Estimating the Number of Bikeshare Trips

Yuta Baba and Saahithi Rao

Abstract

Bike share programs have become increasingly popular in the US, but it is often difficult to know the expected number of riders because cyclists are impacted by factors like weather and familiarity with cycling. We investigated how seasonality, weekends, rider information, such as age, gender and subscribe type, and weather, including temperature, snowfall, precipitation, and wind speed, influenced the number of daily trips in 2017 in Boston, Chicago, and New York City using Bayesian methods. We fit a poisson likelihood on average number of daily trips with weakly informative priors on the predictor variables and ran a Hamiltonian Monte Carlo simulation. Samples from the posterior distribution showed convergence for each parameter, large numbers of effective draws, and low deviance. We concluded that all predictors were associated with number of daily trips with precipitation, snowfall, wind, proportion of males and weekends showing negative associations and temperature and age showing positive associations, while holding all other predictors constant. Fall was the most popular time to cycle, while winter was the least popular time. These factors are useful in explaining the average daily trips in 2017 in these three cities, but this study can be expanded to 2018 and other US cities.

Introduction

Bike share programs are a way for users to rent bikes at a low cost through an app with the small caveat that users must rent and return the bike to any official dock station in the area. Bike share companies are becoming increasingly popular across cities in the US, both for their convenience while remaining environmentally friendly and their promotion of physical activity.

More and more people are cycling to run errands, go sightseeing, and transport themselves between locations. The bike share program at Carleton College, unfortunately, failed due to fewer than expected number of users, so we wanted to determine what factors impact demand of users. Specifically, we were interested in how demographics of users, weather information, and city knowledge influence the number of daily trips in a given city.

We chose to focus on Blue bikes, in Boston, Divvy Bikes, in Chicago, and Citi Bikes, in New York City, which are all popular bike share companies in their respective cities that work with bikeshare.com, a global directory for bike share companies and users. These companies encourage those interested to look into their data and are transparent with their policies.

Previous studies looked at the impact of built environmental factors like street density, street design, and weather on the number of bike share riders. This research stemmed from the idea that some cities are more bike friendly than others. One study found that street density and route connectivity, the orientation of streets, were positively associated with the number of people cycling in a particular city using spatial regression analysis (Zhang 2017). Another study looked at the difference between Canadian cyclists and cyclists in the US and environmental factors that affect both groups. Researchers found that precipitation was associated with lower cycling rates, while temperature was inconclusive (Puehler & Buehler, 2016). We wanted to see how taking into account both the distribution of cities and temporal data would inform the characteristics and number of bike share users in a particular city using Bayesian modeling.

Data

We collected three types of data: bikeshare, weather and city characteristics. Our bike share data was collected for 2017 for each city: Chicago, Boston, and NYC separately (“Bike

Share Portal", 2018). This bike share data, in total of 21,507,445 observations, contained information about each individual trip such as duration of the ride and geolocation of the start and end point. If a user was a subscriber of the program, then their demographic information including gender and birth date were recorded in the data. We summarized this data into daily observations, aggregating the average number of trips per day and the corresponding characteristics of the riders including median age and percentage of subscribers and male riders that day. We chose to use median age rather than average age because it would account for the sizable number of young riders in our data set. We also computed whether the day was a weekend or weekday and its corresponding season based on its month with December through February defined as winter, March through May defined as spring, June through August defined as summer, and September through November defined as fall. Furthermore, we obtained the number of stations in each city as of 2017 from each companies' website ("Bike Share Portal", 2018).

In addition, we collected daily weather data during 2017 at each city from NOAA ("NOAA", 2017). The weather data consisted of average, maximum and minimum temperatures in farenheit, average wind speed in miles per hour, cumulative precipitation and snowfall in inches for each day. Because there was snowfall for a short period of the year, we made this a binary variable of whether or not there was snow. We chose to use maximum temperature as a substitute for average temperature because this information was missing from the New York data and most trips occurred in the afternoon when the maximum temperature was reached.

Lastly, we collected information about how the organization of streets in each city. Specifically, we referred to the study conducted by Geoff Boeing that calculated the Shannon entropy, *H*, of the city's orientations' distribution. Boeing defined it as

$$H_o = -\Sigma^{n}_{i=1} P(o_i) log_e P(o_i)$$

where *n* represents the total number of bins, *i* indexes the bins, and P(o_i) represents the proportion of orientations that fall in the i_{th} bin.[1] Chicago, for example, was found to have a low street orientation entropy with its grid like street structure (Boeing 2018).

We combined all of the datasets into one comprehensive data set containing 1091 cases (365 daily observations per city with four days missing from NYC). Variables in the data set consist of what season the day falls in, whether the day is weekend or weekday and rider demographics including the percentage of subscribers, median age, and percentage of males. The number of trips that day (the response variable), city, precipitation (in), snowfall (binary), average wind speed (mph), and maximum temperature (℉) are also variables in the data set. In addition, the dataset also includes the number of stations and street entropy. The following tables show summaries of each variable of the 1091 observations.

