen City Bike performs the strategy of adding more bikes or removing more bikes, and if the total bikes to be added or removed are more than the vehicle capacity constraint, traveling to more stops is more suitable due to the high unit unsatisfied demand cost.

For the repositioning route optimization model, including both the transportation cost and the unsatisfied demand cost is suggested for Bergen's BSS, because omitting either of them will lead to the higher total cost.

7. Limitation and Suggestions for Future Work

7.1 Limitation

In the thesis, given that all the models are developed on a laptop, it is computationally infeasible to obtain a long-term average result in this case. Moreover, a limited amount of data prevents us from reaching a particular conclusion.

Additionally, we only examine two models for the demand forecasting model, which are random forests and regressions with ARIMA errors, in this thesis. Many models can be used to forecast time-series data, such as the TBATS model, which is suitable for multi-seasonality time series data. However, these methods are not tested in this thesis. Therefore, we cannot conclude the ARIMA model performs the best among all the possible models. Thus, we can only state that the ARIMA model performs better than random forests for BSS in the Bergen case.

Furthermore, according to the p-value from the Ljung-Box test results in Figure 5. 3, demand forecasting models can be enhanced for some stations. Therefore, more factors can be considered in the ARIMA model to improve it.

Also, we use the same set of independent variables to develop the ARIMA models for different stations. However, some factors might affect the demand for a specific station, but they are not relevant to other stations. For example, if the bike station is near a bus stop, the bus schedule will also be relevant to that specific station. When the bus arrives at the stop, more potential users might rent bikes from the bike station nearby. On the other hand, if the bike station is not near the bus stops, whether the bus arrives at the bus stop might not affect that bike station's demand. Therefore, different sets of independent variables for each station can also be considered for future studies.

We set the penalty of unsatisfied demand equal for both pickups and returns regarding the repositioning route optimization model. However, in reality, there might be a different effect between not finding a slot to return the bike and not finding a bike when users need it. Furthermore, the penalty of unsatisfied demand is set as the same value for all stations. Nevertheless, the effect of unsatisfied demand might be different for different stations. For example, failing to return a bike to a remote station from other stations may have a more substantial effect because it is more difficult for the users to reach another station to park the bike. Likewise, failing to rent a bike from a specific station where there is no other public transportation nearby might also cause a more significant effect, compared to the stations located at the place where there are many other alternatives in terms of transportation modes. Therefore, the conclusion will be more complete if we can include these factors in the repositioning route optimization model. Unfortunately, we do not have related data to address these facts in this thesis.

Another concern of the repositioning route optimization model is the cost of transporting the bikes. We set the unit transportation cost to be the same in every meter the vehicle travels. Nonetheless, carrying different amounts of bikes will also affect the unit cost per meter of transportation. For example, carrying more bikes might make the vehicle to be more oil-consuming, and therefore might lead to a higher cost. However, we do not have related data to analyze the cost structure of unit transportation cost under different circumstances in the thesis.

In conclusion, most of the limitations mentioned above are either due to a lack of data or computationally infeasible problems.

7.2 Suggestions for Future Work

Based on the limitation mentioned in Section 7.1, there are some suggestions for future work.

First of all, if a more powerful computer is available, obtaining a long-term average result of the proposed models can be more solid proof to justify the argument of this thesis.

Secondly, more different types of models can be tested for forecasting the station-level net demand. Also, to improve the station-level net demand forecasting model, involving more data is essential. For example, for developing a more accurate demand forecasting model for every station, each bike station's geographical conditions should be considered.

Besides, more data should be retrieved to analyze whether the penalty of unsatisfied demand should be set as the same value for pickups and returns and for different stations. A customer survey can be conducted to retrieve relevant data. Also, more detailed data needs to be collected to identify a more sophisticated cost structure of the unit transportation cost.

