Introduction

William K Davis III

2022-06-25

Introduction

While the concept of renting transportation has been around for decades in the form of rental car companies, such as those often found at airports, it has also been adapted to the modern "sharing economy" in the form of bikeshare programs. These programs, often offered by local governments, allow users to rent bicycles for individual (point-to-point) trips, such as commuting to and from work. Cycling can be a faster mode of transportation than walking and might even be faster than driving or taking a taxi in the most congested of cities. Finally, bicycles offer a lower-pollution alternative to driving, which can be appealing to cities struggling to contain emissions.

A key challenge faced by administrators of bikesharing programs is the efficient allocation of available bicycles. Bikes must be available in the places that people need them and at the time they are needed in order for the program to be effective. In order to efficiently allocate bikes, administrators can periodically transport bikes from areas of low demand to areas of high demand (Schuijbroek et al. [2017]). Demand in a bikeshare system is a function of both time and location. In this paper we focus on the time component of the demand for bicycles at each hour of the day, ignoring the spatial component of demand. We will use the Seoul Bike Sharing Demand data from UCI (UCI [2020]). Our focus will be limited to predicting demand at each hour of the day, ignoring inferential aspects of the analysis. While much of the recent research using this dataset has focused on machine learning methodologies, we will take a multifaceted approach that incorporates advances in time series modeling in addition to the popular machine learning methodologies.

This paper begins with a review of the relevant literature, including studies of the Seoul data specifically. Next, we present the results of the exploratory data analysis, including a description of the dataset and relevant profiles of the features. In part

Literature

The popularity and accessibility of the Seoul bike dataset has resulted in its use for numerous studies. A majority of these studies have focused on the use of various machine learning algorithms. Sathishkumar and Yongyun found that a CUBIST model, which combines tree- and regression-based methods into a series of rules, performed best on the Seoul data when measured by R^2 and RMSE on the testing dataset (E and Cho [2020]). Gao and Chen found that another tree-based method, random forest (RF), performed best on a similar bikeshare dataset when measured by R^2 and RMSE (Gao and Chen [2022]). Both studies further showed that weather-related variables, such as temperatures and precipitation, where among the most important for predicting demand. Gao and Chen's results highlight the importance of selecting a relevant evaluation metric and explanatory variables. When socioeconomic variables were included in the model, the RF outperformed the support vector machine (SVM) when measured by both RMSE and MAE. However, when the socioeconomic variables were excluded from the model, the RF outperformed the SVM when measured by RMSE, but the SVM performed better when measured by MAE This indicates that without the socioeconomic variables included, the SVM was prone to a few errors that were quite large in magnitude, while the RF was more prone to smaller but more frequent errors. This reinforces the importance of using multiple metrics when evaluating predictions.

Considerably fewer researchers have made use of traditional time series methodologies when predicting

demand of a similar nature to the bikeshare data. Both (E and Cho [2020]) and (Gao and Chen [2022]) make use of temporal variables such as hour, day of the week, and holidays. In each case they were found to be of moderate or high importance. (Gao and Chen [2022]) applied linear regression using temporal variables such as a weekend indicator as a predictor, but this model greatly underperformed the machine learning methods. Further, there is no discussion of any attention paid to stationarity and autoregression, which are common in time series data but may also result in violations of the standard assumptions of linear regression.

Our analysis looks to build on this work by using newer machine learning methods such as long short-term memory and recurrent neural networks, as well as modern time series techniques that leverage the underlying structure of the data.

Exploratory Analysis

The dataset consists of 8,760 hourly observations of 12 variables from 2017-12-01 00:00:00 to 2018-11-30 23:00:00. There are 295 observations where *BikeCount*=0 due to the bikeshare system not functioning. There are no other periods where *BikeCount*=0. @ref(tab:dictionary) contains information on the variables in the dataset, including the

Table 1: Variable definitions

Variable	name	Type	Definition
Hour	Hour	datetime	year-month-day hour:minute:second
Rented Bike count	BikeCount	numeric	Count of bikes rented at each hour
Temperature	Temperature	numeric	Temperature in Celsius
Humidity	Humidity	numeric	% humidity
Windspeed	WindSpeed	numeric	meters/second
Visibility	Visibility	numeric	in 10m
Dew point temperature	Dewpoint	numeric	Celsius
Solar radiation	SolarRadtion	numeric	MJ/m2
Rainfall	Rainfall	numeric	mm
Snowfall	Snowfall	numeric	cm
Seasons	Seasons	categorical	Winter, Spring, Summer, Autumn
Holiday	Holiday	categorical	Holiday/No holiday
Functional Day	FunctionalDay	categorical	NoFunc(Non Functional Hours), Fun(Functional hours)
Workday	Workday	categorical	Workday/Not Workday. A workday is a weekday that is not a holiday.