Table 1: Numeric Summaries for variables:

	Min	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**SD**
Subscriber (%)	0.39	0.82	0.91	0.87	0.96	1	0.12
Male (%)	0.27	0.61	0.69	0.66	0.74	0.97	0.11
Median Age (yrs)	27	31	32	32.76	35	40	2.66
Average Wind Speed (mph)	1.12	5.59	8.50	8.89	11.63	25.28	4.28
Precipitation (in)	0.00	0.00	0.00	0.12	0.05	4.19	0.33
Max Temperature (℉)	5	47	63	61.59	78	95	19.02
Number of trips	38	4468	8801	1971	31352	74580	21391

Table 2: Number of Weekends and Weekdays with respect to

[1] Specifically, in this study, bins are calculated by dividing the street network's edges' individual compass bearings into 36-equal-sized bins (each represent 10°). See more details at page 5~6 of the study.

	Weekends	Weekdays
Frequency	315	776

Table 3: Number of Days with and without Snowfall

	Days with Snow	Days without Snow
Frequency	47	1050

Table 4: Number of Stations and Street Orientation of each City

City	Number of Stations	Entropy
Boston	194	3.55
Chicago	580	2.89
New York City	706	2.65

We conducted exploratory data analysis and found right skewness of the response variable as we imagined, since daily trips is a count. We also saw a positive association between logged number of daily bike trips and maximum temperature. Furthermore, we observed a proportion of male riders to be negatively correlated with logged number of daily bike trips with different intercepts for weekends and weekdays (*Appendix A*).

Methods

Since we are assessing counts of daily trips, we fit a poisson model. In addition, because counts are discrete and cannot be negative, we implemented a Monte Carlo simulation rather than a quadratic approximation. Specifically, we used a Hamiltonian Monte Carlo simulation through map2stan from the rethinking package. For numeric stabilization and to reduce correlation between parameters, we standardized maximum temperature, precipitation, average wind speed, median age, percentage of males, and percentage of subscribers before running the model. We also included fixed intercepts in the form of the four seasons to account for seasonality differences.

The following is the model we fit:

$$Likelihood:\ Y_i \sim Poisson(\lambda_i)$$

$$Link\ Function:$$

$$log(\lambda) = \alpha[winter] + \alpha[spring] + \alpha[summer] + \alpha[fall] + \beta_{boston} Boston + \beta_{chicago} Chicago +$$
$$\beta_{temp} MaxTemp + \beta_{prcp} Precipitation + \beta_{snow} Snow(1: Yes,\ 0: No) + \beta_{wind} AvgWindSpeed +$$
$$\beta_{weekend} Weekends(1: Yes,\ 0: No) + \beta_{age} MedianAge +$$
$$\beta_{male} Male + \beta_{subscriber} Subscriber$$

$$Priors:$$

$$\alpha[winter],\ \alpha[spring],\ \alpha[summer],\ \alpha[fall] \sim dnorm(\mu=9.88,\ \sigma=12)$$

$$\beta_{boston},\ \beta_{chicago} \sim dnorm(0,\ 3)$$

$$\beta_{temp},\ \beta_{prcp},\ \beta_{snow},\ \beta_{wind},\ \beta_{weekend},\ \beta_{age},\ \beta_{male},\ \beta_{subscriber} \sim dnorm(0, 5)$$

Where Y is the average number of daily bike trips, *Male* is a percentage of male users, and *Subscriber* is percentage of subscribers. σ is the standard deviation.

Because we are log transforming our outcome of number of daily trips, we set a weakly informative prior for the seasons to have a normal distribution with mean of the log mean number of trips overall and a standard deviation of 12 to account for varying number of rides especially in the winter. Similarly, for the remaining predictors, we used a mean of 0 and standard deviation of either three or five because all numeric predictors were standardized and three or five on an exponential scale was large enough to account for wide variation for both the numeric and categorical variables.

We simulated the posterior distribution of counts of daily trips and sampled from this distribution to assess convergence of our monte carlo simulation. We determined how well our

model converged through Rhat values and trace plots showing four chains of parameters for each predictor variable. If the Rhat values were smaller than 1.1 and closer to 1, we determined that our model was not completely inaccurate. Subsequently, if the trace plots showed stationary distributions and high numbers of effective draws, we inferred that our model converged. Since this project aims to evaluate what are the contributing factors to the number of bike trips, we will look at coefficients of each variable to determine its impact.

We used a model building process of adding and removing variables and comparing number of effective draws and WAIC values between models to assess which model was better at describing the data and making predictions on new data. We simulated and sampled from a poisson model across the three cities and models with and without seasons and demographics of the users before arriving at the model described above (Appendix B: Figure 3, Table 1). In the discussion section, we mention other models that we tried but were unsuccessful including a model with city characteristic variables and a multilevel model.