Moreover, even considering that some studies point out that lost rentals are minor and, therefore, neglectable as mentioned in Section 2.2. The validity of this assumption for BSS in Bergen could be a direction for future work. Historical data of the number of bikes and slots accessible at each station is not currently available. However, Bergen City Bike has the opportunity to read and store the hourly data for a week to analyze how often stations are getting empty or full based on this data. With the help of collecting this data, the lost demand problem can be addressed more specifically, given that information regarding the time period the station was empty or full can indicate the period users could not rent or return the bikes. Additionally, we can ask users in the mobile application to which station they plan to return a bike when they unlock it. By collecting and studying this data we could observe if there is a difference between the station which a user indicated as a desired final destination and where the bike was actually returned. If such a difference occurs, the company could also check whether it was due to the fact that the station did not have empty slots at the time.

Furthermore, how employing the models will affect users is an interesting direction for potential future research. Ideally, after adopting the model developed in our thesis, the lost demand should be reduced due to consideration of unsatisfied demand in the optimization model. Moreover, if the station data could be collected in the future as mentioned in the previous passage, the more accurate forecasts based on real demand data is achievable. Also, the lost demand reduction might change the rentals' distribution after adopting the models. The effect on the future rentals after adopting the model is also interesting to examine.

Lastly, it would be beneficial to make the model more applicable to different bike-sharing companies. It is also helpful to extend the model for the dockless bike-sharing system and to discover how to solve the problem if there are no specific docks for users to pick up or return the bike.

8. Conclusion

The BSS in Bergen provides an environmental-friendly way of mobility. One of the main challenges is that the bike stations do not balance themselves, leading to difficulties in renting bikes or returning bikes. The imbalance problem causes a significant impact on the service level and attractiveness of the service for users. Therefore, the repositioning is needed to be executed to maintain the appropriate number of bikes at each station. However, Bergen City Bike performs repositioning intuitively. There might be inefficient repositioning routes and failure to deliver an adequate number of bikes for Bergen City Bike.

A repositioning model is proposed in this thesis. We start by forecasting the station-level net demand using the ARIMA model and the random forest model. Subsequently, we select the better-performed model, the ARIMA model in our thesis, and then apply its point forecast to determine the target number to be repositioned for each station. A Mixed-Integer Nonlinear Programming model is developed for deciding the optimal route of repositioning, considering the transportation cost, the unsatisfied demand cost, and the capacity of the vehicle and each station. The model is applicable in terms of cost concern for business, given that most of the tools we use are free of charge and computationally efficient for commercial software. However, if more powerful computers can be employed, the model can be tested in a longer time horizon, and the model's applicability can be more justified. Moreover, if there is more data available, the current model can be improved.

With the implementation of the proposed model, a better repositioning could be performed because the forecasting decides the adequate number to be repositioned and the optimization model makes sure the optimal route to be efficient in terms of meeting demand and reducing transportation cost. Hence, the availability of bikes and empty slots should increase, which leads to higher customer satisfaction.

There are some findings in the thesis for Bergen's BSS. First, the station correlated factors are not significant for Bergen's BSS. Second, traveling to more stops during repositioning can satisfy more demand and lead to a lower total cost. However, after reaching a particular value of unit transportation cost, further repositioning is not suggested. Furthermore, due to the relatively low unit transportation cost, traveling to more stops is also suggested if Bergen City Bike employs the strategy to add or remove more bikes than the point forecast.

More studies can be conducted in the future. In addition to improving the computation ability of the hardware and data availability, analyzing the trip data after adoption of the model, and extending the model for the dockless bike-sharing system are some of the possible directions of future work.

Acknowledgments

First of all, we gratefully acknowledge our supervisor Stein W. Wallace for his valuable suggestions and discussions throughout the thesis. Without your help, this work would never have been possible.

We are indebted to Deloitte's The Thesis that Matters Programme and consultants Mohammed Alhayek and Sondre Høyland. We highly appreciate your support amid your tight schedule.

We also thank Roald Sandstad, the operational manager of Bergen City Bike, for his ongoing collaboration with us in replying to our questions about Bergen's bike-sharing system's repositioning.

Lastly, thanks for all the mental support from our friends and families.

Norwegian School of Economics

Bergen, December 2020

Aida Suleimenova

Ping-Yi Chiang

References

Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T., & Shah, S. (2019). Forecasting traffic congestion using ARIMA modeling. *2019 15th International Wireless Communications & Mobile Computing Conference*, 1227-1232.