Bike Count

The data shows increasing demand and variability during the summer months. The bike demand data are counts, meaning they are technically discrete. Based on the histogram and the count nature of the data, it appears that a poisson distribution would be most appropriate for the data. If we were to model the bike demand on a continuous scale, a log-normal distribution might be appropriate.

@ref(fig:timeplot) highlights the incredible variation in demand over time. The variance in demand appears to increase during the summer months and then decrease again in the autumn. This heteroskedasticity violates the constant variance assumption of ordinary least squares regression and will have to be corrected if regression-based methods are to be used. We can also see that the mean of the series appears to increase during the summer months before decreasing again in the autumn.

A KPSS test yields $p = 0.01$, meaning there is strong evidence in favor of the presence of a unit root and the data is likely non-stationary (Kwiatkowski et al. [1992]). A Ljung-Box test was conducted up to 24 lags which resulted in a test statistics of 39743 with $p = 0$, indicating strong evidence that there is serial correlation in the hourly bike count (LJUNG and BOX [1978]). The acf plot in Figure A shows that autocorrelation is

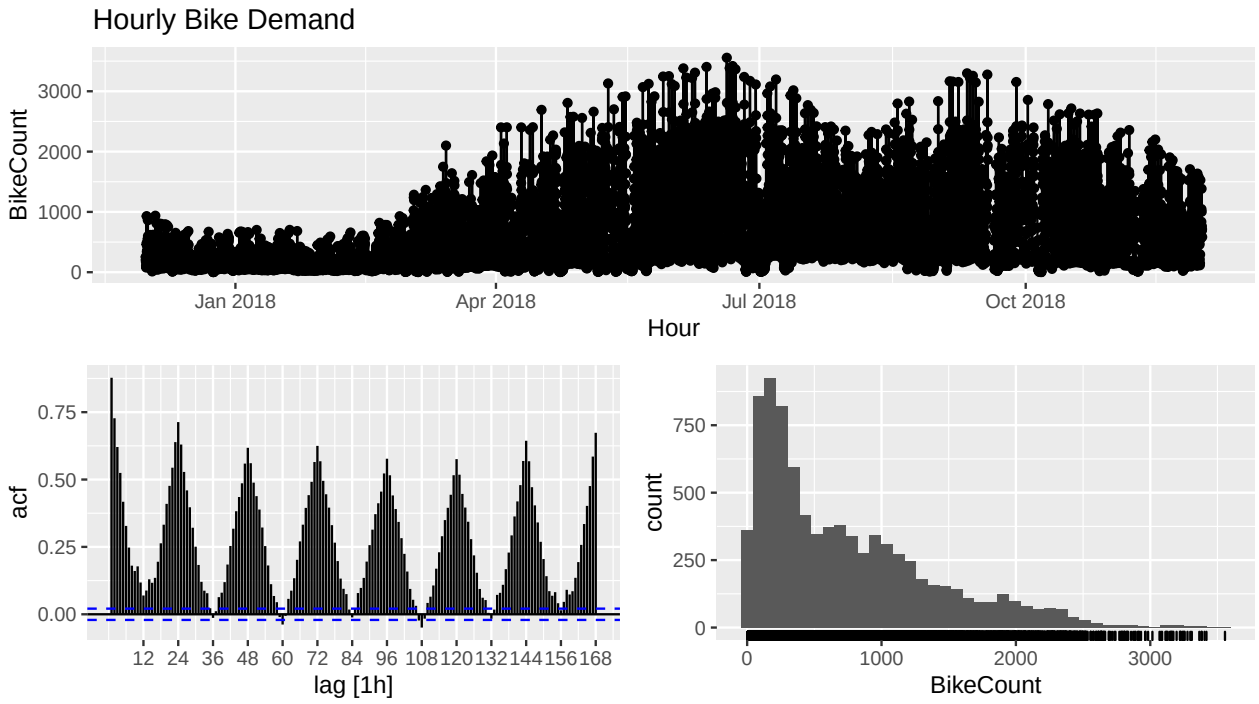

Figure 1: Hourly bike demand

significant at most lags out to 168 hours, which represents the same hour of the same day in the previous week. The strong serial correlation makes this dataset a good candidate for time series techniques.