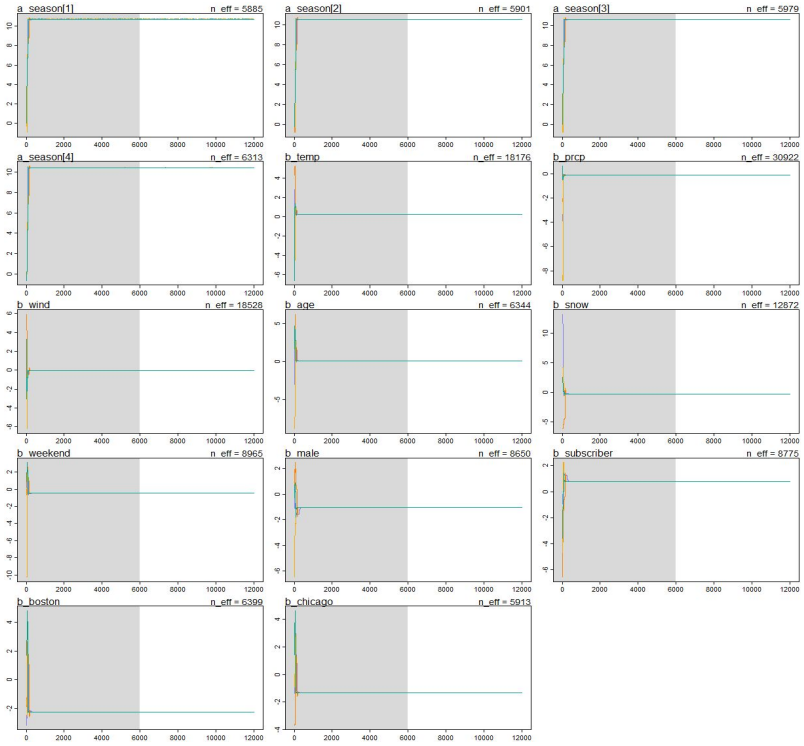
Results

Details of Hamiltonian Monte Carlo Simulation

As mentioned previously, we ran a Hamiltonian Monte Carlo simulation with 12,000 iterations 6,000 of which were used for the warm up region and sampling for each of the four chains respectively. The warm up region was specified at 6,000 iterations to allow for parameters that started at extreme values to stabilize. The Rhat value for each variable was exactly 1; trace plots showed that each variable with four chains converged at a value without symptoms of correlation (Figure 1, Table 5). Therefore, we confirmed the stationary and well-mixed nature of

the Markov chains. In addition, the number of effective draws was well over 6,000 for each parameter (Figure 1, Table 5). Thus, we concluded that the Monte Carlo simulation converged.

Figure 1: Model Trace Plots



Summary of Posterior Distribution

As described above, we created several models and compared their WAIC values to determine the most preferred model (Appendix B: Figure 3, Table 1). In total, we tried three models with one without season and rider information, and another without season. Figure 2 shows the summary of the difference in each model's estimation for each variable (Figure 2). Looking at three models, the coefficients for each variable are similar, suggesting that the these models estimated the impact of each variable to be similar.

Specifically, we found season, city, temperature, precipitation, wind speed, snowfall, weekends, median age, percentage of males, and percentage of subscribers to be significant in

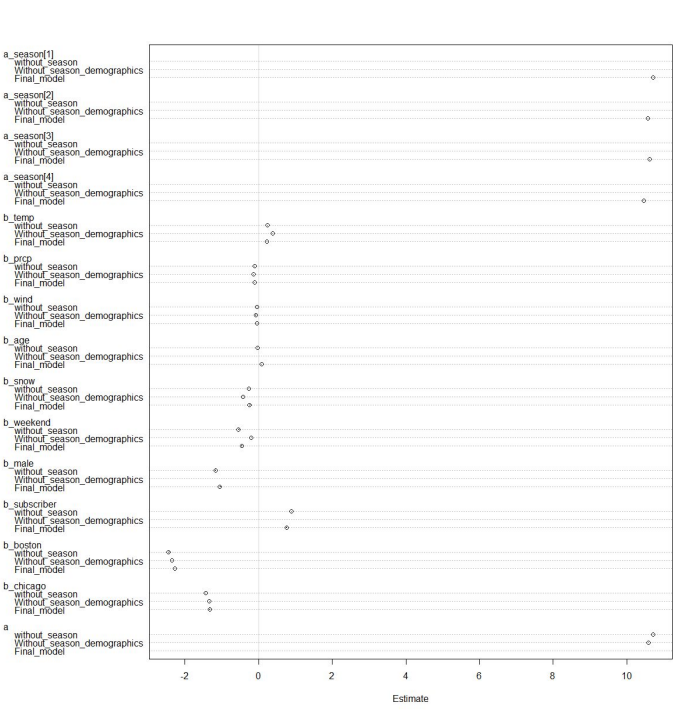
explaining the average number of daily bike share trips. During Fall, the average number of daily trips is 44,801, while in Winter the average daily number of trips is 34, 891 while holding all other predictors constant (Table 5). The average number of daily trips is 0.10 times lower in Boston compared to New York and Chicago combined, while the the average number is 0.26 times lower in Chicago compared to New York and Boston combined while holding all other predictors constant (Table 5). A one unit standard deviation increase in maximum temperature is associated with 1.26 times increase in the daily trips while holding all other predictors constant (Table 5). A one unit standard deviation increase in precipitation is associated with 0.90 times lower number of estimated daily trips while holding all other predictors constant (Table 5). Figure 3 illustrates the 89% credible and prediction intervals for number of bike trips against precipitation (Figure 3). Daily number of bike trips is 0.79 times lower on a day with snowfall than a day without snow while holding all other predictors constant (Table 5). A one unit standard deviation increase in proportion of males is associated with a 0.35 times less number of estimated daily trips while holding all other predictors constant (Table 5). A weekend day is associated with 0.63 times less number of bike trips while holding all other predictors constant (Table 5).