Alvarez-Valdes, R., Belenguer, J. M., Benavent, E., Bermudez, J. D., Muñoz, F., Vercher, E., & Verdejo, F. (2016). Optimizing the level of service quality of a bike-sharing system. *Omega, 62*, 163-175.

Andersen, M. (2016). *How much does each bike share ride cost a system? Let's do the math.* Retrieved 11 19, 2020, from Better bike share partnership: http://betterbikeshare.org/2016/08/16/

Andrysiak, T., Saganowski, Ł., Choraś, M., & Kozik, R. (2014). Network traffic prediction and anomaly detection based on ARFIMA model. *In International Joint Conference SOCO'14-CISIS'14-ICEUTE'14* (pp. 545-554). Springer.

Benchimol, M., Benchimol, P., Chappert, B., De La Taille, A., Laroche, F., Meunier, F., & Robinet, L. (2011). Balancing the stations of a self service "bike hire" system. *RAIRO-Operations Research, 45(1)*, 37-61.

Bergen City Bike. (2020). *Bergen City Bike*. Retrieved from Bergen City Bike: https://bergenbysykkel.no/en/

Bergen City Bike. (2020). *Open data*. Retrieved from Bergen City Bike: https://bergenbysykkel.no/en/open-data

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 192-213.

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

Brinkmann, J., Ulmer, M. W., & Mattfeld, D. C. (2015). Short-term strategies for stochastic inventory routing in bike sharing systems. *Transportation Research Procedia, 10*, 364-373.

Bullock, C., Brereton, F., & Bailey, S. (2017). The economic contribution of public bike-share to the sustainability and efficient functioning of cities. *Sustainable cities and society, 28*, 76-87.

Caggiani, L., & Ottomanelli, M. (2012). A modular soft computing based method for vehicles repositioning in bike-sharing systems. *Procedia-Social and Behavioral Sciences, 54*, 675-684.

Cagliero, L., Cerquitelli, T., Chiusano, S., Garza, P., & Xiao, X. (2017). Predicting critical conditions in bicycle sharing systems. *Computing, 99(1)*, 39-57.

Chemla, D., Meunier, F., & F., R. W. (2013). Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization, 10(2)*, 120-146.

Chen, S., Chung, W., & Chen, R. (2018). Using SVM and Random forest for different features selection in predicting bike rental amount. *International Conference on Awareness Science and Technology* (pp. 1-5). IEEE.

Chopra, S., Meindl, P., & Kalra, D. (2013). *Supply chain management: strategy, planning, and operation.* Pearson Boston, MA.

Chujai, P., Kerdprasop, N., & Kerdprasop, K. (2013). Time series analysis of household electric consumption with ARIMA and ARMA models. *Proceedings of the International MultiConference of Engineers and Computer Scientists* (pp. 295-300). Hong Kong: IMECS.

DeMaio, P. (2008). The Bike-sharing Phenomenon-The History of Bike-sharing. Carbusters Magazine, November.

European Commission. (2017). *Digital Economy and Society Index (DESI) 2017*. Retrieved from European Commission: https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-desi-2017

Fricker, C., & Gast, N. (2013). Incentives and redistribution in bike-sharing systems with stations of finite capacity.

Froehlich, J. E., Neumann, J., & Oliver, N. (2009). Sensing and predicting the pulse of the city through shared bicycling. *Twenty-First International Joint Conference on Artificial Intelligence.*

Gauthier, A.; Institute for Transportation and Development Policy. (2013). The bike-share planning guide. *ITDP Institute for Planning & Development Policy*.

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. *Transportation, 41(6)*, 1205-1225.

Gleditsch, M. D., & Hagen, K. (2018). *A Column Generation Heuristic for the Dynamic Rebalancing Problem in Bike Sharing Systems.*

Gray, A. (2017). *China's 'Uber for bikes' model is going global.* Retrieved from World economic forum: https://www.weforum.org/agenda/2017/06/china-leads-the-world-in-bike-sharing-and-now-its-uber-for-bikes-model-is-going-global/?utm_content=buffer697dc&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer&fbclid=IwAR00eLZ97m7dLU3rvJ3JjoYotlThAveMi

Heaney, A., Carrión, D., Burkart, K., Lesk, C., & Jack, D. (2019). Climate Change and Physical Activity: Estimated Impacts of Ambient Temperatures on Bikeshare Usage in New York City. *Environmental health perspectives*.