Figure 2 highlights various seasonal and temporal patterns in the data. There is certainly an effect based on the hour of day, with demand increasing sharply in the morning, presumably during the morning commute, before decreasing into the lunchtime hour. From there, demand rises steadily through the evening commute before peaking around dinner time and then falling through the nighttime hours. Variability also appears greater during the evening hours, which is consistent with the idea heteroscedasticity; the variance increases with the level of the series. Demand appears slightly higher during workdays (where workdays are weekdays that are not holidays) and weekdays, though the difference does not appear particularly large. There are a number of large positive outliers during the weekdays. Finally, demand appears largest during the summer months and smallest during the winter, as biking would be a less desirable option in the cold.

Figure 3 provides insight into the seasonality of the bike demand. Each panel is a 4-week sample of hourly bike demand, with non-workdays (weekends and holidays) highlighted in red. While these days appear to follow a consistent hourly pattern much like the workdays, we can see that their is a difference between the seasonality of workdays and the seasonality of non-workdays. Most notably, the troughs in demand appear consistent between the types of days while the peaks are consistently higher for workdays as compared to non-workdays. This changing seasonality based on type of day will need to be captured in our model.

Continuing with the impact of seasonality and temperature on demand, we will next explore covariates included with the dataset.

Covariates

This dataset includes a number of covariates to aid in modeling bike demand. All of these covariates are listed in Table 1, along with their definitions. The covariates fall into one of two broad categories: weather and social. Weather covariates include variables such as temperature, humidity, precipitation, and others. Social variables capture the impact of calendar-based human behavior, such as holidays and weekends.

Time plots of the covariates (Figure D) show a dynamic set of variables. Most of the covariates appear