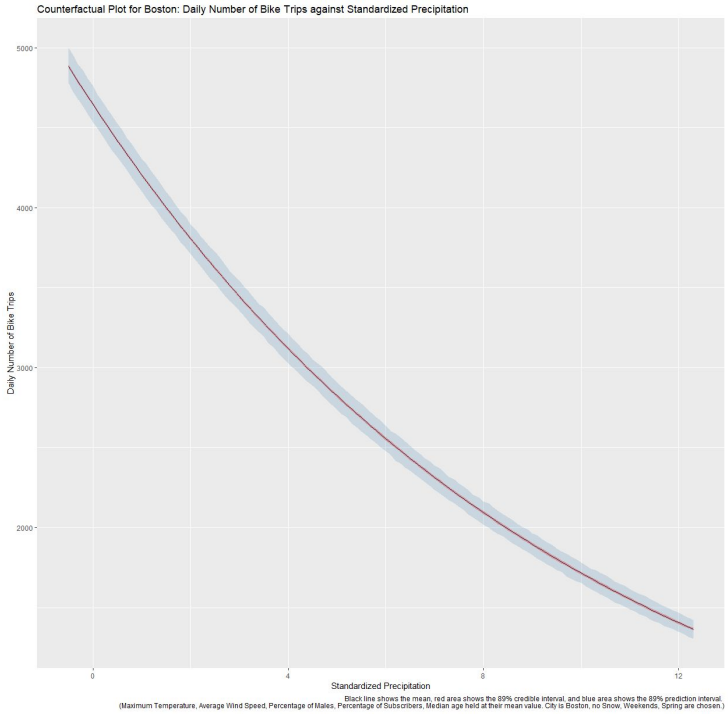
Table 5: Model Estimates

	Mean	**Standard Deviation**	**5.5%**	**94.5%**	**Number of Effective Draws**	**Rhat**
Fall	10.71	0	10.71	10.71	5885	1
Spring	10.57	0	10.57	10.58	5901	1
Summer	10.62	0	10.62	10.62	5979	1
Winter	10.46	0	10.45	10.46	6313	1
Boston	-2.27	0	-2.27	-2.26	6399	1
Chicago	-1.33	0	-1.33	-1.33	5913	1
Max Temperature (℉)	0.23	0	0.23	0.23	18176	1

Precipitation (in)	-0.10	0	-0.10	-0.10	30922	1
Snow	-0.24	0	-0.24	-0.23	12872	1
Average Wind Speed (mph)	-0.04	0	-0.04	-0.04	30922	1
Weekends	-0.46	0	-0.46	-0.46	8965	1
Median Age (yrs)	0.08	0	0.08	0.09	6344	1
Male (%)	-1.05	0	-1.05	-1.04	8650	1
Subscriber (%)	0.76	0	0.76	0.77	8775	1

Figure 2: Parameter Estimates with Credible Intervals for Each Model (Left)
Figure 3: Counterfactual Plot for Daily Number of Bike Trips against Precipitation (Right)





Discussion

Data on bike share trips for the year 2017 was collected from Boston, Chicago, and New York. Average number of daily trips, demographic information on users including age, gender, and subscriber type, and weather data including season, wind speed, temperature, snowfall and precipitation was collected. We used Bayesian methods of setting weakly informative priors on

predictors and a poisson likelihood on average number of daily trips to run a Monte Carlo simulation of the posterior distribution of average number of daily trips while accounting for demographic and weather influence.

The final model demonstrated that season, whether the city was in Boston, Chicago, or New York, temperature, wind speed, precipitation, snowfall, weekend days, proportion of males, proportion of subscribers, and median age all helped explain the average number of daily trips in 2017. Intercept coefficients of fall and summer showed higher counts of daily trips compared to winter and spring (Table 5). This could be explained by higher temperatures during these seasons, which was positively associated with number of trips, and less precipitation and snow, which both had negative associations. This supports the findings of Pucher and Buehler who determined that precipitation was negatively associated with bike share users. However, we found that temperature did help explain average daily trips, which they found inconclusive (Pucher & Buehler, 2017). The negative coefficients for Boston and Chicago demonstrated much higher number of trips in New York than in the other cities, which was most likely due to its large population (Table 5). This model showed convergence of parameters and high numbers of effective draws through trace plots of posterior draws (Figure 1). The final model also had the lowest WAIC value, which made it the most preferred model (Appendix B: Figure 3, Table 1). Lastly, by looking at the magnitudes of coefficients among the temporal variables, snowfall had the most impact on the number of bike trips, followed by temperature.

Limitations

Our model contains three major limitations. The model only takes into account three cities and although bike share data exist for cities beyond Boston, Chicago and New York, we

did not use them for two reasons: other cities do not contain information about gender, which we wanted to account for and adding cities caused the modeling process to be very inefficient. Therefore, we are interested in expanding our model to include other cities in the U.S.

The second limitation is due to the fact that we could not implement a mixed effect model with each city as a level even though we believed that the number of bike trips in each city was highly correlated. Although our analysis focused on estimation of the impact of each variable on the number of daily bike trips, a mixed effect model would be more informative because the city influence would not be restricted to the three cities modeled. In other words, we could apply this model to other cities that are not included in the model to predict number of daily trips in that city. We attempted to implement a mixed-effect poisson model with fixed intercepts (season) and random slopes (cities), but it did not converge after 100,000 iterations (Appendix B: Figure 1). In addition, the effective number of draws for both intercepts and random intercepts were extremely low compared to the data size, which shows a deficiency of the Markov chains for these variables. Including other cities might have helped the model converge; however, with more data entries, we suspect that with our processing power, the simulation would take too much time. Therefore, future studies can look into implementing a multi-level poisson model.

Thirdly, this model does not take into consideration city characteristics because the model did not converge when we included variables such as number of stations in each city and street orientation (Appendix B: Figure 2). In addition, the effective number of draws for each variable was much lower than the number of data samples. We suspect that this was due to a limited number of cities used in the model. Since previous research shows a positive association between city characteristics and number of daily bike trips, future studies can also look into their

impacts on the number of trips. Our model helped determine contributing factors and their impact on daily number of trips. Addressing limitations of our analysis will help construct a more informative model that will be useful for prediction in multiple cities in the U.S. This opens up a door for a new business to implement bike rental services in cities where services are currently not available.

References

Bike Share Data Portals. (2017). *Bike Share Data* [hubway-tripdata, Divy_Trips_2017, citibike-tripdata]. Retrieved from https://www.bikeshare.com/data/.

Boeing G. (2018, August). Urban Spatial Order: Street Network Orientation, Configuration, and Entropy. Retrieved November, 2018, from https://poseidon01.ssrn.com/delivery.php?ID=546004070078004089025109115007124089100051017087011048030016023006113096113119126127032110017005027000016114122119078119026025020034057083043027122029092087095004070086083071122124094069085124005118090025000008005120108119086125097004001090066027002114&EXT=pdf

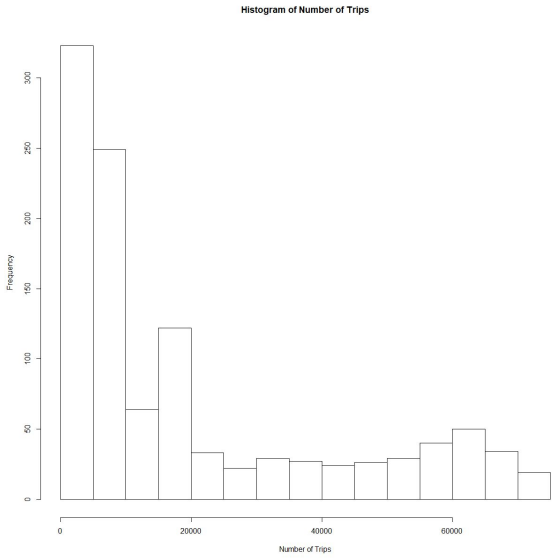
NOAA Climate Data Online. (2017). *Daily Summaries* [Boston, Chicago, New York]. Retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND.

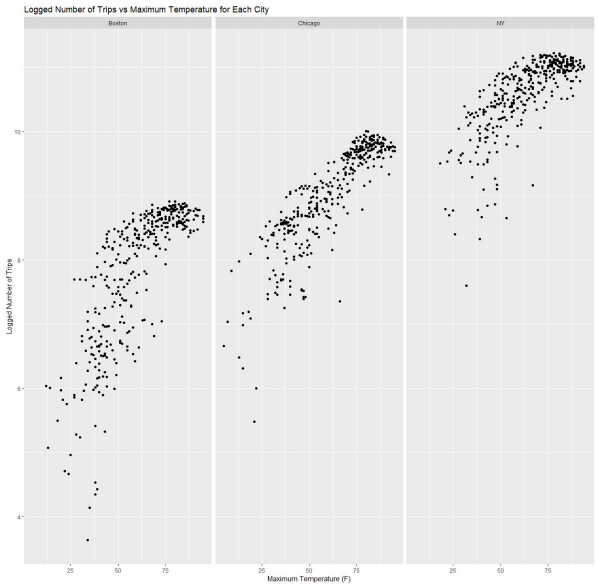
Pucher, John & Buehler, Ralph. (2006). Why Canadians cycle more than Americans: A comparative analysis of bicycling trends and policies. *Transport Policy, Elsevier, vol. 13(3)*. Retrieved November, 2018 from https://www.sciencedirect.com/science/article/pii/S0967070X05001381.

Zhang, Ying. (2017). Bike Sharing Usage: Mining on the Trip Data of Bike Sharing Usage. *University of Twente*. Retrieved November, 2018 from https://webapps.itc.utwente.nl/librarywww/papers_2017/phd/yzhang.pdf

Appendix

Appendix A: Exploratory Data Analysis





Appendix B: Model Exploration

Figure 1: Mixed Effects Trace Plot (Number of iteration: 100,000, Warm-up: 80,000)

a[1]: Season (Fall), a[2]: Season (Spring), a[3]: Season (Summer), a[4]: Season (Winter), b_4: Maximum Temperature, b_5: Precipitation, b_6: Snow (Yes:1, No: 0), b_7: Average Wind Speed, b_8: Weekends (Yes:1, No: 0), b_9: Median Age, b_10 * per_male: Percentage of Males, b_1: Percentage of Subscribers, a_city[1]: Boston, a_city[2]: Chicago, a_city[3]: NY, sigma_city: standard deviation for City (random slope)

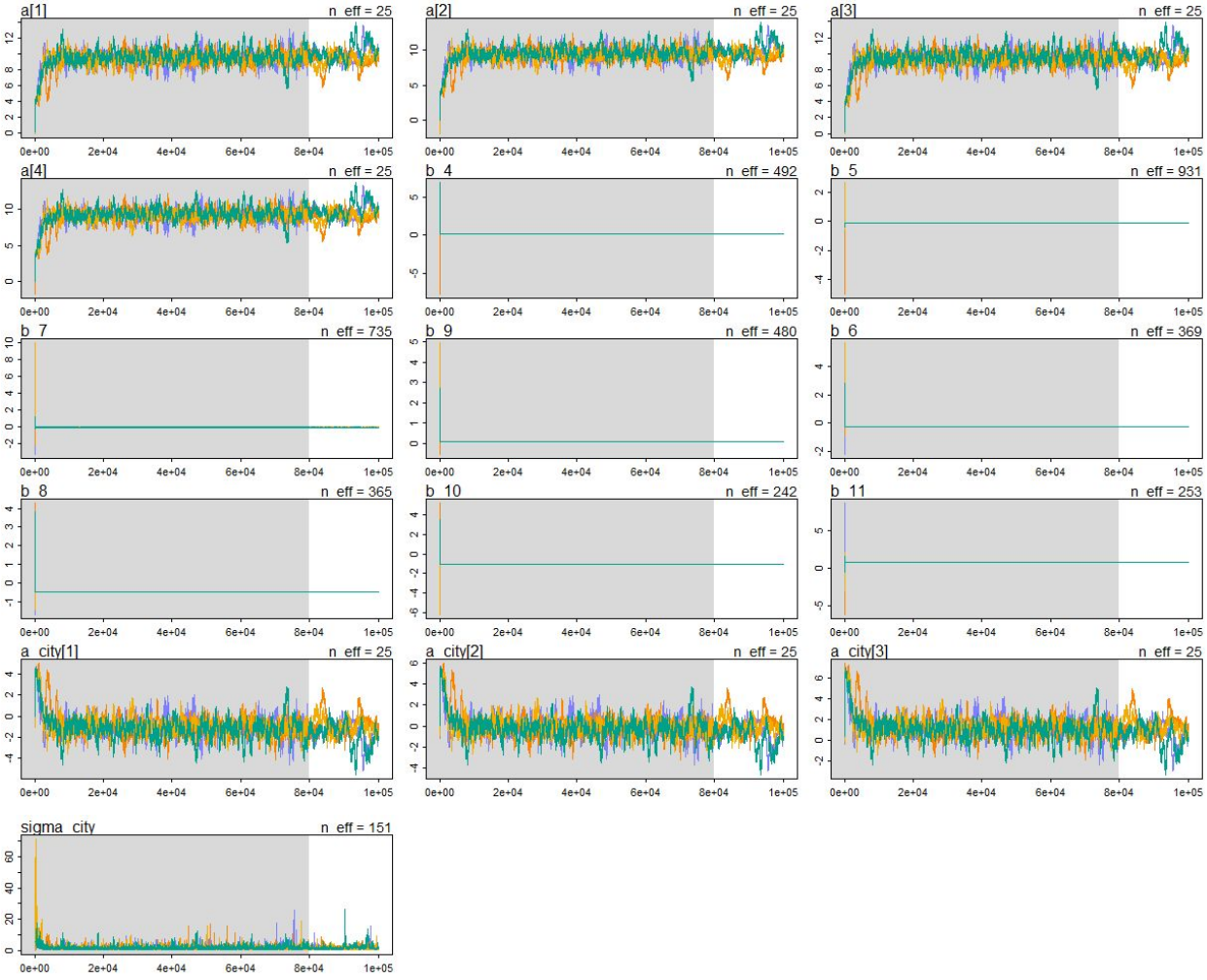


Figure 2: Trace Plot of Model with Stations and Street Orientation (Number of iteration: 30,000, Warm-up: 25,000)
a: intercept, b_1: Boston (Yes:1, No: 0), b_2: Chicago (Yes:1, No: 0), b_14: Street Orientation, b_15: Number of Stations, b_6: Precipitation, b_7: Median Age, b_8: Percentage of Males, b_9: Percentage of Subscribers, b_10: Snow (Yes:1, No: 0), b_11: Maximum Temperature, b_12: Weekends (Yes:1, No: 0), b_13: Average Wind Speed

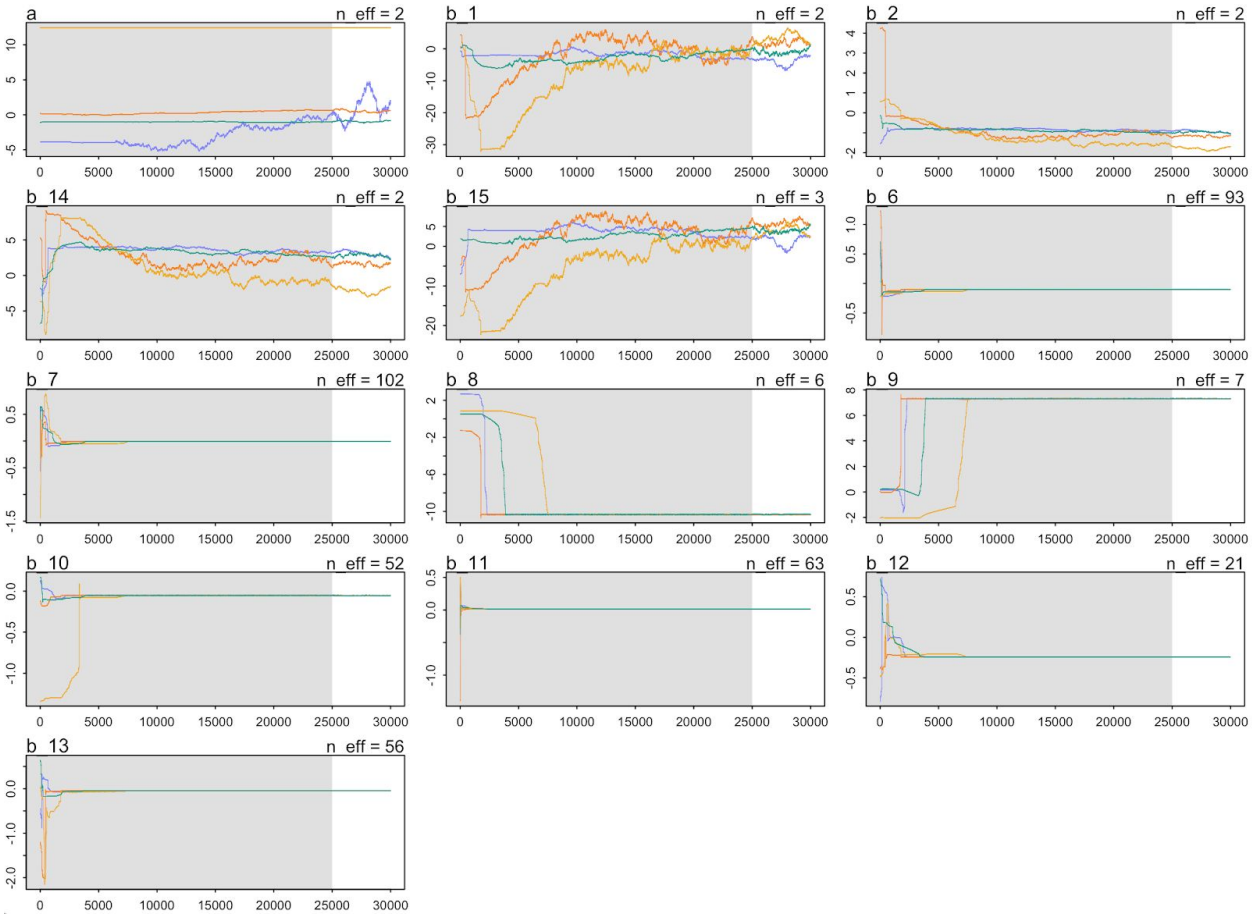


Figure 3: WAIC Plot

*Comparing in-sample and out of sample deviance between final model, model without seasonal data, and model without seasonal and demographic data

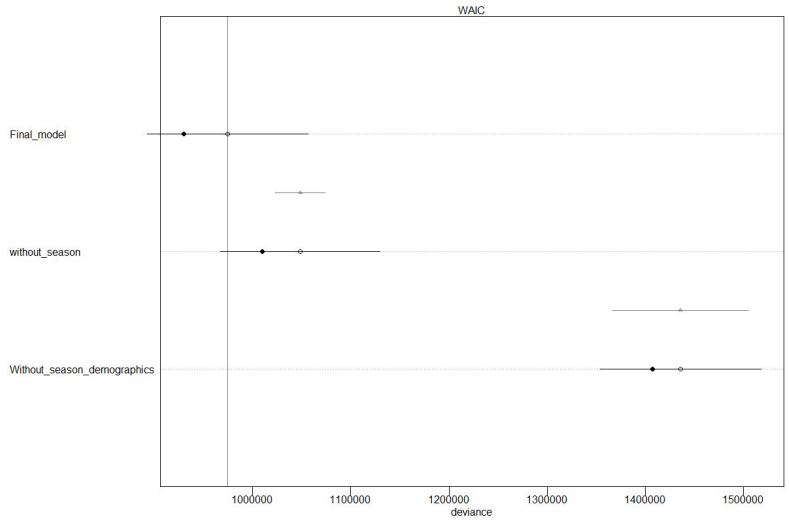
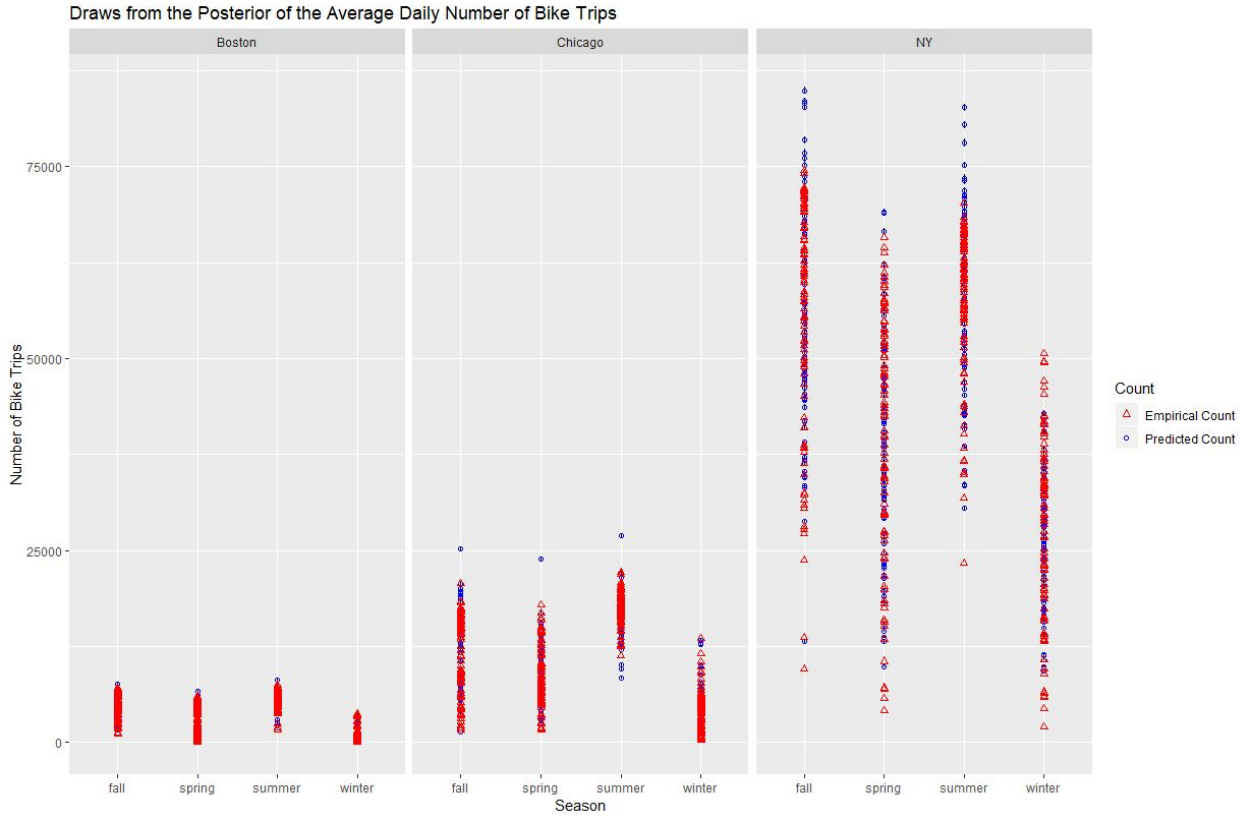


Table 1: Model Comparison

	WAIC	pWAIC	dWAIC	weight	SE	dSE
Final Model	974703	22476	0	1	82045	NA
Without Seasons	1048611	19254	73908	0	80959	25099
Without Seasons and Demographics	1435860	14325	461157	0	81995	69441

Appendix C: Posterior Prediction

Although this project does not aim to predict the daily number of bike trips, we created a posterior prediction.



This illustrates that winter showed lower empirical counts and summer showed higher empirical counts for average number of daily trips. In addition, it demonstrates that New York had much higher empirical counts of daily trips than Chicago and Boston, which was most likely due to its large population.