Hughes, C. (2017). *Bike Share: The Dawn of the Smartbike (and the Death of Dock-Blocking).* Retrieved from Medium: https://medium.com/social-bicycles/bikeshare-the-dawn-of-the-smartbike-and-the-death-of-dock-blocking-9f52bb642ae

Hulot, P., Aloise, D., & Jena, S. D. (2018). Towards station-level demand prediction for effective rebalancing in bike-sharing systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 378-386).

Hyndman, R. J., & Athanasopoulos, G. (2020). *Forecasting: principles and practice.* OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: springer.

Jensen, P., Rouquier, J. B., Ovtracht, N., & Robardet, C. (2010). Characterizing the speed and paths of shared bicycle use in Lyon. *Transportation research part D: transport and environment, 15(8)*, 522-524.

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.

Krishna, V. B., Iyer, R. K., & Sanders, W. H. (2015). ARIMA-based modeling and validation of consumption readings in power grids. *International Conference on Critical Information Infrastructures Security* (pp. 199-210). Springer, Cham.

Le Figaro. (2011, 9 29). *Vélos: 200 M de locations pour JCDecaux.* Retrieved from Vélos: 200 M de locations pour JCDecaux.: https://www.lefigaro.fr/flash-actu/2011/09/29/97001-20110929FILWWW00759-velos-200-m-de-locations-pour-jcdecaux.php

Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. . *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (pp. 1-10).

Lin, L., He, Z., & Peeta, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, 258-276.

Lukáš, A., & Branda, M. (2016). Nonlinear chance constrained problems: optimality conditions, regularization and solvers. *Journal of Optimization Theory and Applications*, 419-436.

Maggioni, F., Cagnolari, M., Bertazzi, L., & Wallace, S. W. (2018). Stochastic optimization models for a bike-sharing problem with transshipment. *European Journal of Operational Research*.

Mahmoud, M. S., El-Assi, W., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand : Station level analysis of commercial bike sharing in Toronto. *Transportation Research Board Annual Meeting*, (pp. 589-613).

Raviv, T., & Kolka, O. (2013). Optimal inventory management of a bike-sharing station. *Iie Transactions, 45(10)*, 1077-1093.

Regue, R., & Recker, W. (2014). Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. *Transportation Research Part E: Logistics and Transportation Review*, 192-209.

Rudloff, C., & Lackner, B. (2014). Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks. *Transportation Research Record, 2430(1)*, 1-11.

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

Shaheen, S. A., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: past, present, and future. *Transportation Research Record, 2143(1)*, 159-167.

Shu, J., Chou, M., Liu, Q., Teo, C. P., & Wang, I. L. (2010). Bicycle-sharing system: deployment, utilization and the value of re-distribution. *National University of Singapore-NUS Business School, Singapore.*

Sun, F., Chen, P., & Jiao, J. (2018). Promoting public bike-sharing: A lesson from the unsuccessful Pronto system. *Transportation Research Part D: Transport and Environment, 63*, 533-547.

The Local. (2019). *Why Norway's bike-sharing schemes outperform those in Sweden and Denmark*. Retrieved from The Local: https://www.thelocal.no/20190516/why-norways-bike-sharing-schemes-outperform-those-in-sweden-and-denmark

Thomas, T., Jaarsma, C. F., & Tutert, S. I. (2009). Temporal variations of bicycle demand in the Netherlands: The influence of weather on cycling.

Van der Zee, R. (2016). Story of cities# 30: how this Amsterdam inventor gave bike-sharing to the world. *The Guardian, 26.*

Vogel, P. (2016). Service network design of bike sharing systems. *Service Network Design of Bike Sharing Systems* (pp. 113-135). Springer, Cham.

Vogel, P., & Mattfeld, D. C. (2010). Modeling of repositioning activities in bike-sharing systems. *World conference on transport research (WCTR).*

Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences, 20*, 514-523.

Wang, W. (2016). Forecasting Bike Rental Demand Using New York Citi Bike Data.

Wang, Y., & Szeto, W. Y. (2018). Static green repositioning in bike sharing systems with broken bikes. *TransportationResearchPartD*, 438-457.

Waserhole, A., & Jost, V. (2012). Vehicle sharing system pricing regulation: Transit optimization of intractable queuing network. *HAL Id: hal-00751744*, 1-20.

Yin, Y. C., Lee, C. S., & Wong, Y. P. (2012). *Demand prediction of bicycle sharing systems*. Retrieved from Demand prediction of bicycle sharing systems: http://cs229. stanford. edu/proj2014/YuchunYin, ChiShuenLee, Yu-PoWong, DemandPredictionofBicycleSharingSystems. pdf.

Yoon, J. W., Pinelli, F., & Calabrese, F. (2012). Cityride: a predictive bike sharing journey advisor. *2012 IEEE 13th International Conference on Mobile Data Management* (pp. 306-311). IEEE.

Yr. (2020). *Hourly forecast for Bergen (Vestland)*. Retrieved from Yr.: https://www.yr.no/place/Norway/Vestland/Bergen/Bergen/hour_by_hour.html

Zhang, J., Pan, X., Li, M., & Philip, S. Y. (2016). Bicycle-sharing system analysis and trip prediction. *2016 17th IEEE international conference on mobile data management (MDM)* (pp. 174-179). IEEE.

Appendix A. The ARIMA Models for May 8, 2019

The backshift notation B is used to formulate the model. As a result, only the coefficients of the trained model in every station are listed. “Day”, “temp”, “pre” and “wind” are the abbreviation of the weekend dummy, temperature, precipitation, and wind speed, respectively. “ar”, “sar”, “ma”, “sma” are the abbreviation of autoregressive, seasonal autoregressive, moving average, and seasonal moving average. The number followed by “ar”, “sar”, “ma”, “sma” denotes the order for the specific component.

The regressions model with ARIMA errors for station-level pickups forecasting.

Station id	Coefficient of each variables							
131	ar1	ar2	sar1	inter	day	temp	pre	wind
	0.30	0.33	0.24	(6.13)	0.18	0.49	0.07	0.05
150	ar1	sar1	inter	day	temp	pre	wind	
	0.36	0.38	(5.63)	0.93	0.47	0.04	0.02	
151	ma1	sma1	day	temp	pre	wind		
	0.34	(0.53)	1.62	(0.01)	(0.03)	0.25		
217	ar1	sar1	inter	day	temp	pre	wind	
	0.27	0.34	(6.70)	1.68	0.19	0.10	0.25	
219	ma1	sar1	day	temp	pre	wind		
	0.23	(0.47)	1.37	(0.26)	(0.15)	0.15		
220	ma1	day	temp	pre	wind			
	0.16	0.56	(0.06)	(0.05)	(0.01)			
3	sar1	day	temp	pre	wind			
	(0.44)	1.28	(0.10)	0.06	0.21			
301	ar1	sar1	inter	day	temp	pre	wind	
	0.52	0.37	(3.92)	0.96	0.10	(0.17)	0.04	
34	ar1	sar1	drift	day	temp	pre	wind	
	0.11	(0.42)	0.00	0.82	(0.19)	(0.07)	0.11	
368	ar1	sar1	sar2	inter	day	temp	pre	wind
	0.82	0.25	0.03	(6.51)	1.83	0.21	0.05	0.15
58	ma1	sar1	day	temp	pre	wind		
	0.33	(0.48)	1.16	(0.17)	(0.23)	0.13		

640	sar1	inter	day	temp	pre	wind				
	0.66	(3.99)	0.49	0.16	0.14	0.09				
641	sma1	sma2	day	temp	pre	wind				
	(1.01)	0.23	1.13	(0.09)	0.00	0.16				
644	ar1	sar1	sar2	inter	day	temp	pre	wind		
	0.37	0.23	0.17	(5.12)	1.40	0.20	0.04	0.13		
75	ar1	ma1	sar1	sma1	sma2	inter	day	temp	pre	wind
	0.72	(0.31)	0.05	0.20	0.38	(4.30)	0.98	0.08	0.24	(0.00)
794	day	temp	pre	wind						
	1.19	0.27	(0.08)	0.09						
814	ar1	sar1	drift	day	temp	pre	wind			
	0.05	(0.53)	0.00	0.27	(0.05)	(0.01)	0.04			
815	day	temp	pre	wind						
	2.10	0.03	0.01	0.12						
82	ar1	sar1	inter	day	temp	pre	wind			
	0.19	0.19	(6.77)	0.59	0.42	0.16	0.13			
83	ma1	sar1	sar2	day	temp	pre	wind			
	0.16	(0.80)	(0.40)	0.92	(0.08)	0.08	0.14			

The regressions model with ARIMA errors for station-level returns forecasting.

Station id	Coefficient of each elements											
131	ar1	ma1	sar1	sar2	inter	day	temp	pre	wind			
	0.69	(0.27)	0.16	0.19	(3.52)	0.49	(0.01)	(0.04)	0.18			
150	ma1	sar1	inter	day	temp	pre	wind					
	0.34	0.49	(4.88)	0.34	0.38	(0.15)	0.14					
151	sar1	drift	day	temp	pre	wind						
	(0.34)	0.01	1.17	0.04	(0.01)	0.25						
217	ma1	sar1	sma1	day	temp	pre	wind					
	(0.60)	(0.00)	0.34	1.45	0.53	0.22	0.29					
219	ar1	sar1	day	temp	pre	wind						
	(0.35)	0.40	0.61	(0.26)	0.07	0.49						

220	ma1	sar1	drift	day	temp	pre	wind					
	0.25	(0.48)	(0.00)	1.01	(0.16)	0.05	0.06					
3	sar1	sar2	sma1	day	temp	pre	wind					
	(0.33)	(0.18)	(0.62)	0.77	(0.22)	0.08	0.17					
301	ar1	ar2	ar3	ar4	sar1	sar2	sma1	day	temp	pre	wind	
	(0.65)	(0.63)	(0.28)	(0.19)	(0.05)	(0.24)	(0.77)	0.26	0.01	(0.02)	(0.09)	
34	ar1	ar2	ar3	ar4	ar5	ma1	sar1	sar2	day	temp	pre	wind
	(0.22)	(0.23)	(0.16)	(0.09)	(0.06)	(0.84)	(0.49)	(0.37)	1.05	0.03	(0.05)	(0.00)
368	ar1	ma1	sar1	sar2	inter	day	temp	pre	wind			
	0.83	(0.01)	0.29	(0.22)	(5.99)	1.36	0.24	(0.14)	0.09			
58	sar1	day	temp	pre	wind							
	(0.39)	1.49	(0.09)	(0.12)	0.13							
640	sar1	inter	day	temp	pre	wind						
	0.72	(4.63)	0.71	0.24	(0.10)	0.08						
641	ma1	sar1	day	temp	pre	wind						
	0.22	(0.47)	1.10	(0.18)	0.03	0.16						
644	ar1	ma1	sar1	day	temp	pre	wind					
	0.42	(0.98)	0.36	1.30	0.31	0.01	0.28					
75	ar1	ar2	ma1	sar1	sar2	inter	day	temp	pre	wind		
	(0.47)	0.37	0.84	0.27	0.16	(4.60)	0.67	0.12	0.13	0.15		
794	ma1	ma2	ma3	sma1	day	temp	pre	wind				
	0.17	0.26	0.13	(0.78)	1.59	(0.14)	(0.09)	0.15				
814	ma1	ma2	sar1	sar2	sma1	sma2	day	temp	pre	wind		
	0.11	(0.10)	(0.68)	(0.31)	0.03	(0.53)	0.78	(0.24)	(0.02)	0.12		
815	ma1	ma2	sar1	sma1	day	temp	pre	wind				
	0.04	0.17	(0.01)	(0.59)	1.42	(0.12)	0.02	0.11				
82	ar1	ma1	inter	day	temp	pre	wind					
	0.74	(0.44)	(6.18)	0.14	0.49	(0.07)	0.05					
83	sar1	sar2	sma1	day	temp	pre	wind					
	0.10	(0.00)	(0.74)	0.90	(0.26)	(0.08)	0.25					