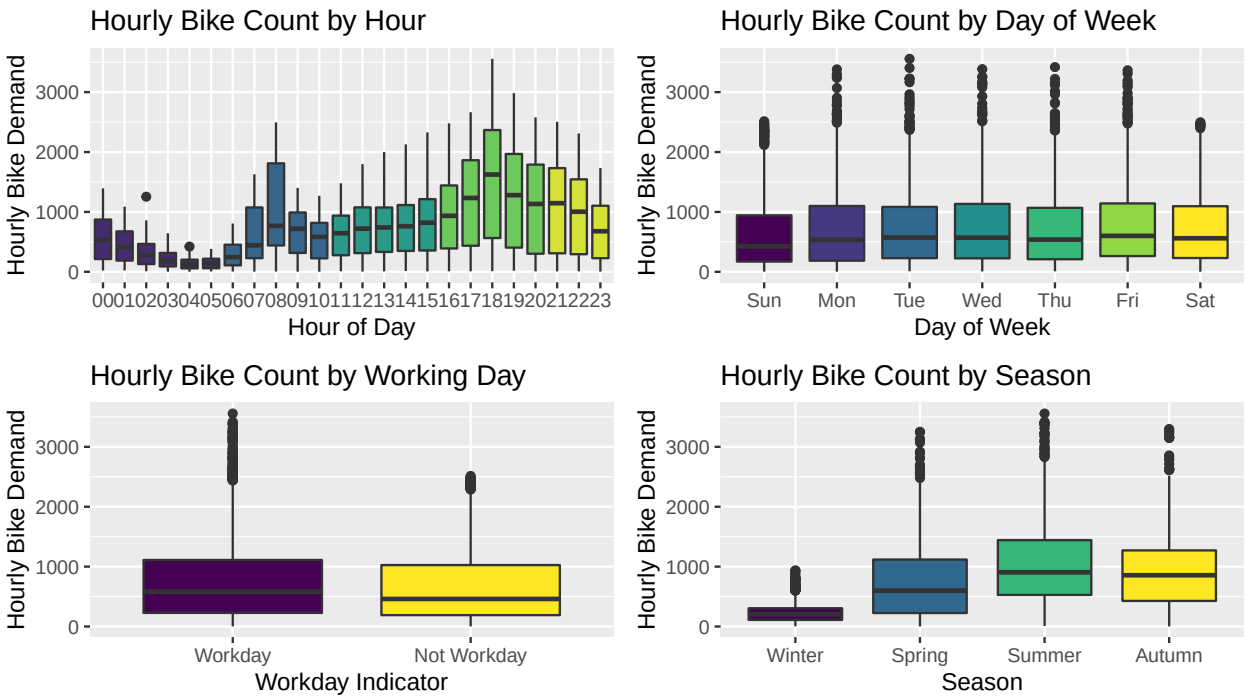

Figure 2: Hourly bike demand by time period

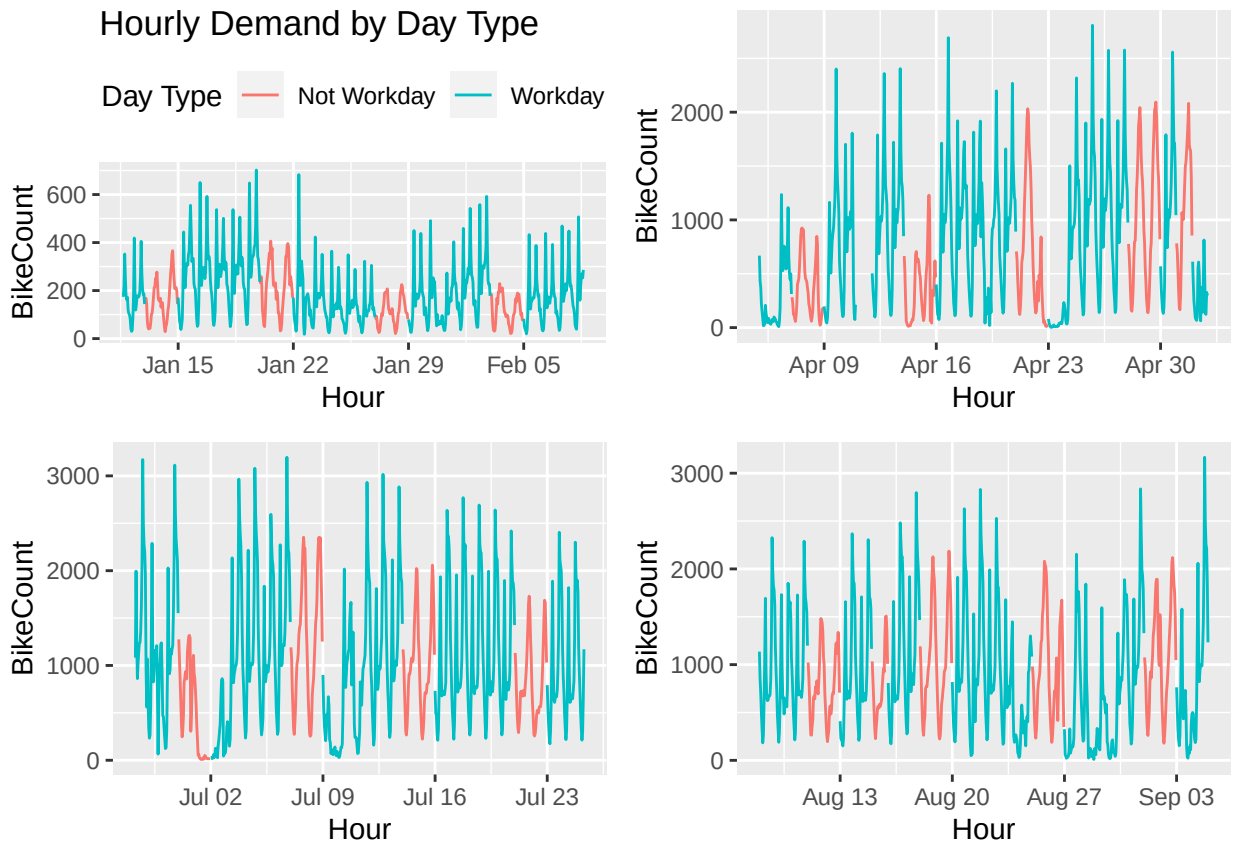

Figure 3: Hourly demand by day type

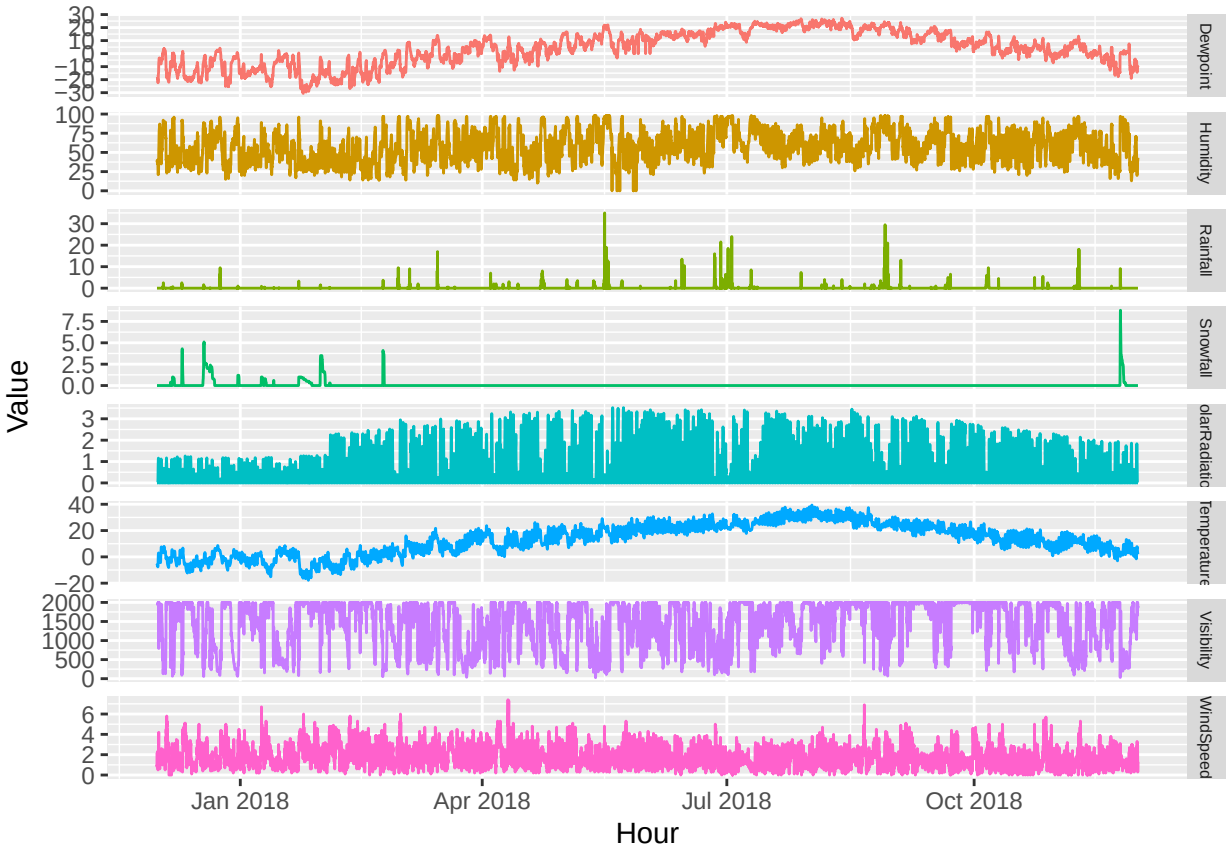

Figure 4: Covariate timeplots

to be typical time series data with varying degrees of trend, seasonality, cyclicality, autocorrelation, and heteroskedasticity. Dewpoint and Solar Radiation follow a predictable pattern that mirrors temperature throughout the year, with an upward trend peaking in the summer months and downward trend that hits a trough in the winter months. Humidity, visibility, and Wind Speed appear to have less of a trend throughout the year. Precipitation (rainfall and snowfall) appears to have a much more random pattern throughout the year, with a large number of periods having no precipitation. Depending on the predictive performance of the raw continuous precipitation features, it may prove more performent to convert them to binary variables that simply indicate if precipitation occurred during that period.

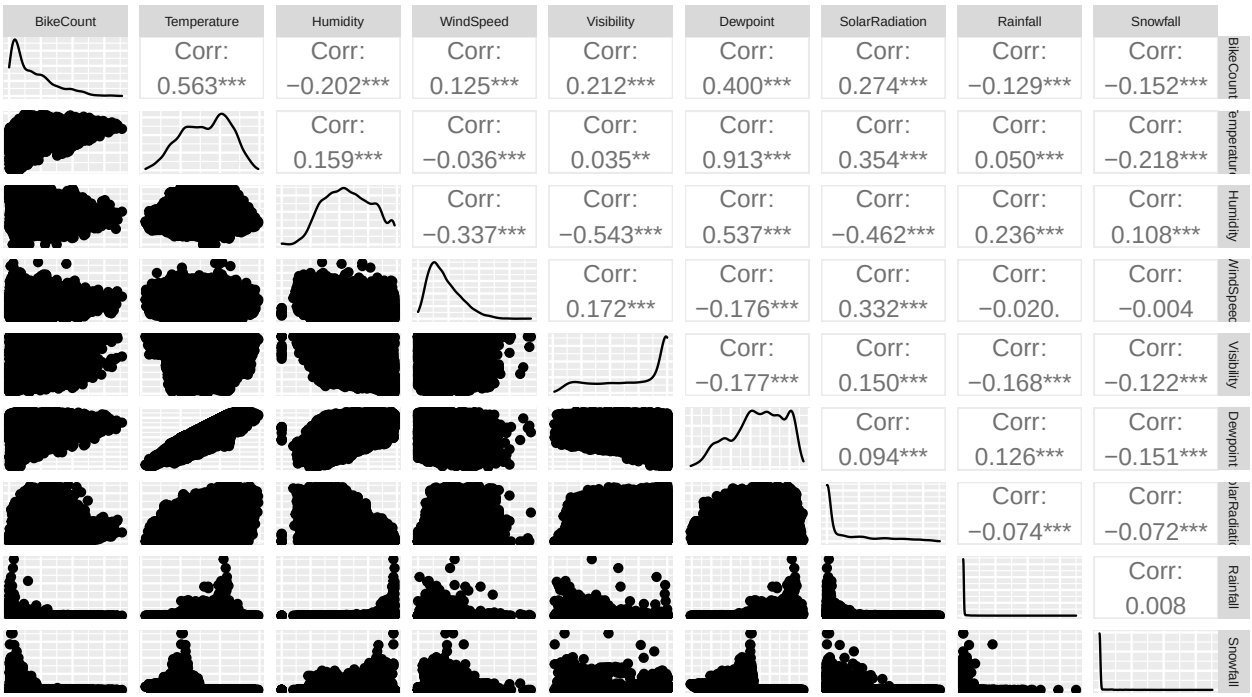

Figure 5: Covariate correlation plot

The correlation plot highlights a number of important relationships in the data. Bike count appears moderately linearly correlated with both temperature ($\rho = 0.563$) and dewpoint ($\rho = 0.400$). However, this is significant multicollinearity between temperature and dewpoint ($\rho = 0.913$), so additional analysis will be required to isolate the effect of each variable. Finally, we can see that a number of the covariates, such as wind speed, visibility, solar radiation, rainfall, and snowfall all appear to follow non-normal distributions. As such, their relationship to bike count might be best represented with a higher-order polynomial.

References

Sathishkumar V E and Yongyun Cho. A rule-based model for seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53(sup1):166–183, 2020. doi: 10.1080/22797254.2020.1725789. URL https://doi.org/10.1080/22797254.2020.1725789.

Chang Gao and Yong Chen. Using machine learning methods to predict demand for bike sharing. In Jason L. Stienmetz, Berta Ferrer-Rosell, and David Massimo, editors, *Information and Communication Technologies in Tourism 2022*, pages 282–296, Cham, 2022. Springer International Publishing. ISBN 978-3-030-94751-4.

Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y.

G. M. LJUNG and G. E. P. BOX. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 08 1978. ISSN 0006-3444. doi: 10.1093/biomet/65.2.297. URL https://doi.org/10.1093/biomet/65.2.297.

J. Schuijbroek, R.C. Hampshire, and W.-J. van Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257(3):992–1004, 2017. URL https://EconPapers.repec.org/RePEc:eee:ejores:v:257:y:2017:i:3:p:992-1004.

UCI. Seoul bike sharing demand data set, Mar 2020. URL https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand.